Lecture Notes in Artificial Intelligence 5345

Edited by R. Goebel, J. Siekmann, and W. Wahlster

Subseries of Lecture Notes in Computer Science

Takahira Yamaguchi (Ed.)

Practical Aspects of Knowledge Management

7th International Conference, PAKM 2008
Yokohama, Japan, November 22-23, 2008
Proceedings

Springer

Series Editors

Randy Goebel, University of Alberta, Edmonton, Canada
Jörg Siekmann, University of Saarland, Saarbrücken, Germany
Wolfgang Wahlster, DFKI and University of Saarland, Saarbrücken, Germany

Volume Editor

Takahira Yamaguchi
Faculty of Science and Technology, Keio University
3-14-1 Hiyoshi Kohoku-ku, Yokohama 223-8522, Japan
E-mail: yamaguti@ae.keio.ac.jp

Library of Congress Control Number: 2008938967

CR Subject Classification (1998): I.2, H.2.8, H.3-5, K.4, J.1

LNCS Sublibrary: SL 7 – Artificial Intelligence

ISSN 0302-9743
ISBN-10 3-540-89446-2 Springer Berlin Heidelberg New York
ISBN-13 978-3-540-89446-9 Springer Berlin Heidelberg New York

Typesetting: Camera-ready by author, data conversion by Scientific Publishing Services, Chennai, India
Printed on acid-free paper SPIN: 12566006 06/3180 5 4 3 2 1 0

Preface

The biennial PAKM Conference Series offers a communication platform and meeting ground for practitioners and researchers involved in developing and deploying advanced business solutions for the management of knowledge in organizations. PAKM is a forum for people to share their views, exchange ideas, develop new insights, and envision completely new kinds of knowledge management solutions.

PAKM 2008, the 7th International Conference on Practical Aspects of Knowledge Management, was held in Yokohama, Japan, for the first time. Although all past PAKM conferences were held in Europe (Basel and Vienna), the PAKM Steering Committee decided two years ago that the PAKM conference should be "on tour": it should be organized by different people and be hosted in different places all over the world.

For this year's conference we received 62 submissions from 23 countries and 3 reviewers were assigned to one paper from the members of the Program Committee and the additional reviewers. Thus 23 good papers were selected. They cover a great variety of approaches to knowledge management, which tackle the topic from many different angles. It is this very diversity that makes PAKM unique, while at the same time focusing on the one issue of managing knowledge within organizations.

Many people were involved in setting up PAKM 2008. We would like to express our warm thanks to everybody who contributed to making it a success. First of all, this includes all the authors who submitted a paper to the review process, the members of the Program Committee and the additional reviewers who made such an effort to select the best papers and to ensure a high-quality program. My thanks also go to Ulrich Reimer, who gave me so many useful comments for all the organizational work based on his excellent experience from past PAKM conferences, and to Takeshi Morita, who devoted himself to all the aspects involved in planning and setting up the conference program.

November 2008 Takahira Yamaguchi

Organization

Steering Committee

Chairs

Dimitris Karagiannis	University of Vienna, Austria
Ulrich Reimer	University of Applied Sciences St. Gallen, Switzerland

Members

Irma Becerra-Fernandez	Florida International University, USA
John Davies	British Telecom, UK
Rose Dieng	INRIA, France
Michael Huhns	University of South Carolina, USA
Daniel O'Leary	University of Southern California, USA
Eric Tsui	The Hong Kong Polytechnic University, China
Mary-Anne Williams	The University of Technology, Sydney, Australia
Takahira Yamaguchi	Keio University, Japan

Conference Chair

Takahira Yamaguchi	Keio University, Japan

Program Committee

Irma Becerra-Fernandez	Florida International University, USA
Xavier Boucher	Ecole des Mines de St. Etienne, France
Kemal A. Delic	Hewlett-Packard, France
Juan Manuel Dodero	University of Cádiz, Spain
Joaquim Filipe	Escola Superior de Tecnologia Setubal, Portugal
Naoki Fukuta	Shizuoka University, Japan
Fabien Gandon	INRIA Sophia, France
Aldo Gangemi	CNR-ISTC, Italy
Ulrich Geske	Fraunhofer Gesellschaft FIRST, Germany
Enrico Giunchiglia	Università di Genova, Italy
Norbert Gronau	University of Potsdam, Germany
Yoshinori Hara	Kyoto University, Japan
Yusuke Hayashi	Osaka University, Japan
Remko Helms	Universiteit Utrecht, The Netherlands
Melanie Hilario	University of Geneva, Switzerland
Knut Hinkelmann	University of Applied Sciences Nordwestschweiz, Switzerland

Achim Hoffmann	University of New South Wales, Australia
Gaku Ishii	Toshiba Corporation, Japan
Takayuki Ito	Nagoya Institute of Technology, Japan
Noriaki Izumi	AIST, Japan
Manfred Jeusfeld	University of Tilburg, The Netherlands
Masakazu Kanbe	NTT Data, Japan
Byeong Ho Kang	University of Tasmania, Australia
Josef Küng	Johannes Kepler University Linz, Austria
Ronald Maier	University of Innsbruck, Austria
Vladimir Marik	Czech Technical University, Czech Republic
Michele Missikoff	Italian National Research Council, Italy
Takeshi Morita	Keio University, Japan
Nikos Mylonopoulos	ALBA, Greece
Alun Preece	Cardiff University, UK
Sven-Volker Rehm	WHU - Otto Beisheim School of Management, Germany
Peter Reimann	University of Sydney, Australia
Debbie Richards	Macquarie University, Australia
Bodo Rieger	University of Osnabrueck, Germany
Gerold Riempp	European Business School, Germany
Yoshihisa Shinozawa	Keio University, Japan
Steffen Staab	University of Koblenz-Landau, Germany
Rudi Studer	University of Karlsruhe, Germany
Motoyuki Takaai	Fuji Xerox Co., Ltd., Japan
Ulrich Thiel	Fraunhofer Gesellschaft, Germany
A Min Tjoa	Technical University of Vienna, Austria
Klaus Tochtermann	I-Know Center Graz, Austria
Eric Tsui	Hong Kong Polytechnic University, China
Roland Wagner	Johannes Kepler University Linz, Austria
Hideo Watanabe	IBM Research, Japan
Frithjof Weber	EADS, Germany
Rosina Weber	Drexel University, USA
Mary-Anne Williams	University of Technology Sydney, Australia
Shuichiro Yamamoto	NTT Data, Japan
Takeshi Yoshioka	Fuji Xerox Co., Ltd., Japan

Additional Reviewers

Juergen Bock	Alexandros Kalousis	Alexander Stocker
Thomas Franz	Antonio De Nicola	Francesco Taglino
Gisela Granitzer	Christoph Ringelstein	Valentin Zacharias
Hans-Jörg Happel	Marco Spruit	

Local Organization

Chair

Yoshihisa Shinozawa	Keio University, Japan

Members

Naoki Fukuta	Shizuoka University, Japan
Yoshinori Hara	Kyoto University, Japan
Tadashi Iijima	Keio University, Japan
Mitsuru Ikeda	JAIST, Japan
Noriaki Izumi	AIST, Japan
Takeshi Morita	Keio University, Japan

Sponsors

International Communications Foundation
Support Center for Advanced Telecommunications Technology Research, Foundation

Table of Contents

Invited Papers

Knowledge Sharing

Collaboration Platforms

Content-Oriented Retrieval

Knowledge Acquisition

Knowledge Management Solutions

Knowledge Mining from Data, Text and the Web

Ontology

Knowledge Utilization

Knowledge Exploratory for Service Management and Innovation

Yoshinori Hara

Graduate School of Management, Kyoto University
Yoshida-Honmachi, Sakyo-ku, Kyoto 606-8501, Japan
hara@gsm.kyoto-u.ac.jp

Abstract. In this talk, various trends in the current service economy are explained. Global economy has turned out to be a service-oriented economy, in addition to the fact that the economy is going along with more digitized information rather than just labor intensive human service activities. In Japan's case, for example, more than 70% of GDP has been created by service sectors in a broader sense recently, and the ratio is still going up.

However, the problem we are facing to the service-oriented economy emerges due to the uncertainness of service management and the low productivity of service businesses in general. We explain how knowledge exploratory could contribute to providing a solution or an insight to improve the situation as "service innovation", and to discuss framework and/or design process of service innovation, illustrating with some of the actual research and education activities. They include service literacy management, service blueprinting based on knowledge management framework and common data models such as UML. The goal of this talk is to provide a common view of the knowledge management framework that will support service innovation for the service-oriented economy. We believe that this kind of systematic approach, having human beings located in the system, will have more meaningful implications for the new economy.

T. Yamaguchi (Ed.): PAKM 2008, LNAI 5345, p. 1, 2008.

From Corporate Memory to Corporate Knowledge

Dennis Tsichritzis

Ricoh, Research and Development Group
2, Route de Florissant 1206 Geneva, Switzerland
dennis.tsichritzis@gmail.com

Abstract. Technology enables us to store and recuperate with a Google like interface all documents produced within the boundaries of a company. Does this mean we have complete Corporate Memory? Yes and No because there are problems to be solved in Accountability, Authority and Responsibility. This is not a free for all environment where people can co-create material of nonuniform quality and value. We are also far for having Corporate Knowledge since the rules, the processes, the acceptable scenarios, the past lessons and the successful cases are missing. We will discuss what type of environment will be needed and outline what problems need to be solved. This talk does not represent in any way the positions of Ricoh as a company and should be considered as an individual opinion.

T. Yamaguchi (Ed.): PAKM 2008, LNAI 5345, p. 2, 2008.

Behavior and Social Influence in Knowledge Sharing: Intention Formation and the Moderating Role of Knowledge Type

Joseph C. Shih[1,2] and C.K. Farn[2]

[1] Lunghwa University of Science and Technology, Taiwan
joseph@mail.lhu.edu.tw
[2] National Central University, Taiwan
ckfarn@mgt.ncu.edu.tw

Abstract. The primary purpose of the paper is to examine the ways in which social information affects knowledge-sharing behavior in an organization. Based on theory of reasoned action and focusing on knowledge sharing setting, we propose that subjective norms and attitudes influence behavioral intention. Three processes drawing from social information processing theory (i.e., internalization, identification, and compliance) are postulated as antecedents of the intention to share knowledge. We also posit that knowledge type (as a moderator) intervenes the forming patterns of sharing behavior. Structural equation modeling was used to test hypotheses. Empirical data are collected from 229 respondents and our arguments were statistically supported. Some theoretical and practical implications are also discussed.

Keywords: Knowledge sharing, TRA, social influence, social information process, social norms, tacit knowledge, explicit knowledge.

1 Introduction

Research Motivations. Recently, the need for further examination of the social factors of knowledge management has been emphasized [11, 13, 23, 30, 33]. Besides, knowledge types, namely tacit and explicit knowledge, are not alike in many aspects [26]. Many scholars have noticed that tacit knowledge is different from explicit knowledge in many ways, including characteristics, hoarding, distribution, and so on [2, 10, 15, 20, 21, 31, 34].

Research Question. Social information is important for people within a workgroup because team members collect cues on what others do and opinion what others think to guide their behavior [14, 24]. This study connects the employee's knowledge sharing behaviors with the sources of social information to understand the formation of this pro-social behavior. This is the first goal of this study. Additionally, a premise stirs the research that tacit and explicit knowledge are distinctive in their nature and sharing approaches [8, 16]. Based on this notion, the second goal of our research is to better understand how knowledge type intervene the effects of social information on sharing behavior.

T. Yamaguchi (Ed.): PAKM 2008, LNAI 5345, pp. 3–13, 2008.

2 Conceptual Background

Perspective of Knowledge Management. Nonaka [26] explicates two dimensions of knowledge in organizations—tacit and explicit knowledge. From the perspective of knowledge distribution, explicit knowledge is the knowledge presented easily by language and documents, whereas tacit knowledge is not. Tacit knowledge is difficult to codify and can never be made explicit. Accordingly, explicit knowledge can be transferred across individual and organizational boundaries by codification. For example, an owner contributes his knowledge to the organizational storage, and then a beneficiary takes it from the repository. However, tacit knowledge can not be taught by reading manuals but must be learned through experience [8] or be absorbed by means of owner's impartation with great enthusiasm.

Social Information Processing Theory. The social information processing (SIP) perspective proceeds from the underlying premise that individuals, as an organism, adapt attitudes, behavior, and beliefs to their social context and environmental situation [29]. Deutsch and Gerard [17] distinguish two types of social influence, informational and normative social influences. The categorization of social influence and its type mapping are listed in Table 1. In this study, we argue that one's behavior of knowledge sharing is changed because of environmental conformity which is brought about from social pressures. For example, a workgroup with cooperative norms may invoke individual members' helping behavior (e.g., knowledge sharing).

Table 1. Influences of social information processing

	Internalization	Identification	Compliance
Type	Informational influence	Normative influence	Normative influence
Accepting reasons	The content of the induced behavior is intrinsically rewarding.	To establish or maintain a satisfying self-defining relationship to another person or a group. Associating with the desired relationship	To achieve a favorable reaction from another person or a group.
Occurrence	The behavior is congruent with his value system.	Taking over the role of the other or taking the form of a reciprocal role relationship.	Not because of believing in content but because of expecting to gain specific rewards or approval and avoid specific punishment or disapproval by conforming.
Satisfaction due to	The content of the new behavior	Act of conforming	Social effect of accepting influence.

Theory of Reasoned Action. Drawn from social psychology, TRA suggests that the best predictor of behavior is intention, which is determined by attitude and subjective norm [19]. According to TRA, sharing behavior is determined by sharing attitudes toward sharing and subjective norms of sharing. Due to the diverse nature and sharing approach of tacit and explicit knowledge, the intention, subjective norms, and attitude are distinctive in the light of knowledge type. For example, subjective norms of tacit knowledge and subjective norms of explicit knowledge are two constructs in this paper.

3 Model Development and Hypotheses

3.1 Research Model

To build up the research model, this study combines aspects of the Theory of Reasoned Action [19] with social information processing theory [29]. Three types of social information processing, namely, compliance, identification, and internalization, influence subjective norms and attitude toward knowledge sharing. Ajzen and Fishbein [1] state that subjective norm is an individual takes into account the normative expectations of various others in his or her environment. "Normative expectations" are socially agreed upon rules, moral, and value that people perceived whereas "subjective norm" refers to a specific behavioral prescription attributed to a generalized social agent.

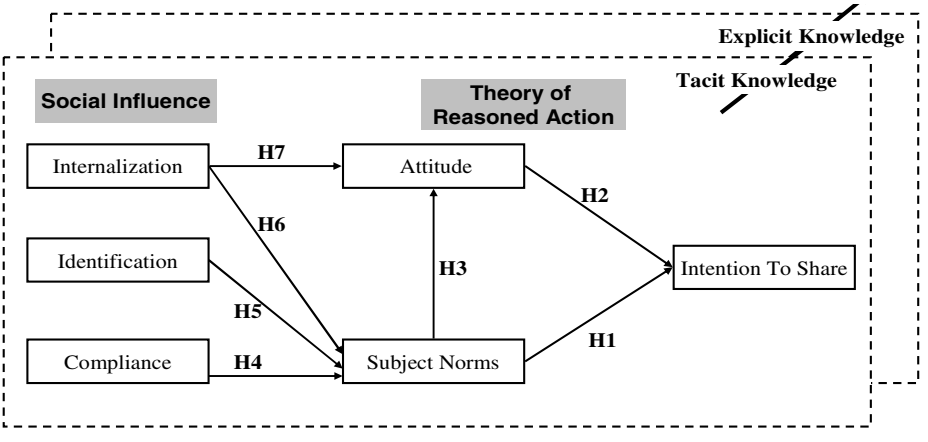


Fig. 1. Research Model

The research model is based on TRA and is presented in Figure 1. We propose that subjective norm of knowledge sharing is affected by the general processes of social influences. In addition, we also proposed that type of knowledge, tacit and explicit, may have different affections on attitudes and subjective norms. Because of the nature and sharing approach, type of knowledge may be a factor moderating the entire causal relationships.

3.2 Research Hypotheses

TRA argues that the best predictor of behavior is intention which is determined by attitude and subjective norm, and there exists positive relationship between attitude and subjective norm [19]. Using attitude and subjective norms to predict knowledge sharing behavioral intention formation is demonstrated in the past study [9], and the empirical evidences also support the notions. Thus,

H1: Intention of knowledge sharing is positively associated with subjective norm of knowledge sharing.

H2: Intention of knowledge sharing is positively associated with attitude toward knowledge sharing.

H3: Attitude toward knowledge sharing is positively associated with subjective norm of knowledge sharing.

Social influence reflect social pressure from significant others to perform a focal act [4]. In this study, we argue that general social influences (i.e., compliance, identity, and internalization) can fashion subjective norms of knowledge sharing. Some researchers have accented that general social influences act on knowledge contribution [4, 13, 15, 22, 23, 25, 33]. Likewise, some have emphasized the specific term—subjective norms of knowledge sharing [11, 15, 13, 23, 33].

According to the definition of compliance, a person thinks he should share knowledge not because he concerns the benefit of work team but because he expects to gain specific approval and avoid punishment [24]. In other words, conforming to share knowledge is motivated by the need for approval from significant other. This social affect of accepting influence—somewhat blind obedience, leads the perception that the focal person has to share knowledge in order to be liked by others. Thus,

H4: Subjective norm of knowledge sharing is positively associated with compliance.

The core tenet of identification is that a person derives a part of his self-concept from the work groups and categories they belong to. Bergami and Bagozzi [7] manifested that identification fosters loyalty and citizenship behaviors in the group setting. Wasko and Faraj [3] also argue that commitment to a group conveys a sense of responsibility to help others within the collective on the basis of shared membership. To sum up, they construe that identity can potentially contribute more knowledge to group. The notion that group identification can affect knowledge contribution is because a person may engage in more pro-social behavior (i.e., knowledge sharing) in order to benefit the group [13]. Thus,

H5: Subjective norm of knowledge sharing is positively associated with identity.

Internalization is occurred when an individual accepts influence because the substantial content of behavior is congruent with his values [24]. For example, a member who shares a common value of team will be more likely to become partners sharing and exchanging their resources [13]. Similarity of values reflects the extent to which members of an organization possess joint goals and interest, thus, the social influence of internalization may associate with knowledge contribution [23]. If an individual is of internalization, the reasons that he or she attempts to share knowledge are not only because everyone is part of the collective, but also "all have a collective goal orientation" (page 42) [33]. If members of a group have interdependent goals, the behaviors of helping or teaching the needed skills each other will be enhanced [22]. In light of internalization, it is not sufficient for a person to merely perceive reference group influence in order to consider the obligation to donate knowledge. Rather, the person perceives that he has duty or obligation to donate knowledge because the shared group values motive the willing of knowledge sharing [4, 13].

H6: Subjective norm of knowledge sharing is positively associated with internalization.

Because group shared value embodies the collective goals and aspirations of the members of a workgroup, an influence will be internalized if the person accepts the common principle shared by the workgroup, in turn, behaves pro-social behavior [13]. In this case, the attitude toward some particular behaviors is fashioned neither from concern about being punished nor from establishing favored relationship with others. Instead, the focal person links up group goal and vision with personal value system. Since the personal value corresponds with group value, it is more likely that the focal person shares or exchanges sources with other team member to provide better decision making for workgroup or to enhance the performance of task. The focal person should feel good and right if he can provide knowledge to his workgroup.

H7: Attitude toward knowledge sharing is positively associated with internalization.

According to social influence theory, compliance is that an individual hopes to attain a favorable relationship or avoid punishment from other but he does not regard to the nature of the behavior itself [24]. Besides, tacit knowledge is shared primarily through the established relationship between individuals [6]. Combining both notions, we are developing the last hypothesis.

First, Nonaka and Takeuchi [27] suggested four primary processes for creating knowledge that are based on converting tacit and explicit knowledge. They are externalization (tacit to explicit), socialization (tacit to tacit), internalization (explicit to tacit), and combination (explicit to explicit). Socialization refers to conversion of tacit knowledge to new tacit knowledge through social interactions and shared experience among organizational members. Social interaction is the primary way to sharing tacit knowledge between individuals, usually through joint activities and closure relationship rather compliance only. Second, evading punishment from others [24] is the primary reason for compliance. However, a person who has tacit knowledge but hoard will not be punished because other people never know the truth. In this case, there will no punishment if someone conceals tacit knowledge intentionally because other people are not able to sense the focal person's private knowing.

Moreover, explicit knowledge sharing is more coercive than tacit knowledge sharing because of organizational institution and supervisor control. This condition is very clear that an employee may be punished and lost favorable relationship if he does not share what he knows. Concealing one's own explicit knowledge, such as working report or documents, is aware easily by other people. Thus, a team member who hides explicit knowledge as private collection is easily accused, in turn, be punished. Compliance, therefore, may increase the perception of important people think one should share explicit knowledge but it does not act on tacit knowledge. Accordingly, we argue that the knowledge type moderates the relationship between subjective norm and compliance.

H8: The relationships among the variables are moderated by knowledge type.
H8a: In the case of explicit model, compliance is associated with subjective norm; in the case of tacit model, compliance is not associated with subjective norms

4 Research Method

Research Design and Data Collection. A web-based system was developed to distribute our questionnaire so that the scores of items could be collected and recorded into database automatically. Respondents were from 69 companies. Each participant was invited via email or MSN in advanced to assure the willingness to take part in the study. Respondents were informed that they would get a free gasoline ticket valued NTD$50 after finishing the questionnaire. There were totally 300 respondents joining the survey and 251 (83.33%) respondents answered the questionnaire. We dropped uncompleted questionnaires (they quitted midway) and some arbitrary answers judging from reversed items. Eventually, 229 (76.33%) valid questionnaires were offered for further data analysis. The female and male percentages of total respondents were 50.22% and 49.78%. The working contents included IT-related, administration, finance, sales, and others, and the percentages were 64%, 18%, 3%, 3%, and 12% respectively.

Scale Development. The items measuring intention, attitude, and subjective norm were adapted from Fishbein and Ajzen [1]. These constructs were measured in both tacit and explicit knowledge sharing settings separately. For example, there are two items for measuring subjective norms of tacit knowledge sharing as well as another two items for subjective norms of explicit knowledge sharing. Moreover, all items of subjective norms of both tacit and explicit knowledge sharing were combined as indicators of the attitude toward knowledge sharing (used in the main model). Likewise, we developed scales of attitude and intention through the same way. Sample items are listed in Appendix. Since the social influence constructs are unrelated to knowledge type, they are common between knowledge types. The items measuring compliance were adapted from Algesheimer et al. [3]. We followed the definition of social identification consisting of cognitive, affective, and evaluative components [4, 24]. Hence, we adopted the measuring strategy of Bagozzi and Dholakia [4] that identification is a construct with three dimensions. To measure internalization, we operationalized the construct as group shard value [28], consistent with previous studies [4]. All items mentioned above were modified into work place setting, seeking respondent's agreement on a 7-point Likert-like scale (strongly agree: 7 and strongly disagree: 1). Content validity and pre-test were performed to assure the quality of the scales.

5 Data Analysis and Results

5.1 Test of Measurement Model

LISREL 8.50 is employed to test the measurement model with maximum-likelihood estimation [9]. We examined the LISREL output and conducted some proper modifications, deriving a new measurement model with improved model fits. Overall model fit indices showed that the main model was a realistic representation of the data [$\chi^2(319)=729.35$, NFI=0.89, NNFI=0.92, CFI=0.93, GFI=0.82, RMSEA=0.075] whereas explicit model [$\chi^2(133)=233.40$, NFI=0.92, NNFI=0.95, CFI=0.96, GFI=0.90, RMSEA=0.058] and tacit model [$\chi^2(135)=305.49$, NFI=0.91, NNFI=0.93, CFI=0.94, GFI=0.88, RMSEA=0.074] were also fit excluding the GFI of tacit model, which was slightly lower. Due to the limitation of spaces, the details of output are omitted in this paper.

A confirmatory factor analysis was conducted. Our results showed that all loadings of each latent variable are above the cut-off value all item loadings are greater than or equal to the recommended cut-off level of 0.70 [5]. Convergent validity is thus established. We then applied the chi-square difference test to assess the discriminant validity of the measurement model [5]. We conducted 15 pair-wise tests (six constructs) for each model respectively. The results are showed that all $\Delta\chi^2$ differences are significant above the level of $Pr[\chi^2(1)\geq 3.84]=0.05$, indicating strong support for discriminant validity [5, 32]. Overall, the evidences of good model fit, reliability, convergent validity, and discriminant validity suggested that the measurement model was appropriate for testing the structural model at a subsequent stage.

5.2 Test of Structural Model

The results of the structural equation analysis are reported in Figure 2 to Figure 4, for the main model, explicit model, and tacit model respectively. Results show that H1 to H7 are supported, as depicted in the beta loading of the respective arrows in Figure 2. H8, examining the knowledge type as a moderator in the research model, is supported if the path from compliance to subjective norms is significant in explicit model while it is not significant in tacit model. Referring to the results of SEM, the path coefficient is $\beta=0.18$ ($t=2.93$, $p<0.01$, significant) for explicit model and $\beta=0.12$ ($t=1.93$, $p>0.05$, insignificant) for tacit model. Accordingly, H8 was also supported.

6 Discussion

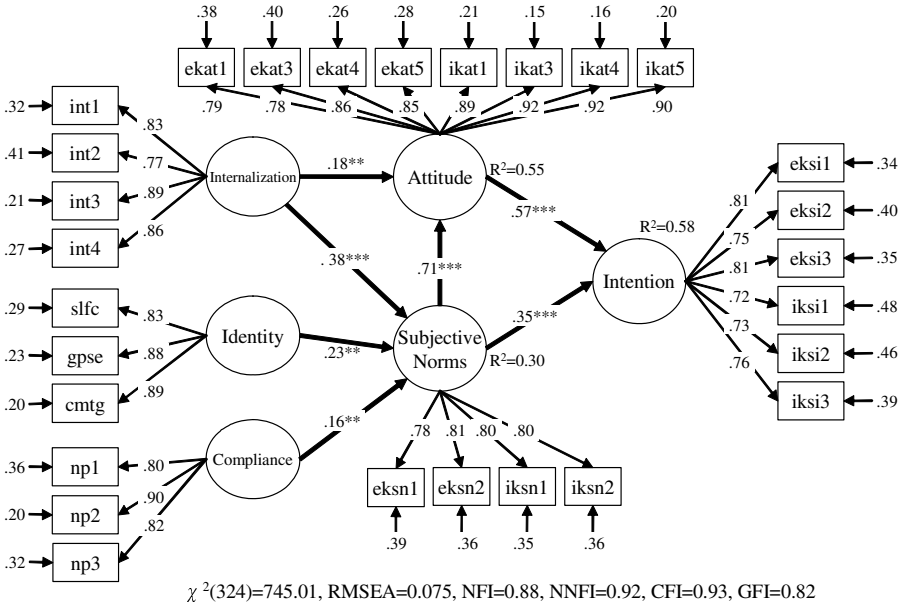
First, as a special case of TRA, knowledge sharing intention indeed associates with both attitude and subjective norms, and our results is consistent with previous studies, e.g. Bock et al.'s [11]. More specifically, TRA is also effective in explaining the forming of sharing intentions in both tacit and explicit knowledge settings.

Second, internalization, identification, and compliance are determinants of subjective norms of knowledge sharing. The finding is consistent with the results of Kankanhalli et al. [23] that identification toward workgroup is helpful for knowledge contribution. Our findings contribute to more understanding how knowledge sharing behaviors are aroused by social influence.

Third, compliance, identification, and internalization can shape the social conformity of knowledge sharing. Based on our findings, the relationship between compliance and subjective norms is significant in the explicit knowledge model but it not significant in the tacit knowledge model. Simply, the pressure of blind obedience is ineffective to foster tacit knowledge sharing intention.

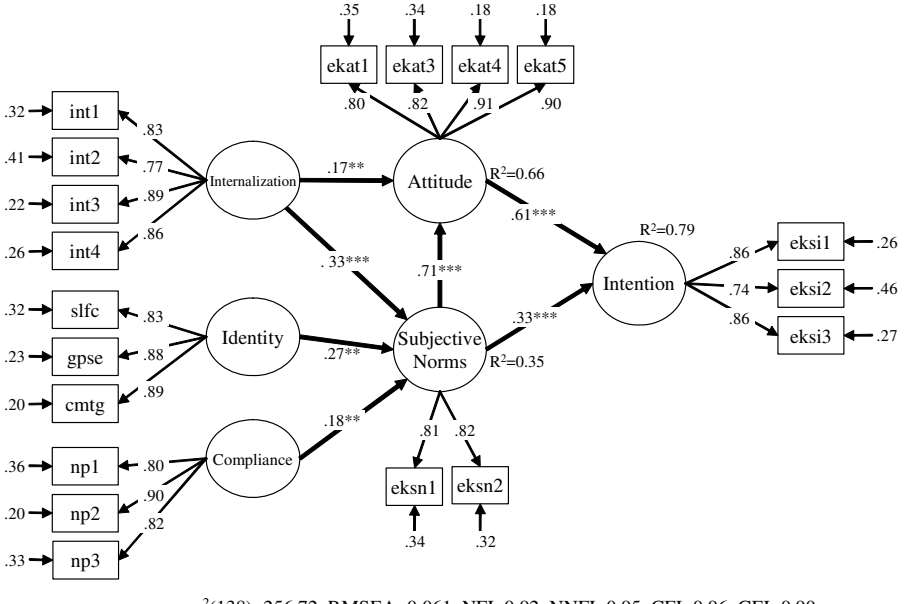
Fourth, our model is different from Bock et al. [11] in measuring intention of knowledge sharing. We treated tacit knowledge sharing and explicit knowledge sharing as two different settings, and measured them separately. We used both tacit and explicit items to measure intention, attitude, and subjective norm respectively. This strategy goes beyond the generic knowledge sharing behavior and looks more specifically into both tacit and explicit knowledge sharing intention forming.

Lastly, a surprising result is the non-significant two relationships in tacit model, between attitude and intention, and between intention and attitude. According to the results, we may draw out the finding that intention of tacit knowledge sharing is determined by subjective norms primarily.



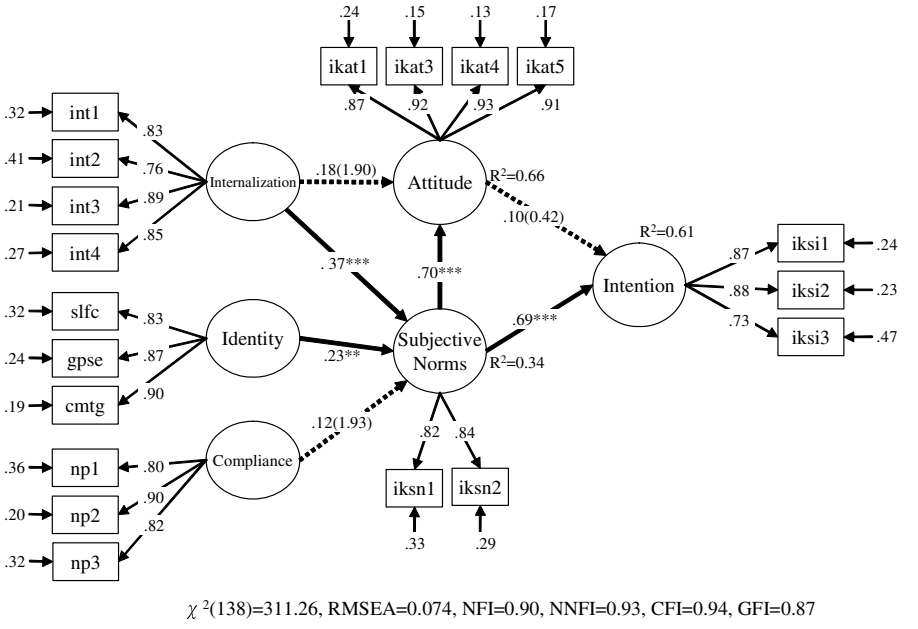
$\chi^2(324)=745.01$, RMSEA=0.075, NFI=0.88, NNFI=0.92, CFI=0.93, GFI=0.82

Fig. 2. Main Model



$\chi^2(138)=256.72$, RMSEA=0.061, NFI=0.92, NNFI=0.95, CFI=0.96, GFI=0.90

Fig. 3. Explicit Knowledge Model



$\chi^2(138)=311.26$, RMSEA=0.074, NFI=0.90, NNFI=0.93, CFI=0.94, GFI=0.87

Fig. 4. Tacit Knowledge Model

7 Conclusion

In this study, we examined how social information processing shapes knowledge sharing intention through the perspective of TRA. Particularly, the patterns forming tacit knowledge sharing intention is different from that of explicit knowledge sharing. Knowledge sharing intention can be changed and influenced unobtrusively and imperceptibly whereby knowledge owners behave in conformity with the social norms, which are derived from the combination of internalization, identification, and compliance. Due to the distinctive sharing nature and approach of tacit and explicit knowledge, knowledge type moderates in the formation of sharing behavior in two aspects: (1) compliance takes no effect on intention of tacit knowledge sharing; and (2) group shared value affects attitude through subjective norms in tacit context. Given the significance of knowledge sharing, we hope our findings could provide some social-psychological insights to enrich our understanding about employees' collective willingness of knowledge sharing.

References

1. Ajzen, I., Fishbein, M.: Understanding Attitudes and Predicting Social Behaviour. Prentice-Hall, Englewood Cliffs (1980)
2. Alavi, M., Leidner, D.E.: Review: Knowledge Management and Knowledge Management Systems: Conceptual Foundation and Research Issues. MIS Quarterly 25(1), 107–136 (2001)

3. Algesheimer, R., Dholakia, U.M., Herrmann, A.: The Social Influence of Brand Community: Evidence from European Car Clubs. Journal of Marketing 69(3), 19–34 (2005)
4. Bagozzi, R.P., Dholakia, U.M.: Intentional Social Action in Virtual Communities. Journal of interactive Marketing 16(2), 2–21 (2002)
5. Bassellier, G., Benbasat, I., Reich, B.H.: The influence of Business Managers IT Competence on Championing IT. Information System Research 14(4), 317–336 (2003)
6. Becerra-Fernandez, I., Sabherwal, R.: Organizational Knowledge Management: A Contingency Perspective. Journal of Management Information Systems 18(1), 23–55 (2001)
7. Bergami, M., Bagozzi, R.P.: Self-Categorization, Affective Commitment and Group Self-Esteem as Distinct Aspects of Social Identity in the Organization. British Journal of Social Psychology 39(4), 555–577 (2000)
8. Berman, S.L., Down, J., Hill, C.W.L.: Tacit Knowledge As A Source of Competitive Advantage in the National Basketball Association. Academy of Management Journal 45(1), 13–31 (2002)
9. Bollen, K.A.: Structural Equations with Latent Variables. John Wiley and Sons, Chichester (1989)
10. Bordia, P., Irmer, B.E., Abusah, D.: Differences in Sharing Knowledge interpersonally and Via Databases: The Role of Evaluation Apprehension and Perceived Benefits. European Journal of Work and Organizational Psychology 15(3), 262–280 (2006)
11. Bock, G.W., Zmud, R.W., Kim, Y.G., Lee, J.N.: Behavioral intention Formation in Knowledge Sharing: Examining the Roles of Extrinsic Motivators, Social-Psychological Forces, and Organizational Climate. MIS Quarterly 29(1), 87–111 (2005)
12. Chin, W.W.: Commentary: Issues and Opinion on Structural Equation Modeling. MIS Quarterly 22(1), vii–xvi (1998)
13. Chiu, C.M., Hsu, M.H., Wang, E.T.G.: Understanding Knowledge Sharing in Virtual Communities: An Integration of Social Capital and Social Cognitive Theories. Decision Support Systems 42(3), 1872–1888 (2006)
14. Cialdini, R.B., Goldstein, N.J.: Social Influence: Compliance and Conformity. Annual Review of Psychology 55, 591–621 (2004)
15. Constant, D., Kiesler, S., Sproull, L.: What's Mine Is ours, or Is It? A Study of Attitudes about Information Sharing. Information Systems Research 5(4), 400–421 (1994)
16. Cross, R., Baird, L.: Technology Is Not Enough: Improving Performance by Building Organizational Memory. Sloan Management Review 41(3), 69–78 (2000)
17. Deutsch, M., Gerard, H.B.: A Study of Normative and Informational Social Influences upon Individual Judgment. Journal of Abnormal and Social Psychology 51(3), 629–636 (1955)
18. Doll, W.J., Raghunathan, T.S., Lim, J., Gupta, Y.P.: A Confirmatory Factor Analysis of the User Information Satisfaction Instrument. Information Systems Research 6(2), 177–188 (1995)
19. Fishbein, M., Ajzen, I.: Belief, Attitude, intention and Behavior: An Introduction to Theory and Research. Addison-Wesley, U.S.A. (1975)
20. Grover, V., Davenport, T.H.: General Perspectives on Knowledge Management: Fostering a Research Agenda. Journal of Management Information Systems 18(1), 5–21 (2001)
21. Hansen, M.T., Nohria, N., Tierne, T.: What's Your Strategy for Managing Knowledge. Harvard Business Review 77(5), 106–116 (1999)
22. Janz, B.D., Prasaphanich, P.: Understanding the Antecedents of Effective Knowledge Management: The Importance of a Knowledge-Centered Culture. Decision Sciences 34(2), 351–383 (2003)

23. Kankanhalli, A., Tan, B.C.Y., Wei, K.K.: Contributing Knowledge to Electronic Knowledge Repositories: An Empirical investigation. MIS Quarterly 29(1), 113–143 (2005)
24. Kelman, H.C.: Compliance, Identification, and internalization: Three Processes of Attitude Change. Journal of Conflict Resolution 2(1), 51–60 (1958)
25. Levin, D.Z., Cross, R.: The Strength of Weak Ties You Can Trust: The Mediating Role of Trust in Effective Knowledge Transfer. Management Science 50(11), 1477–1490 (2004)
26. Nonaka, I.: Dynamic Theory of Organizational Knowledge Creation. Organization Science 5(1), 14–35 (1994)
27. Nonaka, I., Takeuchi, H.: The Knowledge-Creating Company. Oxford University Press, New York (1995)
28. Netemyer, R.G., Boles, J.S., Mckee, D.O., Mcmurrian, R.: An investigation into the Antecedents of Organizational Citizenship Behaviors in A Personal Selling Context. Journal of Marketing 61(3), 85–98 (1997)
29. Salancik, G.R., Pfeffer, J.: A Social Information Processing Approach to Job Attitudes and Task Design. Administrative Science Quarterly 23(2), 224–253 (1978)
30. Sveiby, K.E.: Disabling the Context for Knowledge Work: The Role of Managers' Behaviours. Management Decision 45(10), 1636–1655 (2008)
31. Tuomi, I.: Data Is More than Knowledge: Implications of the Reversed Knowledge Hierarchy for Knowledge Management and Organizational Memory. In: Proceedings of the 32nd Hawaii International Conference on System Sciences, Hawaii, USA, pp. 1–12 (1999)
32. Venkatraman, N.: Strategic Orientation of Business Enterprises: The Construct, Dimensionality, and Measurement. Management Science 35(8), 942–962 (1989)
33. Wasko, M.M., Faraj, S.: Why Should I Share? Examining Social Capital and Knowledge Contribution in Electronic Networks of Practice. MIS Quarterly 29(1), 35–57 (2005)
34. Zander, U., Kogut, B.: Knowledge and the Speed of the Transfer and Imitation of Organizational Capabilities: An Empirical Test. Organization Science 6(1), 76–92 (1995)

Appendix

Followings are samples items of explicit and tacit knowledge for subjective norms. Others constructs, attitude and intention, are omitted because of limited space.

Subjective norms of explicit knowledge sharing

People who are important to me think that I should share my work reports.
People who are important to me think that I should share my official documents.

Subjective norms of tacit knowledge sharing

People who are important to me think that I should share my experience.
People who are important to me think that I should share my know-how.

A Lightweight Approach for Knowledge Sharing in Distributed Software Teams*

Walid Maalej[1] and Hans-Jörg Happel[2]

[1] Technische Universität München
Munich, Germany
maalejw@in.tum.de
[2] FZI Research Center for Information Technologies
Karlsruhe, Germany
happel@fzi.de

Abstract. In collocated software development teams, informal communication is the key enabler for sharing knowledge. In distributed teams, development infrastructures have to fill communication gaps with lightweight articulation and sharing facilities for evolving development knowledge. We propose an ontology-based framework to capture, access and share developers' experiences in a decentralized, contextualized manner. Capturing developers' interaction with related artifacts and providing a Wiki-like annotation approach triggers knowledge capture. Integrated semantic search and recommendation fosters knowledge access and sharing. Our framework enables distributed teams to become more effective by learning from each other's experiences, e.g. on reusing specific components and handling semantic errors.

1 Introduction

The success of agile methodologies for collocated teams, together with the increasing distribution of software projects raises new challenges for development infrastructures to support collaboration and knowledge sharing in software development [1,2,3]. To ensure tight collaboration and continuous knowledge sharing, agile methodologies embrace informal communication [4], task overlapping and interdisciplinary teams. This manifests in practices such as daily Scrum meetings [5], pair programming [6] or test driven development. On the other hand, distributed development hinders informal collaboration and knowledge sharing due to reduced communication bandwidth, as well as social and cultural distance. Therefore, development artifacts constitute the main common ground between the team members while experiences and context information are rarely shared. In collocated teams, experiences such as encountered problems and argumentation behind specific decisions are informally shared, enabling members to synchronize their work and learn from each other. Such context information is lost

* This work has been supported in part by the TEAM project, which is funded by the EU-IST programme under grant FP6-35111 and the BMBF-funded project WAVES.

T. Yamaguchi (Ed.): PAKM 2008, LNAI 5345, pp. 14–25, 2008.

in distributed teams. To fill this gap, development infrastructures should provide pragmatic and lightweight support for knowledge sharing. Missing support for knowledge sharing confines the efficiency and productivity of distributed developers, because software development is a knowledge intensive activity, where trial-and-error is a dominant technique for resolving problems. For example, due to high interdependency among software components, discovering of the right source of a runtime error may require several trials, leading to particular experiences in specific situations. The heavy usage of mailing lists, newsgroups and forums shows that developers benefit from reusing positive and negative experiences that other developers have had with similar problems [7].

In the following we discuss the need and enablers for lightweight knowledge sharing in distributed teams (section 2), and introduce a framework for a decentralized, and context-aware capturing and sharing of knowledge, based on semantic technologies (section 3). We conclude by comparing our model to earlier efforts (section 4) and describing future directions (section 5).

2 Lightweight Knowledge Sharing

As a first step towards conceptualizing a lightweight knowledge sharing framework, we describe desired system capability by showing exemplary problems in current practices. We then analyze shortcomings in knowledge access and sharing and identify enablers for this framework.

2.1 Problem Scenarios

The following scenarios – handling errors and reusing components – show how a lack of knowledge sharing facilities negatively affects the efficiency and productivity of distributed development teams.

Handling Errors. In the case of semantically rich errors such as unexpected system behaviours or runtime errors, the state of the practice is to "google" for error message excerpts or keywords describing the context, in order to find relevant hints how to handle that error. It is, e.g., useful to find out where other developers looked for help while having similar problems. Thereby a main issue arises from the different contexts between the developer who consumes knowledge, and the developer who can provide knowledge about the error situation. Typically these situations are never identical, resulting in a "context gap". This gap can either be bridged by consumers of knowledge (adapting a solution to their problem situations) or by the providers of knowledge (generalizing their experiences to other problem situations). Both cases require extra effort. In collocated agile teams this is addressed informally since context is continuously communicated.

Reuse of Components. As the "construction by configuration" approach continues to emerge [8], reuse is undertaken at a higher level of abstraction and complexity. Correct components integration, effective work with powerful frameworks or successful usage of design patterns requires significant background

knowledge and experience about concerned components. Reusing components and frameworks – in particular those, which have been created externally – is a non-trivial task. Documentation, if it exists, is insufficient to describe all ways to use a framework. In addition, such knowledge is scattered across different sources such as emails, forums, specifications or bug reports, especially in open source development. Again, the context is needed in order for the knowledge to be useful, and a match between consumer's and provider's knowledge is required. Research showed that personal knowledge (especially reuse-related experiences) plays a significant role for the success of reuse in software engineering, besides tools and processes [9,10]. Hence, tackling the personal knowledge and experience dimension in distributed development teams can be considered as orthogonal, complementary to existing approaches, with a leverage effect for the usefulness of all other endeavours to increase team productivity.

2.2 Knowledge Exchange Processes

From a process perspective, knowledge sharing can be defined as the "dual problem of searching for (looking for and identifying) and transferring (moving and incorporating) knowledge across organizational subunits" [11]. It mediates between two roles: the consumer of knowledge, who benefits in a specific situation and the provider of knowledge, who contributes to the teams' experience. Accordingly, knowledge access and knowledge sharing are the core processes to support these roles. In the remainder of this section, we provide an analysis of common problems and possible areas of support.

Knowledge Access. Problems in knowledge access prevent developers from reusing past experience. Developers are either not aware that certain knowledge exists, or are not able to find it. Awareness is limited due to large amount of information, information dispersal across several sources, unknown information sources, or access restrictions. Knowledge access can be improved by capturing and using implicit context of knowledge needs. Based on this, search mechanisms are able to finding similar but for the given context relevant knowledge items. Proactive and personalized knowledge delivery, depending on a user's working and personal context can reduce the barrier to explicitly search for useful information.

Knowledge Sharing. Knowledge sharing is ineffective if existing knowledge cannot be accessed by people who require it. This occurs when knowledge is implicit in the "heads" of people, or information is "hidden" in private spaces. The main dilemma is that a knowledge provider has to spend extra effort without immediate reward. It is often not clear which information is worth to be shared, who requires it and what would be a proper level of generalization. The specification of concrete circumstances of a solution is problematic, since it is hard to determine the importance of different aspects. Since developers tend to be reluctant to provide extensive feedback on those issues, methods for implicitly capturing user's knowledge are required. Moreover, knowledge has to be associated to the developers' working context for establishing suitable similarity measures. This

enables to match shared knowledge with the situation of potential consumers. Accordingly, knowledge sharing benefits from an automatic elicitation of the knowledge through a semantic monitoring of developers interactions.

2.3 Enablers for Lightweight Knowledge Sharing

We identified five common enablers for a lightweight knowledge sharing support in distributed development teams.

Contextualization. In order to reuse previous experiences, the contextual gap between knowledge provider and knowledge consumer should be bridged. A lightweight knowledge-sharing environment should be able to identify context similarities, and map syntactically different but semantically similar contexts. By distilling knowledge in a predefined format, subjective and contextual aspects are lost. Consequently, a knowledge artifact should be represented in its native form as an interlinked piece of information. An important characteristic of development knowledge is the traceability of the context, where the knowledge arose from [12]. Problems that cause a knowledge need, such as build or test fails, should be captured explicitly. Knowledge support should be manifested as a knowledgeable action or source (e.g. searching the web, reading a document or writing a request in a forum) required for resolving a development problem. In the course of resolving a problem, developers often use a set of sources that represents valuable information. The result of knowledge support (i.e. changes in the given code) should be represented as an effect on development artifacts.

Personalization. Personalization is a main characteristic of an efficient knowledge support, especially in software engineering settings, where a large amount of development and communication artifacts exists. A recent survey [7] about knowledge support in software development shows that information overload is the main drawback that decreases the efficiency of a search system. Indeed, web search engines usually deliver lots of irrelevant results, which dissatisfies developers and consumes precious time.

Proactive Assistance. Knowledge sharing does not represent a core function for a team, such as testing, building, integrating or implementing. Because the economic added value of sharing and accessing knowledge is not immediate, developers tend to ignore or delay it. Therefore, an environment for lightweight knowledge sharing should trigger both knowledge sharing and access. Such environment should be able to identify knowledge need situations (such as successive building errors) as well as knowledge sharing situations, where developers have just had similar experience [13]. In both situations developers should be guided through an access resp. sharing process, by recommending knowledge resp. preparing knowledge to be shared. Awareness about who should benefit from such knowledge increases the motivation to share [14]. Such awareness is achieved in collocated teams by continuous informal communication and round-trip team buildings.

Decentralization. Decentralization facilitates informal communication and therefore knowledge sharing. A central knowledge base requires common processes and a "feeling" of authority control. An analogy can be found in a university domain. Students learn formally in lectures or informally from each other. While both approaches are important, in our case we aim at sharing informal, personal development experiences, where decentralization reduces sharing barriers. Sharing and accessing knowledge in a decentralized manner enables the association of knowledge to people (similar to sending an email on experiences with installing a component). Developers have a "total-control" on defining, who should access their knowledge and where they can get knowledge from. Furthermore, distributed development communities and especially open source ones strongly differ in terms of size, organization and geographical distribution. This puts a strong scalability requirement on a lightweight knowledge sharing approach. To exchange expertise, knowledge communities can be built around a single interface in an open source component. Installation and maintenance overhead in such situations is not acceptable. Moreover, if teams start to use their own central environments, crucial synchronization and access control problems may arise. In practice, these problems are ignored and artifacts are exchanged through emails, chat rooms or shared documents. Finally, organizing a knowledge sharing process in a decentralized manner enables the support of heterogeneity of used vocabularies, culture as well as physical and temporal distribution, which makes up the standard setting in distributed development.

Integration. As knowledge workers, software developers make most of their experiences in development environments. Capturing, accessing and sharing of knowledge should be integrated into the development environment to ensure a smooth knowledge exchange. Integration reduces context switches caused by changing a tool, a task or a thought in order to share knowledge by writing an email or editing a Wiki page. Integration increases developers' motivation for accessing and sharing knowledge [15].

3 Conceptual Framework

We aim at a knowledgeable software development environment, which enables efficient, personalized and proactive satisfaction of a software developer's knowledge needs, based on captured experiences from distributed team members. We describe how semantic technologies are used as a major enabler and introduce our architecture for lightweight knowledge sharing.

3.1 Semantic Technologies

Semantic technologies are defined as software technologies that allow the meaning of and associations among information to be known and processed at execution time. They are based on ontologies. An ontology is a formal, explicit specification of a domain of interest [16]. Semantic technologies help solving the

problems of application and data interoperability, improve search, discovery and content provisioning in knowledge-centric systems and facilitate the dynamic integration across distributed systems. Due to their formal but extendible nature, semantic technologies provide mechanisms to address previously mentioned problems. First, ontologies provide the proper level of abstraction required for an efficient structuring of knowledge with the ability of extending and changing underlying models. On the one hand, ontologies enable the required flexible representation of a knowledge artifact. Developers experiences can be captured without distillation and new concepts, such as new technologies, error types or knowledge sources can easily be added without a redesign of the underlying semantic models. Semantic information that describes developers' experiences (e.g. the error message "A" resulted from build action "B" using the Build infrastructure "C") can be therefore described or annotated using ontological concepts. On the other hand ontologies enable formal representation of relationship, in particular taxonomies that are required for semantic similarities of required and provided knowledge. For example, a hierarchical schema for classifying errors can be provided in order to enable finding similar situations for which an information support (i.e. experience) already exists. Second, ontologies provide a unified vocabulary that ensures unambiguous communication within heterogeneous community members. Such vocabularies help to integrate and map semantically equal but syntactically different information. Finally, based on their formal and machine-understandable nature, ontology-based representation of development experiences enables reasoning and therefore the externalization of implicit knowledge. For example, it can be deduced if the the underlying development platform of a current problem and possible available solutions is different or similar (e.g. both based on a UNIX Kernel). Reasoning based on logical axioms from development domain also enables the automatic bridging of contextual gaps between different developers' experiences.

3.2 Conceptual Architecture

We present the conceptual framework to enable a lightweight knowledge sharing among distributed developers. As depicted in Figure 1, the framework includes three logical layers: the Distributed Knowledge Model, the Context System and the Knowledge Desktop including the Knowledge Provider and Knowledge Capturer module, which we describe in the following.

Distributed Knowledge Model. This layer constitutes of a set of Ontologies, the Local Metadata Store and the P2P Infrastructure.

Ontologies. The knowledge model should include content, interaction and organizational information on development experiences. The ontologies describe the structure and content of software artifacts, related problems and their solutions, developers' interactions plus organization, infrastructure, roles and projects.

The most relevant ontological concepts are Development Artifacts, problem and knowledge sources. *Development Artifacts*: A relevant code fragment for a

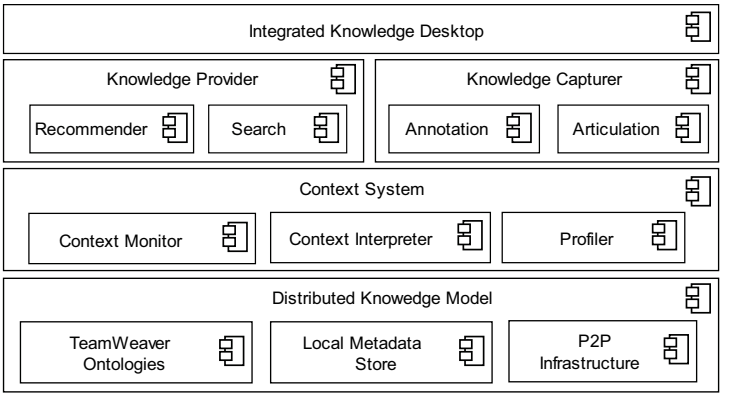


Fig. 1. Lightweight knowledge sharing framework

knowledge artifact is typically not equal to a compilation unit (e.g. a class in Java). Especially in the case of applying an external framework or using an external component, the problem is constituted by a series of source code lines. The code should be described and referenced in a semantic way. *Problem*: A problem is an obstacle, which makes it difficult to achieve a desired goal, objective or purpose. Typical problems to be considered are runtime errors ("stack traces"), failures of unit tests, build failures or a general inability to use the component because of a lack of examples. The ontological representation allows for bridging different levels of detail by classifying problems hierarchically. *Knowledge Source*: The type or content of a knowledge source will be described with a Domain Ontology. The Context Ontology also encompasses personal characteristics of the user, e.g. the level of competence in a certain domain area. This accounts for the fact that problem solving strategies (1) are in general different for different levels of competences (2) can sometimes only be applied if the user has a certain level of competence. Other characteristics can be made up of preferences, e.g. type of sources. Developers may prefer interactive sources like discussion forums while others prefer passive ones.

Local Metadata Store. The basic underlying infrastructure for all semantic services of the proposed framework is an ontology management system, which stores, allows queries to, and performs inferencing over ontology-based metadata describing all knowledge objects in the system, their interrelationships, and the background knowledge denoted in the defined ontologies.

P2P Infrastructure. P2P infrastructure allows a set of Knowledge Desktops and therefore the associated developers to access and provide knowledge sources on a semantic basis, defined on the common language for representing semantic information within the system.

Context System. This layer enables capturing a user's experience and its analysis in order to prepare it for further reuse. It observes a developer's

behavior in order to proactively determine situation in which she needs knowledge support. We treat experience as a context in which a problem is resolved, i.e. the cycle "develop-problem-source-develop-..." that leads to resolving a situation in which a developer requires knowledge support (e.g. a malfunctioning code). Moreover the context system will be able to refine captured interactions and further sensed events to semantically useful context information. For example, from the interaction events the context system describes what the developer has been doing for a period of time (e.g., localizing a bug, refactoring code, implementing authorisation mechanism). The context system includes three modules: Context Monitor, Context Interpreter and Profiler.

Context Monitor. The current content of a knowledge source that a user is reading, as well as the content he is writing is captured by this module. It also detects what the developer is trying to do based on the context of his current work. The contextual data will be derived from methods and classes being used by the code the developer is working on, as well as comments on it. The monitor includes a set of sensors to existing tools and information sources as well as a semantic representation of the sensed information in the ontologies.

Context Interpreter. Captured information should be interpreted according to the above given model of knowledge, i.e. identifying relevant development artifacts, identifying/classifying problems, and the usage of knowledge sources. In addition this module is responsible for the context sessionization, i.e. the determination of the start and the end of a developer session (i.e. a set of actions) that is related to resolving a problem. Session analysis generate two outputs:

- Experience as an instantiation of the knowledge model
- Working context as an input for search, including indicators that a user is in a situation in which she could require knowledge support

Profiler. Since personalization assumes defining general preferences, it is difficult to define general interests of a user regarding development knowledge. For example, if a developer reads several documents about "error X", it does not mean that she is generally interested in that topic. However, this information is useful if another developer is searching for developers that dealt in the past with the "error X". In order to cope with this issue, we introduce two types of personalization: *Local Profile*: This is a specific, quickly adapting, but less accurate profile. It represents information needs during the user's current interaction with the system. Working context can be considered as a local profile. Local profile is a short term one and is adapted after each interaction with the system, which means it has to cope with few and uncertain evidences as well as hard time constraints. *Global profile* It represents a global, long-term view on a user's experience and can be used in the P2P search for identification of "similar" developers. Two sources are used to create this profile and establish similarity among distributed developers:

- All information about a users behavior (e.g. which query he often posts, or in which situations he allows knowledge support) are captured in a log

file. Traditional analyses for frequent pattern discovery (e.g. association rule mining) find important characteristics of a user.
- The local code repository indicates the preferences of a user regarding the development task. By mining this repository, usage history for all the stored classes can be extracted. This can be done once initially for each class and subsequently when a class is added to the repository. Also, for each developer we store a list of components based on the actual usage order.

Knowledge Desktop. This layer represents the integrated interface between the developer and stored knowledge. It enables capturing and accessing knowledge.

Knowledge Capturer. Knowledge Capturer should be seamlessly integrated into a full-featured development environment. Furthermore, it can communicate with other software engineering tools such as a bug tracking system or a developer's mailing list. It includes three capturing modules: annotation, articulation and communication. *Annotation*: With this interface, developers should be able to annotate source code, and other development artifacts using the vocabularies defined in the ontologies. This is a similar tagging functionality to well-known Web 2.0 applications, where the tags represent ontological concepts. This enables the definition of semantic interrelations of knowledge artifacts, which is used by the knowledge provider. *Articulation*: This is an interface for entering knowledge by adding explicitly an instance of the knowledge model. Technically and regarding the user interface, this will offer to developers a Semantic Wiki-like articulation facility for expressing text-based knowledge they wants to share with others. The context system provides default content to the developers, who are then able to edit and update. Auto-completion functionality enables the use of ontological concepts for articulating knowledge enabling the required informal articulation and the required structure for semantic mapping. *Communication*: This module includes the possibility to start synchronous (such as instant messaging) or asynchronous communication between members of a distributed community, related to a specific problem situation. Communication artifacts are considered as further knowledge artifacts that can be used to describe developers' experience.

Knowledge Provider. Knowledge Provider consists of the semantic search and semantic recommendation. *Semantic Search*: This module implements context-sensitive and ontology-based proximity search for relevant knowledge items and represents the central interface for knowledge access. We distinguish between local and global search. Local search uses relevant knowledge within the Local Metadata Store, while global search locates relevant knowledge across the P2P network of distributed developers. In order to precisely specify developers' information need, a query is extended with context information, such as current used artifacts, nature of encountered problem, used platform as well as experience level. In the case that local search results are unsatisfactory, the query and found results will be transferred to the global search. This will be extended with information about preferences in order to enable result ranking

according to profiles of developers who respond to the request. Keyword-based search results in a lot of irrelevant results, such as code examples that developers should try. To deal with this issue, we use two extensions of a traditional keyword-based search: query refinement and query expansion. Query refinement is required in particular with developers who have less experience about the topic they are searching for. In distributed settings, developers are unfamiliar or not aware about knowledge available in different peers. In order to avoid failing queries, developers start searching with a short query and try to exploit the repository in several subsequent refinement steps. Query expansion extends the query with related terms from the common knowledge model, or from the set of top-ranked results. To resolve failing queries, the system automatically suggests the refinement of terms. The search result can be the contact data of an expert, who might be called up (information pull) or the information panel where the context-specific, semantic recommendations are presented to the user (information push). *Semantic Recommender*: This module enables proactive knowledge delivery depending on the actual working and personal context of the user. It supports resolving granular problems, such as instantiating an object, handling a specific error, or understanding a behavior. The semantic recommendation makes use of the context information collected by the monitor, e.g. components used in the current class. One recommendation situation occurs when a developer repeats the same error several times. By analyzing common errors developers made in the past, developers who are creating the same artifacts and getting the same error, can be warned how this error in previous situations has been solved. This is analogous to the peer reviews in Extreme Programming. The knowledge model introduced above supports these analyses, due to its focus on triplets code (problem/knowledge/source)- Another situation for proactive support is the time a developer is spending on writing a program block. It is possible that a developer has been "blocked" in a situation and that small hints will be useful. Contextual information as working context and user's preferences is used to determine implicit knowledge needs. For example, if a user opened a web site while programming, the proactive support takes the content of that document into account.

4 Related Work

Context-aware Environments. Several attempts have been taken toward monitoring developers interaction to achieve context awareness. Mylyn [7] addresses the problem of information overload faced by developers in a single development environment, by using a degree of interest model calculated from the previous modifications of a particular code elements. Mylyn allows a developer to share "task context", which helps to reproduce the working context in collaborative activities as bug fixing. Onmoronyia et al. extended this approach by introducing a "Continuum of Relevance" model that supports distributed developer in distinguishing relevant from irrelevant information [17]. The model orders the relevance of developers, tasks and artifacts by capturing interaction information amongst these entities over a project lifetime. Both approaches aim at reducing

the information overload problem faced by developers by adjusting the user interfaces. None of them targets to connect this data with problems and solutions experienced by developers during development activities.

Knowledge Infrastructures. Existing solutions for knowledge sharing in software engineering are mostly based on centralized approaches. The Experience Factory [18] for example imposes a knowledge extraction and refinement process, whose aim is to eliminate all subjective and contextual aspects of knowledge, and create an objective and general representation that can then be reused by other people in a variety of situations [18]. Besides that, the existence of a centralized knowledge repository is required, which has to be "filled" over time. For small-and middle-sized companies, as well as for open source communities, this is a non-realistic assumption. Lightweight knowledge sharing approaches such as Wikis are more popular, but also lack direct integration into the working processes and context of developers [19]. Recent works suggest using ontologies as a basis to leverage software engineering knowledge. The Dhruv system [20] aims to assist the bug resolution process by recommending relevant information during bug inspection. While nicely integrated into the working environment, Dhruv provides a central, server-based solution, which is specific to bug resolution. However, the underlying concept of semantic technologies is well-suited to support more decentralized knowledge sharing scenarios.

5 Conclusion and Future Work

While current approaches such as configuration management, software evolution management address coordination support related to artifact changes, we introduced a framework for enabling distributed teams to share knowledge in a lightweight manner. The key innovation is the combination of semantic technologies, context-awareness and P2P infrastructures, enabling a personalized, context-aware and a proactive knowledge acquisition and sharing. The introduced framework supports distributed teams to informally exchange experiences and thus better deal with the changes in requirements, technologies, architectures and organizations. This is one major step towards agile distributed teams. The TeamWeaver[1] platform represents a first open source implementation of this framework for Eclipse. We are currently evaluating our approach in several case-studies [21].

References

1. Paasivaara, M., Lassenius, C.: Could global software development benefit from agile methods? In: ICGSE 2006: Proceedings of the IEEE international conference on Global Software Engineering, Washington, DC, USA, pp. 109–113 (2006)
2. Ramesh, B., Cao, L., Mohan, K., Xu, P.: Can distributed software development be agile? Commun. ACM 49(10), 41–46 (2006)

[1] http://www.teamweaver.org/

3. Maalej, W., Panagiotou, D., Happel, H.J.: Towards effective management of software knowledge exploiting the semantic wiki paradigm. In: Herrmann, K., Brügge, B. (eds.) Software Engineering, GI. LNI, vol. 121, pp. 183–197 (2008)
4. Beck, K.: Manifesto for agile software development, `http://agilemanifesto.org`
5. Schwaber, K., Beedle, M.: Agile Software Development with Scrum (2001)
6. Williams, L., Kessler, R.: Pair Programming Illuminated (2002)
7. Kersten, M., Murphy, G.C.: Using task context to improve programmer productivity. In: Proceedings of the 14th ACM SIGSOFT international symposium on Foundations of software engineering, pp. 1–11. ACM, New York (2006)
8. Sommerville, I.: Construction by configuration: Challenges for software engineering research and practice. In: ASWEC 2008: Proceedings of the 19th Australian Conference on Software Engineering (aswec 2008), Washington, DC, USA, pp. 3–12. IEEE Computer Society, Los Alamitos (2008)
9. Griss, M.L.: Software reuse: Objects and frameworks are not enough. Technical Report HPL-95-03, Hewlett Packard Laboratories (January 1994)
10. Kauba, E.: Wiederverwendung als gesamtkonzept - organisation, methoden, werkzeuge. OBJEKTSpektrom 1, 20–27 (1996)
11. Hansen, M.T.: The search-transfer problem: The role of weak ties in sharing knowledge across organization subunits. Adm. Sci. Quarterly 44, 82–111 (1999)
12. Mohan, K., Ramesh, B.: Traceability-based knowledge integration in group decision and negotiation activities. Decision Support Systems 43(3), 968–989 (2007)
13. Happel, H.J., Maalej, W.: Potentials and challenges of recommendation systems for software development. In: Proceedings of the International Workshop on Recommendation Systems for Software Engineering. ACM Press, New York (2008)
14. Happel, H.J.: Closing information gaps with inverse search. In: Yamaguchi, T. (ed.) PAKM 2008. LNCS (LNAI), vol. 5345, pp. 74–85. Springer, Heidelberg (2008)
15. Zeller, A.: The future of programming environments: Integration, synergy, and assistance. In: FOSE 2007: 2007 Future of Software Engineering, Washington, DC, USA, pp. 316–325. IEEE Computer Society Press, Los Alamitos (2007)
16. Gruber, T.R.: A translation approach to portable ontology specifications. Knowledge Acquisition 5(2), 199–220 (1993)
17. Omoronyia, I., Ferguson, J., Roper, M., Wood, M.: A 3-dimensional relevance model for collaborative software engineering spaces. In: ICGSE 2007: Proceedings of the International Conference on Global Software Engineering, Washington, DC, USA, pp. 204–216 (2007)
18. Basili, V.R., Caldiera, G., Rombach, H.D.: Experience Factory. In: Marciniak, J.J. (ed.) Encyclopedia of Software Engineering, pp. 469–476. John Wiley & Sons, Chichester (1994)
19. Chau, T., Maurer, F.: A case study of wiki-based experience repository at a medium-sized software company. In: K-CAP 2005: Proceedings of the 3rd international conference on Knowledge capture, New York, USA (2005)
20. Ankolekar, A., Sycara, K., Herbsleb, J., Kraut, R., Welty, C.: Supporting online problem-solving communities with the semantic web. In: WWW 2006. ACM, New York (2006)
21. Bagnato, A., Maalej, W.: From research to practice: How to txt e-solutions plan to deploy innovations in sharing development knowledge. In: Software Engineering (Workshops). LNI, vol. 122, pp. 43–50 (2008)

Collaboration-Oriented Knowledge Management Using Interaction Patterns*

Ulrich Reimer, Uwe Heck, and Stephan Streit

Institute for Information and Process Management
University of Applied Sciences St. Gallen
Teufener Str. 2
CH-9000 St. Gallen, Switzerland
firstname.lastname@fhsg.ch

Abstract. The paper starts with analysing the requirements of supporting largely unstructured, collaboration-oriented processes as opposed to highly structured, workflow-based processes. The approach derived from those requirements utilizes the concept of interaction patterns to impose some control and guidance on a collaboration process without reducing the needed flexibility. This not only causes a much greater process transparency but also allows to provide and capture knowledge within the collaboration interactions. The application-specific instantiation of the interaction patterns and their enrichment with integrity constraints is done using a model-driven approach. The resulting application model is also a knowledge artefact which represents an important fragment of domain knowledge.

1 Introduction

A classical distinction in knowledge management is the *codification* versus *personalisation* approach [12]. Codification of knowledge facilitates its reuse while personalisation facilitates the creation of new knowledge and innovation in general. In practice usually a mixture of both approaches is needed but the distinction helps to analyse the specific needs within an organisation and to design the appropriate solutions. Most IT-based knowledge management solutions fall within the codification approach, while several IT solutions to facilitate personalisation have emerged only more recently under the paradigm of social software [17].

Another very important and prominent knowledge management approach *integrates process management with knowledge management*. It is based on the idea that it is within the (core) processes of an organisation where knowledge is mainly utilized and created. Several systems have been developed that integrate knowledge provision and knowledge capturing with the corresponding activities in a business process [1, 2, 16, 21]. Process-oriented knowledge management is strongly related to the codification approach and presupposes a sufficiently detailed process description.

* The research presented in this paper was funded by the CTI Swiss Confederation innovation promotion agency under grant number 8617.1.

T. Yamaguchi (Ed.): PAKM 2008, LNAI 5345, pp. 26–37, 2008.

However, there also exist many kinds of applications with only *weakly structured processes*, where each process instance can be quite different in structure without strong rules to predict a-priori what it will look like. Existing approaches to support weakly structured processes either focus on giving ad hoc guidance through a process [13] or allow to handle exceptions or changes during the execution of a predefined workflow [18, 20].

All these approaches still presume a certain amount of process structure that is no longer present when we deal with application scenarios where the interaction between people working together is more a *collaboration* than a workflow-like sequence of predefined activities. Examples of such application areas are governmental processes that involve human decision making, innovation processes, or any other scenario where people need to interact in a not predefined way to obtain a certain result.

The application scenario we came across in a current project on process-oriented, cross-organisational e-government is exactly of the above characterized nature. A high-level process structure occurs together with rather unstructured collaboration activities. The solution to provide a standard collaboration platform where people can (just) exchange documents turned out to be not preferable because of the lack of any process guidance. Instead we decided to facilitate the collaboration process by supporting the high-level process structure and by imposing a micro-level structure in terms of interaction patterns. This approach allows to ensure compliance with regulations and to actively provide the information needed in the collaboration without marring the required flexibility.

Thus, the main contribution of this paper lies in extending the ideas underlying process-oriented knowledge management to application scenarios that are collaboration-oriented. While knowledge management approaches to support collaboration have so far been of the personalisation paradigm our approach makes codification-based approaches applicable to collaboration-based knowledge management.

The remainder of this paper is organised as follows. In Sec.2 we introduce two application scenarios which motivated our approach. Sec.3 presents the concept of interaction patterns as the core element of our solution. Sec.4 describes how this approach can be leveraged using a model-driven approach, while Sec.5 concludes and gives an outlook to future research.

2 Application Scenarios

In this section we discuss two application scenarios which motivated our research and led to the solution presented in Sec.3. The e-government scenario in Sec.2.1 was the main driver and will therefore be used throughout this paper to illustrate our ideas.

2.1 Cross-Organisational e-Government

E-government has two major perspectives: to achieve greater operational efficiency and to offer improved public services. Typical measures undertaken in the former case are the re-engineering of business processes and the move from paper to electronic documents. The typical measure in the latter case is to move citizen services online and make them accessible through a web portal. More recent e-government approaches address a more process-oriented view [3, 15, 19]. These approaches are justified by the following observations:

- Government agencies often do not work independently but need to collaborate, e.g. in order to offer a one-stop-shop portal to citizens.
- Citizen-oriented services often require more than a single stop at a web portal but are rather like a process being carried out, e.g. the submission of a tax declaration which not only includes the submission itself but also the delivery of the tax assessment and maybe further enquiries and requests for additional documents. Depending on the participants in such an e-government process we not only get G2C, G2B and G2G processes but also any combinations.

The need for process-oriented, cross-organisational e-government was the starting point of the HERA (Helvetic E-Government Reference Architecture) project which began in 2007 with the aim of:

- providing a seamless integration of cross-organisational e-government processes into a comprehensive end-to-end process,
- sharing information between process participants while ensuring that the involved parties only see exactly that information they are qualified to see,
- reducing the effort required to implement and maintain cross-organisational e-government applications by adopting a model-driven approach.

As an example of such cross-organisational processes we focus in the HERA project on the declaration of company taxes on profits. The process starts with a company or a trustee which prepares the annual financial statements, then involves an auditor and finally the tax office (or several tax offices if the company has subsidiaries in different cantons). In numerous workshops the following main characteristics of the tax declaration process have been determined:

- The process is clearly divided into different (cross-organisational) sub-processes, each with a different person being responsible and different persons, resp. organisations being involved.
- The transition from one sub-process to the next currently causes a media break because all relevant documents are transferred in paper format, although within the sub-processes the documents are usually handled electronically.
- The process duration can be quite long (several months to more than a year), causing a lack of transparency with respect to the process status, its history, current responsibilities and tasks.
- The interactions within the sub-processes are goal-driven and therefore do not follow a fixed workflow. They can rather be seen as a collaboration between the process participants.

The above findings can be generalized to any cross-organisational e-government process which involves human interaction and decision making. Existing approaches to cross-organisational e-government are either based on a workflow paradigm [3, 15] or employ an automatically orchestrated interplay of services without any human interaction [10] and can therefore not be applied for the above mentioned kinds of processes. Neither can approaches to support weakly-structured or dynamic workflows [13, 18, 20] be applied because they still presume a detailed (albeit varying) process structure while in our case we have nearly no structure at all.

Our characterisation is corroborated by similar findings mentioned in [5, 8, 9]. In [5] it is argued that imposing a rigid structure on e-government processes will result "in IT systems that do not support the situated nature of work".

We suggest to facilitate cross-organisational e-government processes as well as any other processes with similar characteristics (see Sec.2.2) by giving guidance and control for the

- *transitions between sub-processes*, taking the conditions under which they are permitted and the information to be transferred along each transition into account;
- collaboration taking place within each sub-process by *imposing a micro-structure in terms of interaction patterns* while still preserving the required flexibility.

Sec.3 will present our solution in more detail. In the subsequent section we will shortly argument that supporting innovation processes leads to similar requirements than for the e-government processes.

2.2 Innovation Processes

Innovation requires the leveraging of the implicit knowledge of people. The discussion and exchange of ideas is essential so that the personalisation approach [12] is best suited to facilitate innovation processes. Although utilization of implicit knowledge and brainstorming are most import, explicit knowledge is of course also needed.

Innovation processes are not just unstructured brainstorming and collaboration activities but have a clear overall structure. In this respect they are quite similar to the cross-organisational e-government processes discussed in Sec.2.1. Thus, it is to be expected that our solution to support those processes will also be applicable to innovation processes. Moreover, as innovation processes also have a long duration (weeks to months) and involve many participants we predict that their support by an appropriate knowledge management system has the same benefits as for the e-government

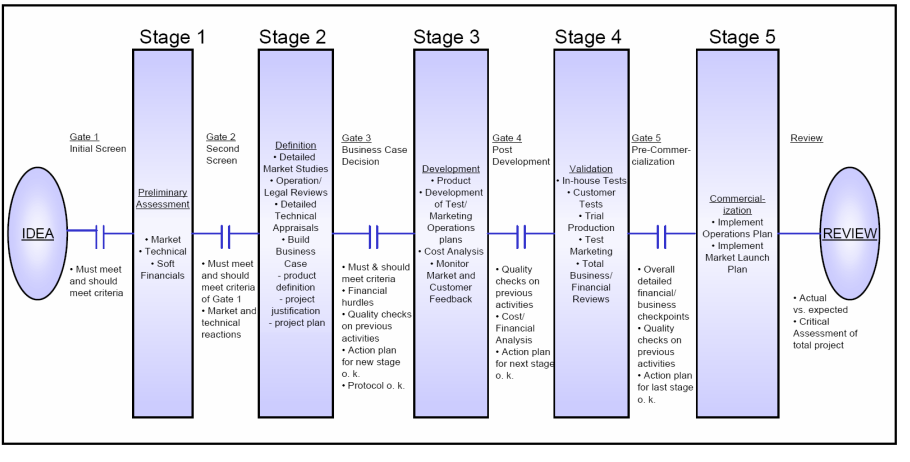


Fig. 1. Example of an innovation process structure (from [7])

processes, namely, among others, process transparency, elimination of duplicate efforts, and greater overall efficiency. Existing research on innovation processes supports this claim.

There exist several models for innovation processes and all of them propose a staged process. An example, taken from [7], is shown in Figure 1. Each stage corresponds to a sub-process as discussed in Sec.2.1, while the activities within each stage form a collaboration, which would benefit from some organizing principle in order to direct and focus the collaboration – just as shown in Sec.2.1. Another, more recent research comes to similar conclusions [4]. The authors show that inter-organisational innovation goes through several consecutive stages, each consisting of several interaction patterns. While their notion of an interaction pattern is more general than ours, purely informal, and not meant to be the basis of some support system the closeness to what we suggest is quite striking.

Moreover, recent innovation models like Open Innovation [6] which further extend the participants involved in the innovation process, e.g. to customers, are more complex and thus require even more process and collaboration support.

3 Collaboration Support Based on Interaction Patterns

Here we describe our approach to support collaboration processes. To simplify the discussion we keep to the e-government application scenario introduced in Sec.2.1.

For each sub-process in the considered application scenario (cf. Fig.2) all permitted kinds of collaboration interactions are predefined. An interaction activity typically consists of several communication steps between an arbitrary number of process

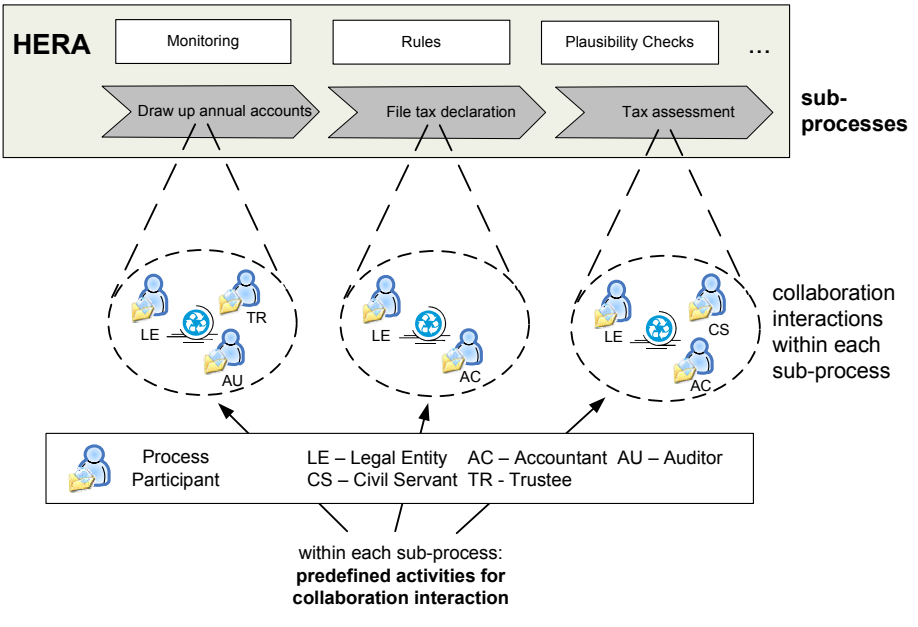


Fig. 2. The HERA platform for supporting the process of company tax declaration

participants, e.g. person A sending an information request to person B, and person B answering that request. Person B might also forward the request to person C which then sends the answer to person A. Moreover, each interaction activity is associated with a set of *integrity constraints* that specify which kinds of process participants (in terms of the role they have) can execute that activity under which preconditions (ensuring compliance to law and regulations). By checking these constraints the system can make sure that only meaningful activities are initiated. For each activity it is also specified which *kind of information* must be sent along to the recipient and what *background knowledge* might be helpful for the sender or recipient in the current context. Moreover, the system can keep track of still unfulfilled requests and can send reminders. Since an e-government or innovation process may last for months this bookkeeping and reminder facility is extremely helpful.

A closer inspection of the collaboration interactions shows that they can all be seen as instantiations from a small set of *generic interaction patterns* (see Table 1). Each pattern defines a prototypical communication structure between two or more participants. For example, the activity of an accountant sending the tax declaration to the tax office and thus initiating the transition to the next sub-process "tax assessment" is an instantiation of the pattern `sendInfoForProcessing`. The activity of an accountant requesting the correction of a book entry from the trustee and receiving from him/her the correct booking is an instantiation of the pattern `requestInformation`.

Table 1. Generic interaction patterns

Pattern Name	Description
`sendInfoForProcessing`	- sender must have sub-process ownership - information is sent to the receiver - sub-process ownership is passed to the receiver - sub-process ownership ends for sender
`sendInfoForNotice`	- information is sent to the receivers
`sendInfoForConfirmation`	- information is sent to receiver for confirmation - receiver must answer the request with either a confirm, reject, or abstain - pending requests may be deleted by the sender
`requestInformation`	- information request is sent to receiver - receiver may pass on the information request, or - receiver must send the requested information back to the sender - pending requests may be deleted by the sender
`publishInformation`	- information related to a certain topic is published - participants can subscribe to topics - from the publisher's point of view, an anonymous set of receivers (resp. subscribers) can consume the information

It can be meaningful to introduce application-specific specialisations of those patterns. For example, the pattern `sendInfoForConfirmation` can be specialised into a pattern that requires the confirmation in the form of a signature of the sent document. Then this specialised pattern is instantiated whenever a signature is needed for the confirmation of a specific information.

Figure 2 illustrates how the collaboration support based on interaction patterns results in a corresponding support system. Shown is the application scenario for the declaration of company taxes. At the core is the web-based HERA platform. It guides through the process by going from sub-process to sub-process and by offering the participants within each sub-process only those interaction activities that are permitted for him/her in the current context. All documents are handled electronically – either as (filled-in) structured forms or as PDF documents. HERA keeps track of each process by maintaining a protocol when which documents have been transferred to whom to carry out which task. Based on the given integrity constraints, HERA checks the documents as well as the whole records file for completeness and consistency.

According to workshops with future users the platform offers the following benefits:

- *Process transparency:* Process participants are always informed of the current process status.
- *Faster communication:* Due to the avoidance of media breaks and the instantaneous sending of documents process duration becomes much shorter. For example, the tax declaration can be submitted electronically directly into the IT system of the tax office.
- *Complete records files:* All the documents involved in the process are in electronic form so that the completeness of a records file is always guaranteed (as opposed to a paper file).
- *Correct processes:* Integrity constraints ensure (to a certain degree) a correct process. For example, incomplete or wrongly filled-in tax declaration forms cannot be submitted.
- *Reduced manual effort and elimination of errors:* This results from the avoidance of media breaks.
- *Eliminating duplicate efforts:* Supporting all aspects of the tax declaration process within one process and one system (such as requests for deadline extension, requests for additional documents, declaration of withholding tax) eliminates otherwise occurring duplicate efforts.
- *Integration with related e-government processes:* By accommodating several e-government processes on one or a set of federated HERA platforms, the benefits multiply (such as combining corporate tax declaration with value-added tax declaration, with the registration of addresses, etc.).

4 A Model-Driven Approach for Generating New Applications

Sec.3 presented our solution to supporting collaboration-based processes in terms of knowledge exchange and proper sequencing of activities. As we have discussed in Sec.2, there is a wide variety of application areas where the application of this

approach brings significant benefits. However, these benefits can only be realised if the corresponding support systems are not too expensive to set up. To this end, we adopt a model-driven approach: The main characteristics of our approach, which are common to all application areas, are modelled in a *meta model* [11, 14]. Various kinds of application areas such as cross-organisational e-government or innovation processes can be accounted for by appropriate specialisations of that meta model (see Fig.3). A specific application is then derived from the corresponding meta model by appropriately instantiating it (see again Fig.3).

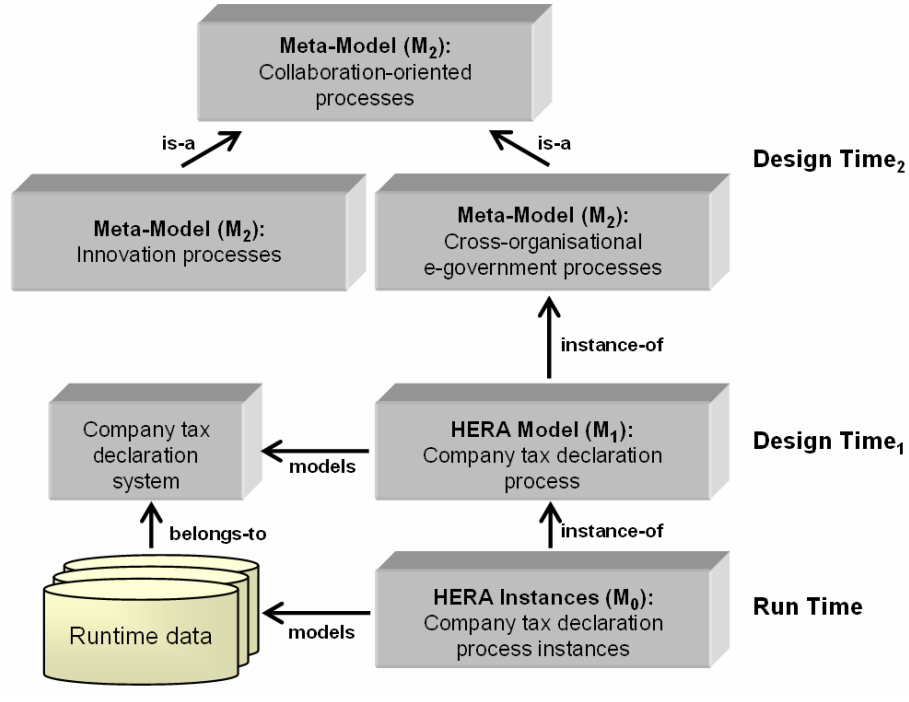


Fig. 3. Illustration of the model-driven approach

The main elements of the meta model (M_2) are:

- *Generic interaction patterns:* The application-specific collaboration interactions are instantiated from them.
- *Sub-processes:* The meta model provides the model element of a process that can be subdivided into sub-processes. The meta model leaves it open if the sub-processes are ordered linearly or partially. Each sub-process stands for a goal-driven collaboration between certain process participants. The sub-process is complete when that goal is achieved.
- *Process participants, their roles and their affiliations to organisations:* The generic organisational model can be instantiated as needed for the specific application.

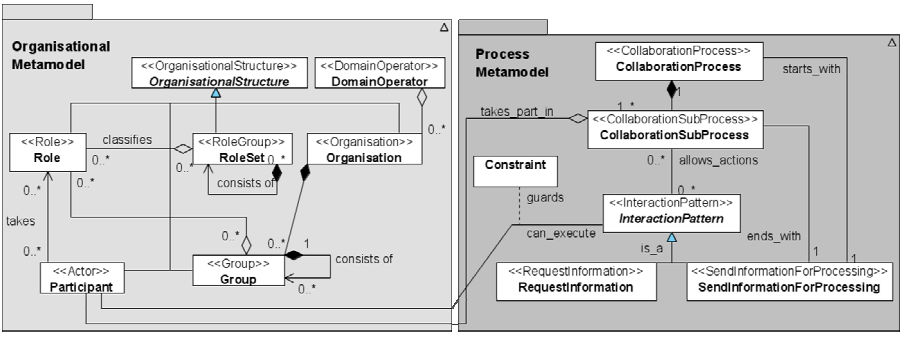


Fig. 4. A fragment of the meta model

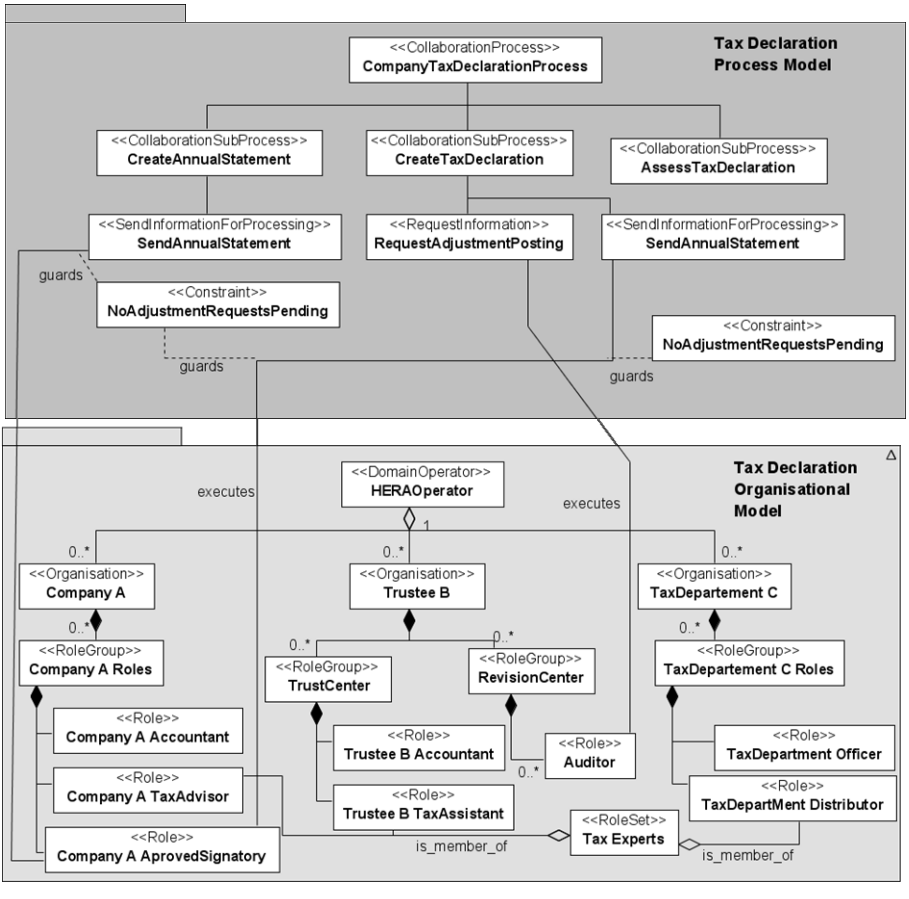


Fig. 5. Instantiation of the meta model in Fig.4 into an application-specific model for the tax declaration process

- *Information entities:* A classification of the information requested or to be provided in the interaction activities is given (in terms of a simple ontology).
- *Background knowledge:* A classification of background knowledge that may be identified as relevant in the context of an interaction activity. This may e.g. be some regulations to be taken account of, similar former cases, etc.

A model editor allows to generate and edit application models (M_1) as instantiations of one of the given meta models. A subsequent generation step maps the application model elements into the corresponding data structures of the underlying runtime architecture. Fig.4 shows a simplified fragment of the meta model for the e-government application scenario while Fig.5 illustrates its instantiation to the tax declaration process.

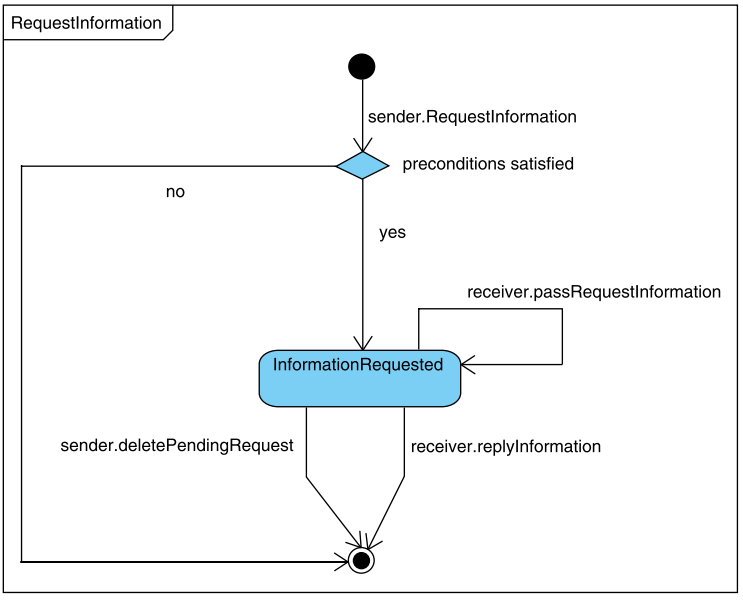


Fig. 6. Finite state machine for the interaction pattern `RequestInformation`

A model-driven approach only works when the elements of the meta model are not purely syntactic structures but have a formal semantics assigned to them. Here we illustrate only how the semantics of the interaction patterns – the core element in our approach – is specified. Each interaction pattern is described by a finite state machine. Other formalisms, like petri nets or the pi calculus, could have been used instead but the required mapping to elements of a process engine makes finite state machines a well suited, yet simpler formalism.

For each pattern a finite state machine defines the possible state transitions. The nodes of the finite state machine represent the states while the links are associated with the user activities. Fig.6 shows the finite state machine for the pattern `RequestInformation`. The pattern is invoked when a process participant initiates an activity which is an instantiation of that pattern. It is then checked if all conditions necessary to execute the activity are satisfied. If this is the case the request will be

delivered to the receiver, which can decide to either reply with the requested information or forward the request to another process participant. When the receiver replies to the request or the sender decides to cancel the request the pattern terminates.

5 Conclusions

The paper started with an analysis of application scenarios that involve largely unstructured, collaboration-oriented processes. Approaches of process-oriented knowledge management developed for supporting structured, workflow-based processes cannot be applied to such collaboration processes because of the lack of process structure. Based on the concept of interaction patterns we introduced a new approach that allows to impose some control and guidance on a collaboration process without reducing the needed flexibility. This allows to provide and capture knowledge within the collaboration interactions and results in a much greater process transparency.

The application-specific instantiation of the interaction patterns and their enrichment with integrity constraints is done using a model-driven approach. The resulting application model is a *knowledge artefact* in itself and represents an important fragment of the domain knowledge.

We have motivated our research mainly with an e-government scenario and a discussion of innovation processes. Although there is strong evidence that our approach is also applicable to innovation processes this still needs to be proved. This will be a subject of our future research. Currently, we are implementing the model-driven approach, which will lay the basis to explore further application scenarios more easily.

References

1. Abecker, A., Bernardi, A., Hinkelmann, K., Kühn, O., Sintek, M.: Context-Aware, Proactive Delivery of Task-Specific Knowledge: The KnowMore Project. Int. Journal on Information Systems Frontiers 2(3/4), 139–162 (2000) (Special Issue on Knowledge Management and Organizational Memory)
2. Abecker, A., Papavassiliou, G., Ntioudis, S., Mentzas, G., Müller, S.: Methods and Tools for Business-Process Oriented Knowledge Management – Experiences from Three Case Studies. In: Weber, F., Pawar, K.S., Thoben, K.-D. (eds.) Proc. 9th Int. Conf. on Concurrent Enterprising, Espoo, Finland, pp. 245–254 (2003)
3. Beer, D., Höhne, S., Petersohn, H., Pönitzsch, T., Rünger, G., Voigt, M.: Designing a Distributed Workflow System for E-Government. In: Proc. 24th IASTED Int. Conf. on Modelling, Identification, and Control, pp. 583–588. ACTA Press, Austria (2005)
4. Bossink, B.A.G.: The Interorganizational Innovation Processes of Sustainable Building: A Dutch Case of Joint Building Innovation in Sustainability. Building and Environment 42, 4086–4092 (2007)
5. Cajander, A., Eriksson, E.: Automation and E-government Services – A Widened Perspective. In: Winckler, M., Noirhomme-Fraiture, M. (eds.) Proc. 1st Int. Workshop on Design & Evaluation of e-Government Applications and Services (2007)
6. Chesbrough, H.W.: The Era of Open Innovation. MIT Sloan Management Review 44(3), 35–41 (2003)

7. Cooper, R.G., Kleinschmidt, E.J.: New Products: The Key Factors in Success. American Marketing Association, Chicago (1990)
8. Dustdar, S.: Caramba – A Process-Aware Collaboration System Supporting Ad Hoc and Collaborative Processes in Virtual Teams. Distributed and Parallel Databases 15(1), 45–66 (2004)
9. Feldkamp, D., Hinkelmann, K., Thönssen, B.: KISS: Knowledge-Intensive Service Support: An Approach for Agile Process Management. In: Paschke, A., Biletskiy, Y. (eds.) RuleML 2007. LNCS, vol. 4824, pp. 25–38. Springer, Heidelberg (2007)
10. Gugliotta, A., Cabral, L., Domingue, J., Roberto, V., Rowlatt, M., Davies, R.: A Semantic Web Service-based Architecture for the Interoperability of E-government Services. In: Proc. Workshop on Web Information Systems Modeling (WISM 2005) in conjunction with the 5th International Conference on Web Engineering (ICWE 2005), Sydney, Australia (2005)
11. Guttman, M., Parodi, J.: Real-Life MDA: Solving Business Problems with Model Driven Architecture. (The OMG Press). Morgan Kaufmann, San Francisco (2006)
12. Hansen, M.T., Nohria, N., Tierney, T.: What's your Strategy for Managing Knowledge? Harvard Business Review, 106–116 (March–April 1999)
13. Holz, H., Rostanin, O., Dengel, A., Suzuki, T., Maeda, K., Kanasaki, K.: Task-Based Process Know-how Reuse and Proactive Information Delivery in TaskNavigator. In: Proc. ACM CIKM Int. Conf. on Information and Knowledge Management, pp. 522–531. ACM Press, New York (2006)
14. Karagiannis, D., Kühn, H.: Metamodelling Platforms. In: Proc. Third Int. Conf. on E-Commerce and Web Technologies 2002, p. 182. Springer, Heidelberg (2002)
15. Punia, D.K., Saxena, K.B.C.: Managing Inter-organisational Workflows in e-Government Services. In: Proc. 6th Int. Conf. on Electronic Commerce, pp. 500–505. ACM Press, New York (2004)
16. Reimer, U., Margelisch, A., Staudt, M.: EULE: A Knowledge-Based System to Support Business Processes. Knowledge-Based Systems 13(5), 261–269 (2000)
17. Schaffert, S.: Semantic Social Software – Semantically Enabled Social Software or Socially Enabled Semantic Web? In: Sure, Y., Schaffert, S. (eds.) Semantics 2006: From Visions to Applications (2006)
18. Sheth, A., Han, Y., Bussler, C.: A Taxonomy of Adaptive Workflow Management. In: Proc. Workshop of the 1998 ACM Conf. on CSCW: Towards Adaptive Workflow Systems. ACM Press, New York (1998)
19. Tambouris, E., Gorilas, S., Kavadias, G., Apostolou, D., Abecker, A., Stojanovic, L., Mentzas, G.: Ontology-Enabled e-Gov Service Configuration: An Overview of the OntoGov Project. In: Wimmer, M. (ed.) KMGov 2004. LNCS (LNAI), vol. 3035, pp. 122–127. Springer, Heidelberg (2004)
20. Weske, M.: Flexible Modeling and Execution of Workflow Activities. In: Proc. 31st Annual Hawaii Int. Conf. on System Sciences, vol. 7, pp. 713–722. IEEE Computer Society, Los Alamitos (1998)
21. Woitsch, R., Karagiannis, D.: Process-Oriented Knowledge Management Systems Based on KM-Services: The PROMOTE Approach. In: Karagiannis, D., Reimer, U. (eds.) PAKM 2002. LNCS (LNAI), vol. 2569, pp. 398–412. Springer, Heidelberg (2002)

The Right Expert at the Right Time and Place
From Expertise Identification to Expertise Selection

Pavel Serdyukov[1], Ling Feng[2], Arthur van Bunningen[3], Sander Evers[1], Harold van Heerde[1], Peter Apers[1], Maarten Fokkinga[1], and Djoerd Hiemstra[1]

[1] Database Group, University of Twente, Enschede, The Netherlands
{serdyukovpv,everss,h.j.w.vanheerde,apers,
fokkinga,hiemstra}@cs.utwente.nl
[2] Database group, Dept. of Computer Science and Tech, Tsinghua University, China
fengling@tsinghua.edu.cn
[3] Teezir, The Netherlands
arthur@vanbunningen.com

Abstract. We propose a unified and complete solution for expert finding in organizations, including not only expertise *identification*, but also expertise *selection* functionality. The latter two include the use of implicit and explicit *preferences* of users on meeting each other, as well as *localization* and *planning* as important auxiliary processes. We also propose a solution for *privacy protection*, which is urgently required in view of the huge amount of privacy sensitive data involved. Various parts are elaborated elsewhere, and we look forward to a realization and usage of the proposed system as a whole.

1 Introduction

Expertise sharing is gaining increasing popularity and importance for enterprises due to the fact that a mass of knowledge has been accumulated in the course of corporate business, and meanwhile employees tend to seek and interact with knowledgeable people for information prior to using some formal sources to solve their daily work related problems. It is even common that users often search for persons rather than for relevant documents [15]. Besides being sources of unpublished knowledge, the experts on the search topic are also able to explain problems and solutions by guiding the user through existing artifacts. However, attempts to identify experts by manual browsing through organizational documents or via informal social connections are impractical in large enterprises, especially when they are geographically distributed. Usually, a specialized *expert finding system* (also known as expert search, expert recommendation or expertise location system) is developed to assist in the search for individuals or departments that possess certain knowledge and skills within the enterprise and outside [26]. It allows either to save time and money on hiring a consultant when company's own human resources are sufficient, or helps to find an expert at affordable cost and convenient location in another organization.

T. Yamaguchi (Ed.): PAKM 2008, LNAI 5345, pp. 38–49, 2008.

Finding an expert is a challenging task, because expertise is a loosely defined and not a formalized notion. It is common to refer to expertise as to "tacit knowledge" [8], the type of knowledge that people carry in their minds and is, therefore, difficult to access. It is opposed to "explicit knowledge" which is easy to capture, describe, document and store. Thus, an expert finding system aims to manage "tacit knowledge" in organizations by inferring it using organizational "explicit knowledge" and finally to transfer it among people by helping their socialization and knowledge exchange. With respect to these missions, it is common to divide the task of expert finding into two stages of equal importance: *expertise identification* and *expertise selection* [27].

At the *expertise identification* stage, all employees of the organization are ranked with respect to the user information need for expertise, usually expressed in *a short text query*. It is often unclear what amount of personal knowledge should be considered enough to name somebody "an expert". It depends not only on the specificity of the user query, but also on characteristics of respective expertise area: on its age, depth and complexity. However, expert finding systems do not actually infer the level of expertise or any quantitative estimate which is easy to semantically interpret or map to a qualitative scale. They just provide some estimate that may be used to rank people by their expertise level.

Since arranging a meeting with even a single expert might be very time-consuming, a practically usable expert finder should help not only to identify knowledgeable people, but also to *select those experts* that are most appropriate for a face-to-face contact with the user [1]. Since expert finding is a tool for improving organizational communication, it must be able to predict various features of a planned communication in order to help it be successful. In the first place, it should aim for a communication that is physically doable. So, the availability and interruptability of experts that may depend on their location and/or occupancy should be considered. In other cases, an intelligent meeting planning, taking into account agenda records of several employees, including the expertise seeker, as well as predictions for their future location, is required. An expert finder should also try to predict whether the communication is likely to be desired by both parts. Various human factors like expert's mood or mental stress may be considered. Preferences of users on communication with certain people (e.g., based on their positions/ranks or reputation in a company) should also be integrated.

Introducing expert finding in an enterprise inevitably results in an environment in which highly precise data about the whereabouts of employees, their behavior, preferences and the way they spent their time will be collected. Although any goal of increasing work efficiency might be of interest for both employers and employees, and employers do have the right to monitor and collect data from their employees, such goal can only be achieved when the right of privacy is not violated in the process [28]. Unlimited and unrestricted collecting of the private data, for example by monitoring the browsing behavior of employees, will infringe the privacy of the employee, with undesirable effects. Hence, when monitoring and collecting data is inevitable in order to enable services like expert finding, a clear corporate privacy policy is needed.

Our Contribution. The expertise identification task developed a lot during last years as a subject of research on Information Retrieval [14]. Recently, we proposed several solutions that were successfully evaluated within widely accepted experimental frameworks [31,32,33,34]. Moreover, the quality of existing research prototypes is currently quite high with an error rate about 50% [35], what motivates to expand the scope of research in expert finding to a broader spectrum of related vital issues. Consequently, this paper mainly seeks to *research on expertise selection.* To the best of our knowledge, this integral stage of expert finding is traditionally neglected in contemporary approaches and practically no academic research has been conducted in this direction, even at the conceptual level. We propose a unified and complete solution for expert finding in organizations, including not only *expertise identification*, but also *expertise selection* functionality. We present methods that solve several identified problems and have recently shown themselves to advantage in the respective problem domains; these methods will hopefully ease the development of a fully fledged expertise sharing system.

Organization of the Paper. The next section gives a brief overview of existing research in expert finding. The follow-up sections propose new solutions for each of the problems that a real-world expert finding system with a complete functionality inevitably faces. Section 3 describes how to make a first step from expertise identification to expertise selection and consider a model of user preferences on meeting certain people in the organization. Section 4.1 continues to explain how to efficiently facilitate expertise selection and describes methods for measuring and monitoring experts' availability (i.e., localization) considering that all users have the implicit preference that their question is answered immediately. Section 4.2 shows how to integrate another implicit desire of most users — to meet experts eventually in the future if the question is left unanswered. Section 5 proposes a unified method for ensuring privacy based on the fact that the proposed additional expertise selection stage needs a lot of private user data for the analysis. Section 6 concludes the paper with a summary of our contribution.

2 Existing Solutions

Expertise identification. In early expert finding systems the prediction of personal expertise was often made through the analysis of textual content of employee profiles. These profiles contained summaries of personal knowledge and skills, affiliations, education and interests, as well as contact information [16,17,9]. However, such profiles are always known to be incomplete and outdated due to serious time investment needed for their maintenance [10]. Therefore, the majority of successor systems, numerously emerged in academia during recent years, regarded any documents the person is related to as possible indicators of his/her expertise. They commonly assumed that the more often a person is related in the documents containing many words describing the topic, the more likely we may rely on such a person as on an expert.

Existing approaches to expertise identification naturally fall into two categories. *Profile-centric* approaches [23,29] merge all documents related to a candidate expert into a single personal profile either prior to retrieval time, or dynamically using only the top retrieved documents [31]. These personal profiles are then regarded as single documents to be ranked using standard measures of document relevance. *Document-centric* approaches first rank documents and then sum their retrieval scores for each related candidate to estimate the degree of candidate's expertise [25,7,19]. It was also proposed to calculate only the relevance score of the text window surrounding the person's mentioning [24] and to propagate relevance from documents to their related candidates not in one-step, but in several steps through utilizing graph structure of the respective expertise domain [34].

Expertise selection. Expertise identification methods mostly develop due to the interest in the academic world; in contrast, expertise selection research is making its marginal progress only due to the existence of industrial expert finding solutions. However, even their assistance in expertise selection is not all-embracing. Some of these systems offer powerful ways to represent and manually navigate search results, what, to a certain extent, simplifies expertise selection. Autonomy (`autonomy.com`), the undoubted market leader, allows the classification of experts in the result list by competency areas and positions in a company. So does the Endeca (`endeca.com`), the third enterprise search market leader after FAST (`fastsearch.com`). In some cases, the searcher's context is not totally ignored and implicit preferences of the user on types of people are considered: Microsoft's Knowledge Network recommends those experts who are found in proximity of the user in organizational social network. Workload aspects are considered by AskMe (`askmecorp.com`) that develops an expert finder on top of the FAST platform: it enables experts to personally control or change the number of questions that they are willing to answer at any given time.

Apparently, neither academic, nor industrial approaches to expert finding are ready to facilitate expertise selection at a full-scale level. While expert finders offered on the market are of a great help to improve organizational communication and knowledge flow, they are still too far from providing a complete solution. Such a long-awaited software that would assist at each step of expertise sharing and acquisition is envisioned in early research on expert finding [27,20], although no real solutions for design and implementation are proposed so far. Our work is the first attempt to not only decompose the expertise selection problem, but also to propose specific ways to overcome each discovered issue.

3 Expertise Selection with Explicit Preferences

While a lot of user preferences on meeting certain people are easy to infer just using common sense assumptions or global statistics of the system's usage, it is still reasonable to start the design of expertise selection component from making up a mechanism for setting up explicit user preferences on persons to communicate with. Such preferences could be of a great help for both sides: the

expertise seeker and the expert who is ready to share the expertise under certain conditions. Although preferences on certain individuals are easily imagined, typically, users do not know everyone in their enterprise and hence should bid their preference only on features they like or dislike in people. In this connection, it is important to draw a line between two types of a person's features: static features whose value do not change or change slowly over time (e.g., age, gender, position in a company, education, etc.) and dynamic, or context-specific, features that may vary even within a minute (e.g., location, emotional state, workload).

In the data management field, there are two well-known approaches to managing users' preferences, namely, quantitative and qualitative [13]. The qualitative approach intends to directly specify preferences between the data tuples in the query answer, typically using binary preference relations. An example preference relation is "*prefer expert A to another expert B if and only if A's rank in company X is higher than B's and A holds no part-time positions in other companies*". These kinds of preference relations can be embedded into relational query languages through relational operators or special preference constructors, which select from its input the set of the most preferred tuples (e.g., Winnow [13], PreferenceSQL [21], Skyline [12]). The quantitative approach expresses preferences using a scoring function, which associates a numeric score with every tuple of the query. A framework for expressing and combining such kinds of preferences functions was provided by Agrawal [3] and Koutrika [22]. The latter approach seems more appropriate for the expertise selection task, as well as for any task with a majority of non-binary or uncertain features.

Our design of a knowledge-based context-aware preference model allows to take set of ranked experts and *select the topmost preferred* of them by *re-ranking* the set. We use a variant of Description Logics [6] to represent preferences and then apply a probabilistic inference mechanism. Let F be a function that for each expert e gives its features: $F(e) = \{f_1, f_2, \ldots, f_n\}$. We use $Prob(f \in F(e))$ to express the probability that feature f holds for expert e. We assume that the features are independent. Similarly, let function G give for each preference p its context features (e.g., $hasStatus.\{Free\}$): $G(p) = \{g_1, g_2, \ldots, g_m\}$. (Since the context may feature different properties over time, whereas experts have features that are relatively stable, we call the latter *static* and the former *dynamic*.) The satisfaction of a context feature usually depends on measurements returned from error-prone (hence *uncertain*) sensors; therefore we use $Prob(g \in G(p))$ to denote the probability that preference p has context feature g. To decide whether a given expert satisfies the user's preference, we need to determine the probability that the expert is ideal for the required context features of the preference. To this end, let σ be the *score function* that for each pair (g, f) in the observed history H returns a score $\sigma(g, f)$: the probability in H that for a preference with context feature g, the user has selected an expert with static feature f. Now, in terms of these concepts, the probability that an expert e is the ideal one according to preference p, can be expressed. Details of our inference mechanism are described elsewhere [37].

4 Expertise Selection with Implicit Preferences

Despite the fact that explicit preferences are usually indispensable for personalized systems, users are often not enough enthusiastic to accurately specify their preference models, although still expecting the system to be efficient. However, some preferences are likely to be assumed by all users by default. When a group of two or more users agrees to meet for expertise sharing, their usual demand from the system which is responsible for arranging such a meeting is to organize it as soon as possible, considering current and future locations. If that is not possible, the next preference is to have such a meeting in any reasonable time in the future, during some common free time slot of all involved users.

4.1 Implicit Preference on Immediate Meetings

Most requests coming from users to experts are short questions awaiting for short answers. In such cases, when a momentary communication is sufficient for the desired knowledge exchange, an expertise selection component needs just to infer current user locations, make a guess about the time all users need to approach each other and, if an immediate meeting is possible, inform all interested persons about such an opportunity. While the inference of a current user activity and his/her level of occupancy is preferable, the location context is usually selective enough to filter out a lot of opportunities. The proximity of the users to communication facilities, e.g., a videoconference system or a phone, could be also considered.

The most important information for the localization of users is usually taken from several types of "sensors". An important observation is that a lot of location information *is already present, but not exploited* in many enterprises, and our goal is to make full use of such information. Possible sources for current location information are:

- GPS-enabled devices are getting more and more wide-spread, especially among mobile phones (e.g. IPhone).
- WiFi access points that register the proximity of an expert's WiFi-enabled mobile devices such as laptop or PDA. Enterprise buildings are usually well-covered by WiFi, but an expert may not always carry his/her mobile device.
- Bluetooth access points that register the proximity of cellphones. A cellphone is more often carried on the body, but Bluetooth coverage is usually sparse. However, more and more PCs are equipped with Bluetooth, so coverage is potentially high in places where people are at work.
- Registration of access cards gives a broad indication of in which (part of a) building a person is located.
- Computer activity provides an accurate indication of where the user is, given that the location of the computer is known.
- Simple webcams or microphones can pinpoint the presence of a person. Face or speech recognition may even identify this person (although the accuracy of these techniques may not be high, the combination with other sensors like Bluetooth may prove useful).

Essential characteristics for this kind of "sensor data" are that it is uncertain, incomplete and heterogeneous. The user locations returned by sensor networks are therefore always probability distributions and hence to deal with this, we use probabilistic models. They model where a person can go, what devices he has with him, and where people or devices can be sensed by several technologies. The *observed variables* consist of the signal strengths of WiFi scans, detections of Bluetooth devices, recorded computer activity, etc. Some of these observations can be modeled as *instantaneous*, others (like Bluetooth or WiFi scans) take a certain *interval* to complete. A *probabilistic model* connects the observed variables to the *query variables*: in this case, these are the locations of the experts at each point in time t. Additional unobservable variables, such as whether expert A has device B with him, may also play a role in the probabilistic model.

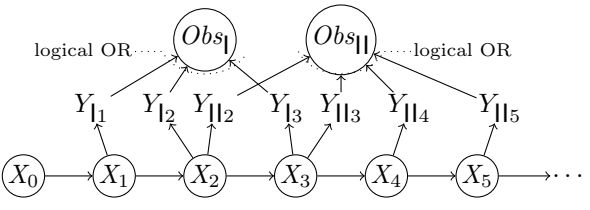


Fig. 1. Probabilistic model with Noisy-OR intervals

A general schema of our probabilistic model is shown in Fig.1. On top is a layer of observed variables; on the bottom, a layer of variables representing the *state* of the experts at each point in time. This state, represented in the figure by variable X_t, is split up into several sub-variables containing location and carried devices of each expert, but this is not shown. The graph in the figure corresponds to the structure of the Bayesian Network that defines the model; informally, an arrow pointing from X to Y means that the probability distribution of Y depends directly on the value of X. In our model, there is a layer of Y variables between states and observations. A variable Y expresses the fact that the observation of a certain device has been influenced by the location of this device at time t. This model is a variant on the Hidden Markov Model with Noisy-OR observations (HMM-NOR), and has the pleasant property that, for binary observations such as Bluetooth scans, the complexity of probabilistic inference stays linear in the length of the interval. The details of this approach are elaborated in a forthcoming paper [18].

4.2 Implicit Preference on Arranging Meetings in the Future

In cases when certain users cannot meet each other at the time of the request for expertise, their communication can be scheduled for the future with the help of an intelligent meeting planning mechanism. It is reasonable to assume that the agendas of all involved users have implicitly the user preference not to meet within occupied time slots. Considering that, the preference to meet with certain

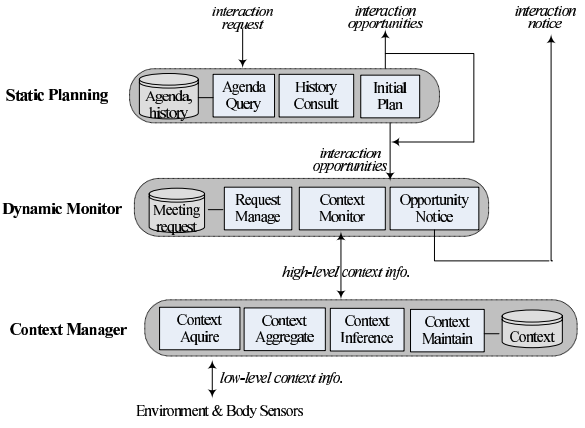


Fig. 2. Intelligent meeting planning

experts in the future –if the user information need will not be satisfied by that time– should not be ignored.

Figure 2 sketches how our meeting planning approach works. It first queries the agendas of both parties to find out possible time slots that satisfy requestor's constraints and preferences. Meanwhile, it also consults the previous conversation history (log) of the wanted person to enforce or complement the possibilities, which will then be returned to the user as an initial answer in the order of time. After that, it starts to monitor these possibilities, and meanwhile keeps eyes open on new unanticipated opportunities. When the behavior of the requested person deviates from the original schedule in his/her agenda, an immediate conversation might be possible as well. For instance, the requested person finishes a meeting and returns to his/her office 10 minutes earlier than scheduled, exhibiting a possibly good opportunity for a short conversation. The context monitor is responsible for checking context information in a timely fashion and decides whether a particular possibility is indeed a good conversation opportunity. Once this is the case, a chance alert will be sent to the user, so that a conversation between the user and the wanted expert can be conducted right away.

To do this, the context manager of the module plays an important role. It gathers low-level context information from various context suppliers such as sensors, and performs context aggregation and context inference so as to derive high-level context. Necessary context is stored into a context database for later retrieval and analysis. Besides, the context manager has the duty to answer pull-context queries, and actively execute push-context actions, in response to the requests of the context monitor. The final answer to the request will be logged (memorized), so that the context manager can do learning and reasoning in order to deliver smart solutions later on.

5 Privacy Control

Our approach to expert selection and meeting uses a lot of data about the people involved. In corporate environments, access control techniques [2,11] are insufficient for privacy protection, since access control can easily be bypassed by system administrators and the employers themselves. Employees have to fully trust their employers and such trust they will not put forever, especially not in cases where there is a conflict between employer and employee. Moreover, in cases when the enterprise is a subject to investigation by governmental organizations, the stored data will be the subject of investigation too, and possibly even disclosed to the public afterwards. The Enron fraud investigation in 2002 exemplifies this [30]. Hence, although it is tempting to store all data, a balance is needed between infinite storage—keeping full potential for new or existing services—and no storage at all to make sure that privacy sensitive data will never be disclosed.

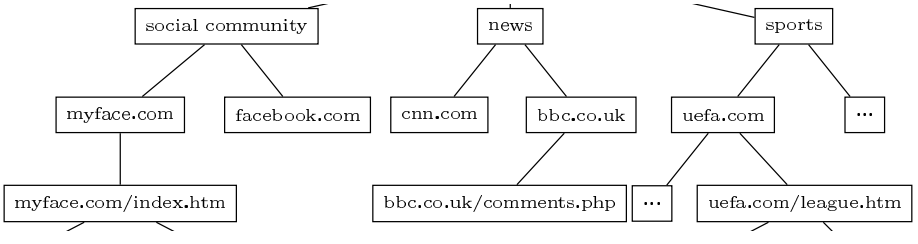


Fig. 3. Example of a generalization tree for the browsed web page attribute.

We propose a new technique termed *data degradation*, which is based on the assumption that long lasting purposes can often be fulfilled with less accurate data. Privacy sensitive data will progressively be degraded from *most accurate* via *generalized intermediate states*, up to complete removal from the system. The data degradation technique can be applied to both the data used at expertise identification stage (e.g., to the user web browsing history) and the data used at the stage of expertise selection (e.g., location traces). Using this technique, the enterprise is urged to carefully think about the form and period they need and want to store data, and gives employees the possibility to express what they find acceptable in terms of privacy, with a useful compromise between data usability and privacy as a result. The privacy benefit for employees is that they do not have to worry that the collected data about them can be misused in the future.

We consider the collected data as being a collection of *trails* of employees. A trail consists of a set of attributes, some of which are considered as privacy sensitive. In our data degradation model, termed the *life cycle policy model*, a trail is subject to progressive degradation from the accurate state to less detailed intermediate states, up to disappearance from the database. The degradation of each piece of information is captured by a *Generalization Tree*. Given a *domain generalization hierarchy* for an attribute, a generalization tree for that attribute

gives, at various levels of accuracy, the values that the attribute can take during its lifetime. Hence, a path from a particular node to the root of the tree expresses all degraded forms the value of that node can take in its domain (see Figure 3). Note however, that when storing privacy sensitive data in regular databases, it must be ensured that at each degradation step, the data will be irreversibly removed from the system, which is not a straightforward task [4,36]. For more details of the model we refer to [5].

6 Conclusion

This paper reports on our joint integral efforts towards a full-fledged expertise sharing solution for enterprises, covering expertise identification and expertise selection stages. The latter includes the use of implicit and explicit preferences of users on meeting each other. We also proposed a solution for privacy protection, which is applicable at all stages of the expert finding process. The work presented here still remains at a rather preliminary stage with a number of interesting issues to be addressed in the near future. So far we have designed the system more from the functionality aspect, and have ignored the efficiency perspective. For instance, a user's expert finding task can follow different execution plans, where all involved modules can execute in different sequential orders so as to minimize expert selectivity. Besides, implementing the whole system and using it in a real setting is necessary for the proof of concept.

Acknowledgments. This work is funded by the Dutch Organization for Scientific Research (NWO-VIDI project), Dutch Ministry of Economic Affairs (MultimediaN-AmbientDB project), and Centre for Telematics and Information Technology of the University of Twente in the Netherlands.

References

1. Ackerman, M.S., Wulf, V., Pipek, V.: Sharing Expertise: Beyond Knowledge Management. MIT Press, Cambridge (2002)
2. Agrawal, R., Kiernan, J., Srikant, R., Xu, Y.: Hippocratic databases. In: 28th Int'l Conf. on Very Large Databases (VLDB), Hong Kong (2002)
3. Agrawal, R., Wimmers, E.L.: A framework for expressing and combining preferences. In: SIGMOD 2000: Proceedings of the 2000 ACM SIGMOD international conference on Management of data, pp. 297–306. ACM Press, New York (2000)
4. Anciaux, N., Bouganim, L., van Heerde, H.J.W., Pucheral, P., Apers, P.M.G.: Instantdb: Enforcing timely degradation of sensitive data. In: Proceedings of the 24th International Conference on Data Engineering, April 2008. IEEE Computer Society Press, Los Alamitos (2008)
5. Anciaux, N.L.G., Bouganim, L., van Heerde, H.J.W., Pucheral, P., Apers, P.M.G.: Instantdb: Enforcing timely degradation of sensitive data. In: Proceedings of the 24th International Conference on Data Engineering (ICDE 2008), Cancun, Mexico, April 2008, pp. 1373–1375. IEEE Computer Society Press, Los Alamitos (2008)

6. Baader, F., Calvanese, D., McGuinness, D.L., Nardi, D., Patel-Schneider, P.F. (eds.): The Description Logic Handbook: Theory, Implementation, and Applications. Cambridge University Press, Cambridge (2003)
7. Balog, K., Bogers, T., Azzopardi, L., de Rijke, M., van den Bosch, A.: Broad expertise retrieval in sparse data environments. In: SIGIR 2007: Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval, pp. 551–558. ACM, New York (2007)
8. Baumard, P.: Tacit Knowledge in Organizations. Sage Publications, Inc., Thousand Oaks (2001)
9. Becerra-Fernandez, I.: Facilitating the online search of experts at NASA using expert seeker people-finder. In: PAKM 2000, Third International Conference on Practical Aspects of Knowledge Management (2000)
10. Becerra-Fernandez, I.: Searching for experts on the web: A review of contemporary expertise locator systems. ACM Trans. Inter. Tech. 6(4), 333–355 (2006)
11. Bertino, E., Byun, J.-W., Li, N.: Privacy-preserving database systems. In: Aldini, A., Gorrieri, R., Martinelli, F. (eds.) FOSAD 2005. LNCS, vol. 3655, pp. 178–206. Springer, Heidelberg (2005)
12. Börzsönyi, S., Kossmann, D., Stocker, K.: The skyline operator. In: Proceedings of the 17th International Conference on Data Engineering, April 2-6, 2001, pp. 421–430. IEEE Computer Society Press, Los Alamitos (2001)
13. Chomicki, J.: Preference formulas in relational queries. ACM Trans. Database Syst. 28(4), 427–466 (2003)
14. Craswell, N., de Vries, A., Soboroff, I.: Overview of the trec-2005 enterprise track. In: Proceedings of TREC 2005, Gaithersburg, USA (2005)
15. Craswell, N., Hawking, D., Vercoustre, A.-M., Wilkins, P.: Panoptic expert: Searching for experts not just for documents. In: Ausweb Poster Proceedings, Queensland, Australia (2001)
16. Davenport, T.: Knowledge Management at Microsoft. White paper (January 1997), http://www.itmweb.com/essay536.htm
17. Davenport, T.: Ten principles of knowledge management and four case studies. Knowledge and Process Management 4(3) (1998)
18. Evers, S., Fokkinga, M., Apers, P.M.G.: Probabilistic processing of interval-valued sensor data. In: Proceedings of DMSN 2008, 5th International Workshop on Data Management for Sensor Networks, To be held in conjunction with VLDB 2008, Auckland, New Zealand, August 24 (to appear, 2008), http://dmsn08.cs.umass.edu/
19. Fang, H., Zhai, C.: Probabilistic models for expert finding. In: Amati, G., Carpineto, C., Romano, G. (eds.) ECiR 2007. LNCS, vol. 4425, pp. 418–430. Springer, Heidelberg (2007)
20. Johansson, C., Hall, P.A.V., Coquard, M.: ¨talk to paula and peter - they are experienced¨- the experience engine in a nutshell. In: Ruhe, G., Bomarius, F. (eds.) SEKE 1999. LNCS, vol. 1756, pp. 171–185. Springer, Heidelberg (2000)
21. Kießling, W.: Foundations of preferences in database systems. In: VLDB, pp. 311–322 (2002)
22. Koutrika, G., Ioannidis, Y.E.: Personalized queries under a generalized preference model. In: Proceedings of the 21st International Conference on Data Engineering, ICDE 2005, Tokyo, Japan, 5-8 April 2005, pp. 841–852. IEEE Computer Society Press, Los Alamitos (2005)

23. Liu, X., Croft, W.B., Koll, M.: Finding experts in community-based question-answering services. In: CIKM 2005: Proceedings of the 14th ACM international conference on Information and knowledge management, pp. 315–316. ACM Press, New York (2005)
24. Lu, W., Robertson, S., Macfarlane, A., Zhao, H.: Window-based Enterprise Expert Search. In: Proceedings of the 15th Text REtrieval Conference, TREC 2006 (2006)
25. Macdonald, C., Ounis, I.: Voting for candidates: adapting data fusion techniques for an expert search task. In: CIKM 2006: Proceedings of the 15th ACM international conference on Information and knowledge management, pp. 387–396 (2006)
26. Maybury, M.T.: Expert finding systems. Technical Report MTR06B000040, MITRE Corporation (2006)
27. McDonald, D.W., Ackerman, M.S.: Just talk to me: a field study of expertise location. In: CSCW 1998: Proceedings of the 1998 ACM conference on Computer supported cooperative work, pp. 315–324. ACM Press, New York (1998)
28. Miller, S., Weckert, J.: Privacy, the workplace and the internet. Journal of Business Ethics 28(3), 255–265 (2000)
29. Petkova, D., Croft, W.B.: Hierarchical language models for expert finding in enterprise corpora. In: ICTAI 2006: Proceedings of the 18th IEEE International Conference on Tools with Artificial Intelligence, pp. 599–608 (2006)
30. Rudman, C.S.G., LLP, R.: The enron fraud (June 2008), `http://www.enronfraud.com/`
31. Serdyukov, P., Hiemstra, D.: Modeling documents as mixtures of persons for expert finding. In: Macdonald, C., Ounis, I., Plachouras, V., Ruthven, I., White, R.W. (eds.) ECIR 2008. LNCS, vol. 4956, pp. 309–320. Springer, Heidelberg (2008)
32. Serdyukov, P., Rode, H., Hiemsta, D.: Exploiting sequential dependencies for expert finding. In: SIGIR 2008: Proceedings of the 31th annual international ACM SIGIR conference on Research and development in information retrieval (2008)
33. Serdyukov, P., Rode, H., Hiemsta, D.: Modeling expert finding as an absorbing random walk. In: SIGIR 2008: Proceedings of the 31th annual international ACM SIGIR conference on Research and development in information retrieval (2008)
34. Serdyukov, P., Rode, H., Hiemstra, D.: Modeling multi-step relevance propagation for expert finding. In: CIKM 2008: Proceedings of the 17th ACM conference information and knowledge management (2008)
35. Soboroff, I., de Vries, A., Craswell, N.: Overview of the trec-2006 enterprise track. In: Proceedings of TREC 2006, Gaithersburg, USA (2006)
36. Stahlberg, P., Miklau, G., Levine, B.N.: Threats to privacy in the forensic analysis of database systems. In: SIGMOD 2007: Proceedings of the 2007 ACM SIGMOD international conference on Management of data, pp. 91–102. ACM Press, New York (2007)
37. van Bunningen, A.H., Fokkinga, M.M., Apers, P.M.G., Feng, L.: Ranking query results using context-aware preferences. In: First International Workshop on Ranking in Databases (DBRank 2007), Istanbul, Turkey, April 2007, pp. 269–276. IEEE Computer Society Press, Los Alamitos (2007)

Semantic and Event-Based Approach for Link Prediction

Till Wohlfarth[1,2] and Ryutaro Ichise[2]

[1] University Paris VI, Pierre et Marie Curie, 4, place Jussieu Paris, France
[2] National Institute of Informatics, Tokyo 101-8430, Japan

Abstract. The scientific breakthroughs resulting from the collaborations between researchers often outperform the expectations. But finding the partners who will bring this synergic effect can take time and sometime gets nowhere considering the huge amounts of experts in various disciplines. We propose to build a link predictor in a network where nodes represent researchers and links - coauthorships. In this method we use the structure of the constructed graph, and propose to add a semantic and event based approach to improve the accuracy of the predictor. In this case, predictors might offer good suggestions for future collaborations. We will be able to compute the classification of a massive dataset in a reasonable time by under-sampling and balancing the data. This model could be extended in other fields where the research of partnership is important as in world of institutions, associations or companies. We believe that it could also help with finding communities of topics, since link predictors contain implicit information about the semantic relation between researchers.

1 Introduction

Research collaborations are always beneficial, as they often yield good results. However, searching for the collaborator who will satisfy the expectations, happens to be an arduous task because of the lack of awareness of the existence of other researchers with whom collaborations might prove fruitful. Even with this awareness, it would be difficult to predict which potential collaborations should be pursued. The researchers finally doesn't know who he should collaborate with.

Assuming that collaborations often happen when the researchers meet each other, for example in a conference, or when they collaborated with the same persons, we can establish a predictor based on the structure of a network where nodes would be the researchers and links - collaborations. Then, we can transform the problem of finding collaborators in a link prediction problem where predicting a link would be like advising a partnership.

However, the graph structure is often very sparse and doesn't allow to use classical graph metrics. Furthermore, considering a huge network of researchers from every fields would increase significantly the computation time and would compel us to considered a smaller graph.

Researchers' semantic descriptions might be very helpful here. If one knew to what extent each researcher is an expert in each field, one could potentially

T. Yamaguchi (Ed.): PAKM 2008, LNAI 5345, pp. 50–61, 2008.

use this knowledge to find researchers with compatible expertise and suggest collaborations. Moreover, an event-based approach where we consider the common conferences where researchers presented their works, or the common journal where they were published, should contribute to identify more precisely the potential future collaborations for the same reasons. We are implicitly finding communities of researchers by topics and by event.

Our method extracts structural attributes from the graph of past collaborations along with semantic and event-based features, and uses them to train a set of predictors using supervised learning algorithms. These predictors can then be used to predict future links between existing nodes in the graph. We test our method on a co-authorship network and the results confirm that the appearance of new collaborations is dependent on the semantic description and the numbers of the common event. We also prove that supervised learning methods can exploit this dependence to make predictions with reasonable accuracy on condition that we preprocess the dataset. The approach is not specific to the domain of co-authorship networks and can be easily applied to virtually all types of networks in which link prediction is desirable.

Section 2 briefly mentions the other researches and compares them to our method. Section 3 formally defines how we approach the link prediction problem. Section 4 describes the experiments we run and discusses their results. Finally, Section 5 concludes.

2 Related Works

Research in the area of network evolution models [14] is closely related to the problem of link prediction. Kashima et al. [9] attempt to fit a parametrized "copy-and-paste" model to a partial snapshot of a network and use this model to predict the full network structure. The same problem (though viewed in slightly differing ways) is tackled by [16] and [18]: they build partial networks from observable data and use them to infer links that are not directly observable but are likely to exist. These approaches all differ from our view of the problem, as they only work with a static snapshot of a network and do not consider network evolution over time.

A similar view to ours is employed in [11], where a collaboration network is cut into time slices and the network structure of one time slice is used to predict the structure in the next slice. A link is considered likely to appear if a structural attribute for the two end-nodes achieves a high score. The authors test and compare the predictive power of several such attributes. While generally achieving a low prediction rate, their approach is promising in the sense that it significantly outperforms a random predictor. This implies there is, in fact, a correlation between the structure of a collaboration network and the places where collaborations are likely to appear in the future.

The more related work is the method for link prediction using *multiple* structural attributes in the prediction of a single link, from Pavlov and Ichise [15]. We will see in Section 3 the detail of their procedure as we decided to use it as a base for our work.

Though these works in social network analysis has often focused on the graph structure of the network, without making any use of the properties of the nodes themselves. Indeed, we can collect more informations than only the link structure, such as titles of the papers for coauthorship networks. An application using these features tries also to predict the future collaborations [7] and the semantic attribute was the top ranked attribute that contribute to well predict a link.

Our approach aims to build an improved method for link prediction by utilizing attributes of the node itself and to improve the computation time of the work of [15]. The following explains how.

3 Method

3.1 Construct the Graph and the Features Vectors

The Structural Attributes. This work is based on the work of Pavlov and Ichise [15], so to understand our contribution let us introduce the previous work. Based on a dataset of papers, we build an undirected graph where nodes are researchers and links represent a coauthorship between two researchers. The links are weighted by the number of coauthorship between the researchers. Then, for every couples of researchers, we construct a feature vector, a vector of numbers which are calculated with different graph algorithms, and apply supervised learning algorithms to predict future edges, that is, if the couple is more likely to write a paper together. A feature vector containing n attributes can be mapped to a point in n-dimensional space (where every dimension corresponds to an attribute). Thus, our feature vectors set F is represented by a set of points in this space. Each point then has its own binary label. The goal is to separate the points into two groups so that points with the same label are in the same group. In the previous work, the features were based on the structure of the graph, using for example the shortest path between two nodes or their PageRank index. Table 1 shows all these attributes.

The *shortest path* between two nodes v_i and v_j is defined as the minimum number of edges connecting them. If there is no such connecting path then the value of this attribute is generally assumed to be infinite. *Common neighbours* counts the number of neighbours that the two nodes have in common. *Jaccard's coefficient* [12] is a normalized measure of common neighbours. It computes the ratio of common neighbours out of all neighbours (common or not). This is sometimes a better measure than the (unnormalized) common neighbours, especially when one end-node has a substantially larger neighbourhood than the other. *Adamic/Adar* [1] measures similarity between two nodes by weighing "rarer" common neighbours more heavily. The rationale is that two nodes that have a common neighbour that no other node has are often more "similar" than two nodes whose common neighbours are common for many other nodes. *Preferential attachment* [13] says that new links will be more likely to connect higher-degree[1] nodes than lower-degree nodes. In a collaboration network, this

[1] The degree of a node is equal to the number of edges linking to it. Thus, $\deg(v_i) = |\Gamma(v_i)|$.

Table 1. Structural attributes computed for each node pair (v_i, v_j)

Attribute name	*Formula*
Shortest path	$\min\{s \mid \text{paths}_{ij}^s > 0\}$
Common neighbours	$\lvert\Gamma(v_i) \cap \Gamma(v_j)\rvert$
Jaccard's coefficient	$\frac{\lvert\Gamma(v_i)\cap\Gamma(v_j)\rvert}{\lvert\Gamma(v_i)\cup\Gamma(v_j)\rvert}$
Adamic/Adar	$\sum_{v_k \in \Gamma(v_i)\cup\Gamma(v_j)} \frac{1}{\log\lvert\Gamma(v_k)\rvert}$
Preferential attachment	$\lvert\Gamma(v_i)\rvert \cdot \lvert\Gamma(v_j)\rvert$
Katz_β	$\sum_{s=1}^{\infty} \beta^s \cdot \text{paths}_{ij}^s$
Weighted Katz_β	Same as above, but $\text{paths}_{ij}^1 = w_{v_i,v_j}$
PageRank_d^{min}	$\min\{\text{PageRank}_d(v_i), \text{PageRank}_d(v_j)\}$
PageRank_d^{max}	$\max\{\text{PageRank}_d(v_i), \text{PageRank}_d(v_j)\}$
SimRank_γ	$\begin{cases} 1 & \text{if } v_i = v_j \\ \gamma \cdot \frac{\Sigma_{a\in\Gamma(v_i)}\Sigma_{b\in\Gamma(v_j)}\text{SimRank}_\gamma(a,b)}{\lvert\Gamma(v_i)\rvert\cdot\lvert\Gamma(v_j)\rvert} & \text{otherwise} \end{cases}$
Link value	w_{ij}

means a new collaboration is more likely to occur between authors who collaborate more often (regardless of who they collaborate with).This (unnormalized) likelihood is reflected by the simple product of the nodes' degrees. *Katz* [10] is a refined measure of shortest path. It considers all paths between two nodes and weighs shorter ones more heavily. The "non-attenuation" parameter $\beta \in [0, 1]$ controls the aggressiveness of weighing. E.g., a very small β yields a measure which is similar to common neighbours, since paths^s-values for higher s will not contribute significantly to the summation. *Weighted Katz* uses the same core formula as Katz, but also observes the weight between linked nodes. Thus, two nodes that are connected by "heavier" paths will achieve a higher score. *PageRank* is the same core algorithm used by Google to rank search results [2]. Using our notation, $PageRank(v_i) = (1-d) + d\Sigma_{v_m \in \Gamma(v_i)} \frac{PageRank(v_m)}{\Gamma(v_m)}$. In effect, the rank of a node in the graph is proportional to the probability that the node will be reached through a random walk on the graph. $d \in [0, 1]$ is a damping factor which specifies how likely the algorithm is to visit the node's neighbours rather than starting over. Note that the original algorithm computes ranks over nodes. Since we need ranks over pairs, we take the minimum and maximum page ranks for the two nodes in a pair. *SimRank* [8] is also recursively defined. It states two nodes are similar to the extent they are connected to similar neighbours. $\gamma \in [0, 1]$ is a parameter that controls how fast the weight of connected nodes SimRank decreases as they get further away from the original nodes. Finally, the *link value* of a pair of nodes is simply the weight of the edge between them. If such an edge does not exist then the value is assumed to be zero.

The "Non-structural" Attributes. The previous approach, only based on structural features implies that researchers are more likely to collaborate with

people of there *entourage*. However, it happens that communities based on the same topic are not related at all, or by very few links, because of the real distance between the people or because of a non-existing partnership between the laboratories. In these many cases, having the structure of the graph is not enough to predict the best partner in a specific domain. Our approach is to combine the structural and non-structural attributes to resolve the problem of link prediction: the most obvious idea for linking researchers is to compare the main topic of their research. Thus, by counting the number of words in common between all the titles of their previous papers, we can have a new feature based on the semantic and not on the network structure. First of all, we preprocess the titles to eliminate the stop words (words that doesn't add a meaning, like "the" or "a"). Then we use the Jaccard's coefficient, as define in previous section, to compare the similarity of two sets of words. We call this feature the "Keywords match count" (KMC) [7] which could be similar to a clustering of the graph by topics. The KMC is defined by

$$KMC(w_i, w_j) = \frac{|w_i \cap w_j|}{|w_i \cup w_j|}$$

where w_i and w_j are sets of words.

The semantic feature doesn't use the structure of the graph, thus, we can have good results even on small datasets which are very sparse.

To be more accurate, we can combine the KMC with a feature that count the number of events in common between researchers [6]. An common event could be a journal where they both wrote a paper or a conference where they both presented their work. Every nodes will have a collection of events that we can after that compare to other sets of events. The idea is also to forget about the structure of the graph and only look at the relation between the works of every researchers. Attending that every conferences and journals focus on an aspect of their main field, we are using a layer that includes the KMC. Thus, we can avoid the recurrent problem of words with several meanings considering the topics of the paper. In our case, to calculate this metric, we calculate the number of conferences in common and then we also use the Jaccard's coefficient.

3.2 Supervised Learning Method for Building Link Predictor

In this work we are using a decision tree as the main supervised learning algorithm. We use the J48 algorithm which is the Weka [19] adaptation of the C4.5 algorithm [17]. A decision tree is a binary tree whose terminal nodes are classification outcomes and non-terminal nodes - decisions. In our case the classification is the prediction of the future link: the couple of researchers is a positive example (they could possibly make a collaboration) or the couple is classified as a negative example (they probably won't make a paper together). Therefore, traversing the tree from the root to the leaf, correspond to progressively separate the set of feature vectors in groups according to a specific feature. At every non-terminal node, we cut the dataset in two, according to the attribute that best differentiates the set. An advantage of this approach is that we can observe the learned

decision tree directly and potentially make meaningful conclusions about which attributes are truly important for the prediction problem. The process terminates when all nodes in a partition have the same label or when no attributes can split a partition any further. After that, we are able to classified a couple of researchers as potentially "linkable" by traversing the tree and choosing the right branch regarding the feature vector of the couple.

After we construct a feature vector for every possible pairs of researchers, we notice that, logically, the huge part of the couples didn't work together. The dataset becomes imbalanced and skew the supervised learning algorithm: by predicting that none of the couples will collaborate, the decision tree will be correct at almost 100%, but we will not be able to predict a future coauthorship. Furthermore, considering the huge size of the dataset, the processing of learning could manage to take a lot more than a day. So we propose to preprocess the dataset to reduce and balance it without modifying the general structure of the network.

3.3 Preprocessing of the Dataset

A dataset is imbalanced if the classification categories are not approximately equally represented. There are two ways of rebalancing the dataset: one is the under-sampling of the majority class examples by deleting some negative examples and the second is to over-sampling the minority class examples by creating synthetics examples. We utilize both methods for proprocessing the data.

First we conduct under-sampling the majority class in training dataset on nodes that are reachable (there is a path between them). The dataset is very sparse because of the many topics from every fields and despite the presence of big sub-graphs, the percentage of non reachable nodes is still huge. During the learning process, we are using feature vectors of researchers who will never meat each other: they don't work on the same topic, they are not part of the same team etc. Thus, we decided to train the learning algorithm, with couples of researchers who could know each other. Nodes that aren't reachable could represent these researchers who won't work together. Deleting them will decrease hugely the training set, giving better results with a better computation time and it will prevent the overfitting: the learning algorithm overfits when it fits so correctly on the training set that it doesn't fit on the testing set. Instead of making a general prediction, it became too specific.

Nevertheless, the set of examples is still imbalanced and we have to oversample the minority class. In [3], many types of method are proposed and analysed, and the conclusion is that the SMOTE algorithm [4] (Synthetic Minority Over-sampling Technique) answers to our demands: to overcome the overfitting and increase the decision region of minority class examples, it generates synthetic examples by operating in "feature space" rather than "data space". Synthetic examples are generated in the following way: Take the difference between the feature vector under consideration and its nearest neighbour. Multiply this difference by a random number between 0 and l, and add it to the feature vector under consideration. This causes the selection of a random point along the line segment between two specific features.

The synthetic examples cause the classifier to create larger and less specific decision regions. More general regions are now learned for the minority class rather than being subsumed by the majority class examples around them. The effect is that decision trees generalize better. Furthermore, Chawla et al. [4] exposed that combining the C4.5 decision tree algorithm and the SMOTE method outperformed other methods.

4 Experiment

In this research, we use the *Digital Bibliography & Library Project* (DBLP) database [5], hosted at Universitat Trier, in Germany, which contains bibliographic information on major computer science journals and proceedings. To be able to compare our result to the previous work we had to use only a part of the DBLP database, which is conferences about Artificial Intelligence. It represents 17623 authors, 18820 papers which creates 81058 links in the graph. but only 0.01% of the DBLP dataset. Though, it has already a massive computation time.

This network contains 22 years of evolution history and it might be difficult to see (or predict) evolution patterns unless we consider some method of time slicing. We split the data set into 16 partitions (from 1987 to 2002, we take the range of six years beginning by this date), each consisting of six years of evolution data, and perform the same experiment on every partitions independently.

For every partition, we take three first years, and create the network $G = \{V, E, W\}$ of these years as follow. We construct the network. V is the set of all authors in the data set.

$$E = \{(v_i, v_j) \in V \times V | v_i \text{ and } v_j \text{ have co-authored at least one paper}\}$$

Finally, the weight of a collaboration between two authors, w_{ij}, is equal to the number of co-authored papers between them.

Before extracting feature vectors we draw attention to several things. First, we note that many collaborations between authors occur only once. This results in many edges of weight one. These "tenuous" edges do not seem to carry much information about collaboration tendencies between authors - in fact, they could often be attributed to chance[2]. For this reason, we find it reasonable to *filter out* all edges whose weight is equal to one.

From the six years of evolution data we need to extract feature vectors and corresponding labels. We use the first three years for the former and the remaining three for the latter. The label of the feature vector is a boolean value that defines whether the example is positive, the label is true, or negative, the label is false. If the example is positive, it means that the couple of researchers collaborated in the three last years of the range. We note, however, that some authors are only active during the first four years but stop publishing after that.

[2] E.g., two authors might never personally collaborate, yet still have their names appear on a paper because of a third party. Unless this happens again, we consider the link between such authors accidental.

Table 2. Structure of the graphs between 1987 and 2002

Years	*Nodes*	*Edges*	*Positive examples*	*Negative examples*	*% of unreachable pairs*
1987	1844	4136	249	1698997	99
1988	2090	4874	258	2182747	98.1
1989	2049	4896	308	2097868	99.1
1990	1970	4604	360	1939105	99.1
1991	1877	4736	361	1760265	99
1992	2005	4820	473	2008537	90
1993	2209	5574	550	2438186	99
1994	2408	5864	739	2897289	99
1995	2763	7636	817	3814886	99
1996	3094	9036	867	4784004	99
1997	3481	10676	998	6055942	99
1998	3753	11648	306	1460889	99
1999	3624	11168	1047	6563829	99
2000	3566	11308	1234	6355161	99
2001	3358	10522	1573	5634830	99
2002	3935	13364	1976	7738169	99

Conversely, others only start publishing in the latter three years and are inactive in the preceding time period. In the graph, this results in a variable nodes set V, which leads to a mismatch between feature vectors and labels. To avoid this, we *trim* the set V to only contain authors that are active during both time periods. In the Table 2, we expose the structure of the resulting networks.

At this point we have an average proportion of positive examples at about 0.0002%. Then we use or not the SMOTE algorithm and compute the J48 algorithm.

For training the basic dataset, we use a subset containing 90% of the vectors, exampled at random. Testing is performed on the remaining 10%. Implementations of the J48 learning algorithm are already available in the Weka software package [19]. We use these implementations for our experiments. Then for the filtered dataset and the over-sampling one, we do the training with all modified dataset and the testing is performed on the regular dataset to compare the final classification.

We have also to define some performance metrics for predictor evaluation. For each feature vector, a predictor p can make either a *positive* (P) or a *negative* (N) prediction concerning the corresponding label[3]. In the positive case, if p is correct, the prediction is said to be *true-positive* (TP); otherwise it is *false-positive* (FP). Conversely, in the negative case a prediction can be either *true-negative* (TN) if correct or *false-negative* (FN) if wrong. We can now define the metric *recall* as the proportion of TP predictions out of all *true* labels. Recall will give us an idea of how well p is able to predict collaborations that will happen in the future. It might also be useful to define the metric *precision* as

[3] To disambiguate: A positive prediction means p thinks the label is *true*.

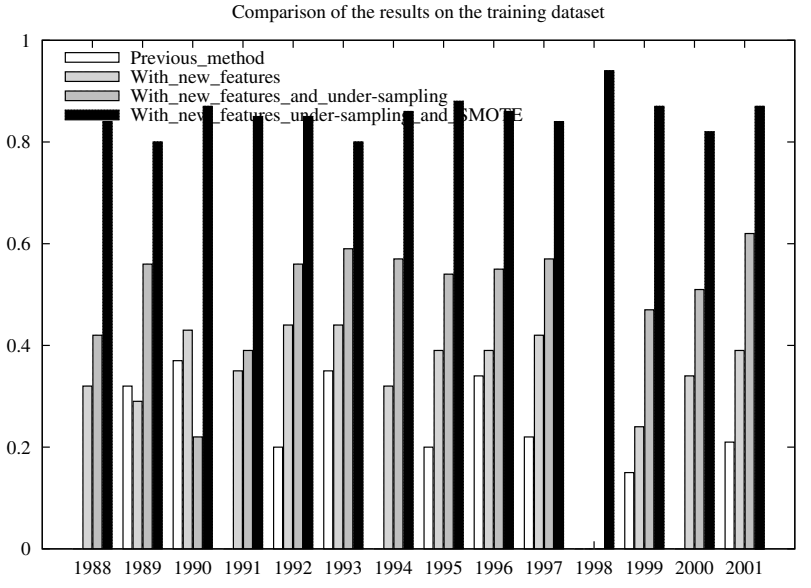


Fig. 1. Evaluation on the training set

the proportion of TP predictions out of all positive predictions. Precision will be useful in determining how well p fits the whole data (as opposed to just always predicting *true*, which guarantees a 100% recall rate).

$$precision = \frac{|TP|}{|TP| + |FP|}, \quad recall = \frac{|TP|}{|TP| + |FN|}$$

Using both precision and recall we are able to introduce a final metric called the F-measure or F-score to numerically evaluate and compare predictors. The F-score can be interpreted as a weighted average of the precision and recall, where an F-score reaches its best value at 1 and worst score at 0.

$$FScore = 2 * \frac{precision \times recall}{precision + recall}$$

4.1 Results

In this section, we will show the improvements of our contribution by testing successively the classification of the previous method [15], then the one with the new features and finally, by examining the results with the under-sampling and with the SMOTE algorithm. For the following experimentations, the parameters of the SMOTE methods are 200 for the percentage of increase of the minority class and 5 for number of nearest neighbours that we use.

On the Figure 1 and Figure 2, we clearly see the gradual increase of the F-Score, both on the train dataset and the test dataset. The new features have a

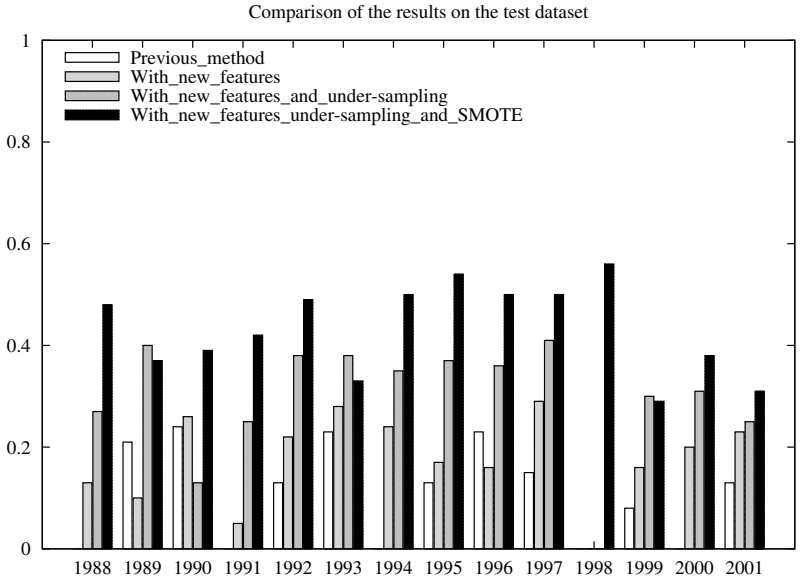


Fig. 2. Evaluation on the testing set

beneficial effect but, because of the imbalanced dataset problem, we still have a low score. The first treatment, which deletes the unreachable pairs of researchers, manages to increase the results almost every years and sometimes doubles the F-Score. If we look closer at the year 1998, we can see that the F-Score is null for the first three methods. Actually, during the learning process, the decision tree classified all the examples as negative because there were not enough positive examples. The correctly classified examples percentage is still near the 99% but because there are no examples classified as positive,the recall is null. By over-sampling the positive examples set, we observe an undeniable effect of the SMOTE algorithm, which has been able to double the F-Score of almost every years, and it performed the best score for the year 1998. Considering the comparison of the computation time, a graphic wouldn't help us to see how we decreased the time of learning because of the massive difference. The reduction factor varies between 50 and 1000.

5 Conclusion

This paper presented a supervised learning method for building link predictors from structural attributes of the underlying network and from non-structural attributes of the nodes as the semantic. It also proposes a way to improve the computation time and the performances at the same time using under-sampling and over-sampling methods. In a network of researchers, where a link represents a collaboration, such predictors could be useful is suggesting unrealized

collaborations and thus help in building and maintaining strong research teams. In addition, by analysing the algorithmic structure of predictors constructed for a specific network, we could gain valuable information about which attributes are most informative for the link prediction problem and use this knowledge as a basis for specifying vocabularies for researcher description. This approach tries to help the world of the research but can also be deployed in other branches where collaborations have this synergetic effect.

There are many future research in this area to take. We think an important next field of investigation could be the time slicing process which usually do not evolve at a constant rate. It would be interesting to see if this rate varies significantly and if we can adjust the durations of the history and evolution time spans to take such variations into account. One way to do this would be by compiling different pairs of history-evolution data and comparing predictor performances on them.

References

1. Adamic, L.A., Adar, E.: Friends and neighbors on the web. Social Networks 25(3), 211–230 (2003)
2. Brin, S., Page, L.: The anatomy of a large-scale hypertextual Web search engine. Computer Networks and ISDN Systems 30(1–7), 107–117 (1998)
3. Chawla, N.V.: Data mining for imbalanced datasets: An overview. In: The Data Mining and Knowledge Discovery Handbook, pp. 853–867. Springer, Heidelberg (2005)
4. Chawla, N.V., Bowyer, K.W., Hall, L.O., Kegelmeyer, W.P.: Smote: Synthetic minority over-sampling technique. Journal of Artificial Intelligence and Research 16, 321–357 (2002)
5. The DBLP computer science bibliography (2008), `http://dblp.uni-trier.de/xml/`
6. Getoor, L., Diehl, C.P.: Link mining: a survey. SIGKDD Explor. Newsl. 7(2), 3–12 (2005)
7. Hasan, M.A.: Link prediction using supervised learning. In: Proceedings of the Workshop on Link Analysis, Counter-terrorism and Security (2006)
8. Jeh, G., Widom, J.: Simrank: a measure of structural-context similarity. In: Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining, pp. 538–543. ACM Press, New York (2002)
9. Kashima, H., Abe, N.: A parameterized probabilistic model of network evolution for supervised link prediction. In: Proceedings of the Sixth International Conference on Data Mining, pp. 340–349. IEEE Computer Society Press, Los Alamitos (2006)
10. Katz, L.: A new status index derived from sociometric analysis. Psychometrika 18(1), 39–43 (1953)
11. Liben-Nowell, D., Kleinberg, J.: The link prediction problem for social networks. In: Proceedings of the twelfth international conference on Information and knowledge management, pp. 556–559. ACM Press, New York (2003)
12. Manning, C.D., Raghavan, P., Schütze, H.: Introduction to Information Retrieval. Cambridge University Press, Cambridge (2008)
13. Newman, M.E.J.: Clustering and preferential attachment in growing networks. Physical Review E 64, 025102 (2001)

14. Newman, M.E.J.: The structure and function of complex networks. SIAM Review 45(2), 167–256 (2003)
15. Pavlov, M., Ichise, R.: Finding experts by link prediction in co-authorship networks. In: Proceedings of the 2nd International Workshop on Finding Experts on the Web with Semantics (November 2007)
16. Popescul, A., Ungar, L.H.: Statistical relational learning for link prediction. In: Proceedings of Workshop on Learning Statistical Models from Relational Data (2003)
17. Ross Quinlan, J.: C4.5: Programs for Machine Learning. Morgan Kaufmann, San Francisco (1993)
18. Taskar, B., Wong, M.-F., Abbeel, P., Koller, D.: Link prediction in relational data. In: Proceedings of Neural Information Processing Systems (2003)
19. Witten, I.H., Frank, E.: Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations. Morgan Kaufmann, San Francisco (1999)

Social Semantic Bookmarking

Simone Braun, Valentin Zacharias, and Hans-Jörg Happel

FZI Research Center for Information Technology, Haid-und-Neu-Str. 10-14,
76131 Karlsruhe, Germany
{braun,zacharias,happel}@fzi.de

Abstract. In this paper we present the novel paradigm of Social Semantic Bookmarking. Social Semantic Bookmarking combines the positive aspects of semantic annotation with those of social bookmarking and tagging while avoiding their respective drawbacks like the cumbersome maintenance of ontologies or the lacking semantic precision of tags. Social semantic bookmarking tools allow for the annotation of internet resources based on an ontology and the integrated maintenance of the ontology by the same people that use it. We introduce Social Semantic Bookmarking and present the SOBOLEO application as an implementation of this paradigm.

Keywords: Social tagging, semantic tagging, social bookmarking, SOBOLEO.

1 Introduction

A big challenge for today's internet users is the focused discovery of new information that is likely to be interesting and useful as well as the rediscovery of information that they had once found and identified as such. Social bookmarking systems (e.g. such as del.icio.us[1]) allow for the collection, management, and sharing of bookmarks, i.e., references to such information entities. The users can easily annotate these bookmarks with arbitrary tags that help in organizing, navigating and searching the bookmark collection.

These tags, however, are completely unstructured. Problems such as homonyms, synonyms, multilinguality, typos or different ways to write words, and tags on different levels of abstraction hamper search and retrieval; in particular in complex domains. Replacing tags with semantic annotations based on an ontology as a controlled vocabulary can help here.

Many systems that allow for annotating documents with respect to ontologies struggle, however, with a number of problems, too. Not only are they cumbersome to use but they also view ontology creation as a process separate from its use, performed by people different from those that use it. These systems also often assume that the ontology stays unchanged for longer periods of time and is updated only seldomly. All this leads to unsatisfied users being confronted with out-of-date, incomplete, inaccurate and incomprehensive ontologies that they cannot easily use for annotation; this problem is particular acute in fast changing domains [9].

[1] http://delicious.com

T. Yamaguchi (Ed.): PAKM 2008, LNAI 5345, pp. 62–73, 2008.

The novel paradigm of Social Semantic Bookmarking combines the positive aspects of semantic annotation with those of social bookmarking while avoiding their respective drawbacks. Social semantic bookmarking tools allow for the annotation of internet resources with respect to an ontology and the integrated maintenance of the ontology by the same people that use it. Through the use of state-of-the-art web technologies such as bookmarklets and AJAX (e.g., for auto complete functionality), these systems make ontology-based annotation of web documents as simple as tagging. Through easy-to-use, lightweight web ontology editors that are integrated into the system, the barrier between ontology creation and use is removed; users who annotate with the help of the ontology are the same who continuously evolve this ontology. Because internet resources are annotated with concepts (and not keywords), the problems of homonyms, synonyms etc. are avoided.

We present Social Semantic Bookmarking using the example of our system SOBOLEO (Social Bookmarking and Lightweight Engineering of Ontologies) – a system combining the above mentioned features with an innovative search engine and functionality supporting the discovery of experts on specific topics based on their interaction with the system. We also shortly discuss other social semantic bookmarking systems such as Bibsonomy, int.ere.st, GroupMe!, Fuzzy, and Annotea.

2 Social Tagging vs. Semantic Annotation

2.1 Social Tagging and Its Problems

Social bookmarking systems allow their users for annotating bookmarks with several arbitrary tags they find most suitable for describing them. In this way – in contrast to the traditional folder structure like browser favorites – users can organize their bookmarks according to more than one category. This facilitates the organization, navigation, and search in the bookmark collection. The popularity of such social tagging applications have shown that this organizing principle with tags and folksonomies evolving from these is much easier accessible for users than structured and controlled vocabularies; in particular for collaborative applications.

These applications, however, often reach their limits because of lacking semantic precision of tags. Folksonomies have only very limited structure. Their missing semantic precision hampers efficient search and retrieval support, in particular in complex domains, because of problems like the following (cf. [6,7]):

- **(Mis-)Spelling:** The most obvious problem is that tags are simply misspelled or written in different ways because of occurring plurals, abbreviations or compound words, e.g. 'spagetti' vs. 'spag**h**etti', 'noodle' vs. 'noodle**s**', or 'spaghetti**C**arbonara' vs. 'spaghetti_**c**arbonara'.
- **Multilingualism:** Tags only relate to one language. That means, especially in Europe with many different languages, users have to annotate a resource with many tags in different languages, e.g. with 'pasta', 'noodles', and 'Nudeln', in order to ensure that other users will find it later on (e.g. to promote their own great spaghetti recipe).
- **Polysemy:** Tags can have several similar meanings. This leads to search results with low precision because of irrelevant resources; e.g. with the tag 'pasta' the users

can think of a dish that contains pasta as its main ingredient or of the aliment itself as shaped and dried dough made from flour and water and sometimes egg.

- **Homonymy:** The problem of homonymy is comparable to the problem of polysemy. However, in this case, one tag can have several totally different meanings. This also leads to irrelevant results as all resources that relate to these different meanings are annotated with the same tag. For instance the word 'noodle' can have the meaning of an aliment but also of a swearword for a human head.
- **Synonymy:** Resources are not found because they are annotated with another tag with the same meaning, e.g. with the tag 'vermicellini' instead of 'spaghettoni'. Similar to mulitlingualism, the users have to annotate the resources with many synonymous tags in order to ensure the retrieval by other users.
- **Mismatch of abstraction level:** Also a typical search problem emerges because tags are specified on different abstraction levels, i.e. either too broad or too narrow. This problem, also known as the "basic level phenomenon" [16], can be traced back to different intentions and expertise levels of the users. For instance, one user tags a resource on the basic level with 'spaghetti', another with 'noodles' and a third differentiates 'spaghetti' from 'bigoli' (thicker spaghetti) and 'vermicelli' (thinner spaghetti). A resource annotated with 'spaghetti', however, cannot be found with the search term 'pasta'.

2.2 Semantic Annotation and Its Problems

Replacing tags with semantic annotations based on an ontology promises to solve the limits of (linguistic) tagging-based applications. Ontologies, as formalizations of a shared understanding of a community [8], contain background knowledge of a certain domain. They improve the description or the retrieval of resources (in its broadest sense) by making subject, creation, usage, relational or other context of these resources explicit.

These semantic annotation approaches also rely on the use of some standardized formal language for representing the ontology, such as RDF [14], SKOS [15], or one of the OWL languages [5]. They have a number of potential benefits:

- **Better Retrieval:** The formally represented relations between the concepts in the ontology can be used to offer superior browse or query facilities. In the case where a powerful language like OWL is used, queries may even be answered using reasoning algorithms.
- **Better Use of Annotation:** The availability of machine understandable context for the used annotation terms can be utilized to make better use of the annotation; e.g. information that some annotations represent geographic locations for which a latitude and longitude is known can be used to show the annotated document in a map or to make them available based on the users current location.
- **Better Quality Assurance:** The information contained in the ontology about concepts used for annotation can enable checks on whether an annotation is likely to make sense; this can help to catch errors early. Also changes in the ontology can be checked whether they violate its integrity.
- **Better (Semantic Web) Integration:** The ontology that is used in the annotation is usually assumed to be also used in other systems and the common usage of the

ontology can enable the integration of data created and managed in these diverse systems. Another related aspect is that semantically annotated data can become part of the Semantic Web and then Semantic Web aware agents and applications can make use of it.

- **Better Support of Vocabulary Management:** Through the use of standardized languages to represent the ontologies, these approaches can rely on a landscape of tools that is available to create, manage and evolve these ontologies.

Many systems that allow for annotating with respect to ontologies, however, have not found widespread adoption yet and struggle with a number of problems, too. To a large extend because the annotation process, i.e. the usage of the ontology, and the creation of the ontology are two separate processes, performed by a different set of people. Annotation is done by the users of a semantic application and the ontologies are created by dedicated knowledge engineering specialists. However, separating the use and the creation of the ontology and involving knowledge engineering specialists is causing a number of problems:

- **High Cost:** Knowledge engineers are highly paid specialists, and their effort comprises not only the actual implementation of the domain ontology, but also learning about and understanding the domain of interest. While in many Web 2.0 scenarios a large amount of work is done for free by users interested in the result, this is unlikely to work when knowledge engineers with little innate interest in the domain in question are involved.
- **Domain Errors:** Knowledge engineers are specialists for the domain of knowledge formalization – not for the domain that is being formalized. For this reason they will not have an understanding of the domain comparable to that of domain experts, this limited understanding may cause errors in the resulting ontology [2].
- **Heavyweight Process and Upfront Investment:** Because annotation cannot start without an available ontology, there needs to be an upfront investment to finance the development of this ontology, which includes a systematic requirements elicitation phase. During the usage phase of the ontology, there also needs to be a accompanying process to collect newly emerging requirements, bugs and other change requests and to implement them into a newer version of the ontology.
- **High Time Lag:** There will always be some time lag between the emergence of a new concept and the time when it is included in the ontology and can eventually be used. This time lag is relatively large, when the users of the ontology cannot make the change themselves but must rely on knowledge engineers understanding the requirement, implementing it and finally rolling out the new version of the ontology. In fast moving domains this time lag can quickly get so big that the ontology as a whole becomes unusable [7].
- **Low Appropriateness and Understandability:** An ontology is appropriate for a task if it enables the users to reach their goals more quickly. However, having different people using and developing the ontology makes reaching appropriateness of the ontology much harder. A particular challenge is to ensure that the ontology is at the right level of abstraction to be understood by the domain experts.

3 Social Semantic Bookmarking

In the previous sections we have seen that (linguistic) social tagging approaches, while popular, struggle with problems such as polysemy, multilingualism or abstraction level mismatches. At the other end many current semantic annotation approaches struggle with the problem of timely updates and appropriateness of the underlying ontology as well as affordable creation. Social Semantic Bookmarking now combines the benefits of tagging with semantic annotation in order to address their respective weaknesses.

Social semantic bookmarking systems allow for the annotation of resources (e.g. web pages, documents) with concepts whose definition and description also evolves collaboratively within the same system. Similar to tagging approaches, they allow for creating new concepts whenever a need arises. Unlike these approaches, concepts can have powerful descriptions and can be interlinked; for example allowing the system to understand that 'swimming bath' and 'swimming pool' are synonyms for the same concept. These powerful concept descriptions are similar to those used in traditional semantic annotation, but social semantic bookmarking allows for adding and changing concepts permanently and easily at the time the concepts are used.

The SOBOLEO[2] system [17] is a particular social semantic bookmarking system that will be used to further illustrate this approach in this section. SOBOLEO is based on AJAX technology and works in most current browsers – thus does not require any local installation. It consists of four application parts: an editor for the modification of the shared ontology, a tool for the annotation of internet resources, a semantic search engine for the annotated internet resources, and an ontology browser for navigating the ontology and the bookmark collection.

SOBOLEOs functionality and the concept of Social Semantic Bookmarking will be further described with an example of a user who annotates an internet resource with a new concept 'sphaghetti', then adds some information about this new concept. A different user will then search for the annotated resource at a different level of abstraction and find it using the semantic search feature.

3.1 Annotation

The annotation process starts when a user finds an interesting resource that she wants to add to the shared repository. In this example a user discovers a tasty pasta recipe. In order to annotate the document the user clicks on a bookmarklet in her browser which opens up the small dialog window (see Fig. 1). The user can annotate the web document using any of the concepts already known to the system and is supported by auto completion in doing that. Here the user also adds a new concept named 'Spaghetti' – adding a concept is seamlessly done by simply entering a term that is not yet known to the system.

Once the user clicks save, the system stores the URL of the document with all assigned concepts; any new concepts are also added to the shared ontology of the repository. The SOBOLEO system crawls the content of the annotated web page that is added to a full text index associated with a repository.

[2] http://www.soboleo.com

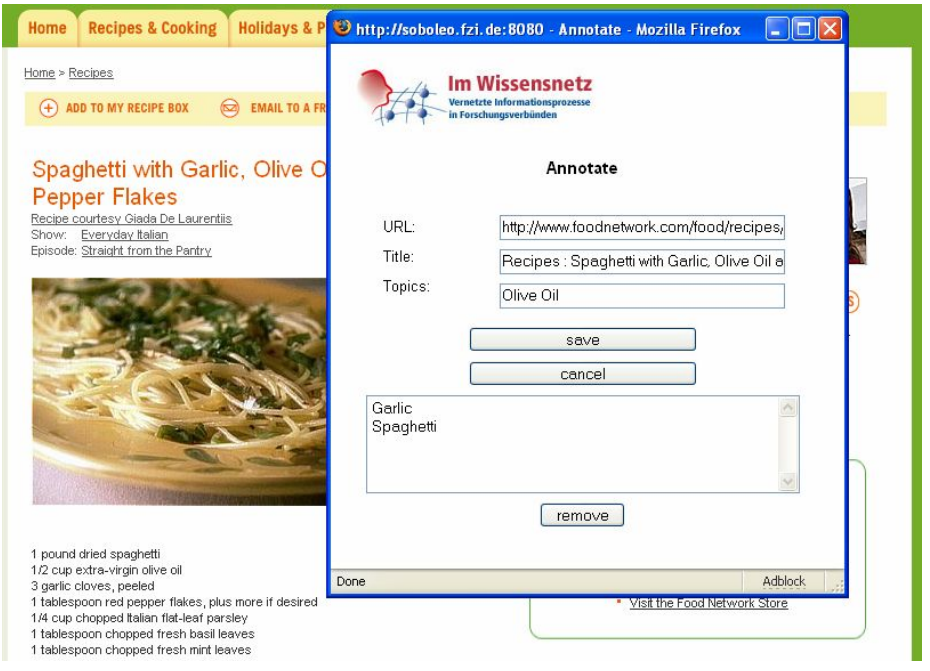


Fig. 1. Annotating a web page

3.2 Ontology Editing

Each user of SOBOLEO belongs to a user group that has a shared repository containing the annotations and the ontology. Such a user group consists of people working on the same topic, such as a department in a large company or a special interest group spanning continents.

The ontology in the shared repository is represented using a subset of the SKOS standard; it allows for concepts with a preferred label, a description and any number of alternative labels. It also allows for broader, narrower, and related relations between concepts. The ontology in this shared repository is edited using the AJAX editor (see Fig. 2). The editor is a collaborative realtime AJAX editor; i.e., it can be used by multiple persons simultaneously in their respective browsers with edits showing up for the others in realtime.

In the example the user opens the editor to add more information about the new 'Spaghetti' concept. First the user uses the mouse to drag the 'Spaghetti' concept onto the 'Pasta' concept, quickly establishing the relation that 'Spaghetti' is a narrower concept than 'Pasta'. She also adds a short description to 'Sphaghetti' and 'Spaghetto' as synonym.

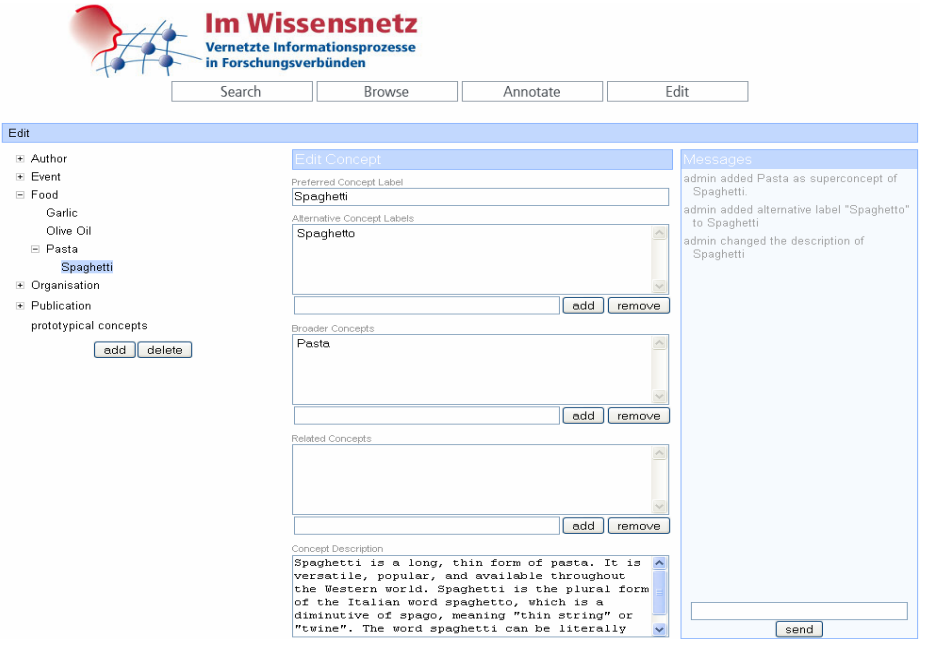


Fig. 2. Collaborative realtime onotology editor

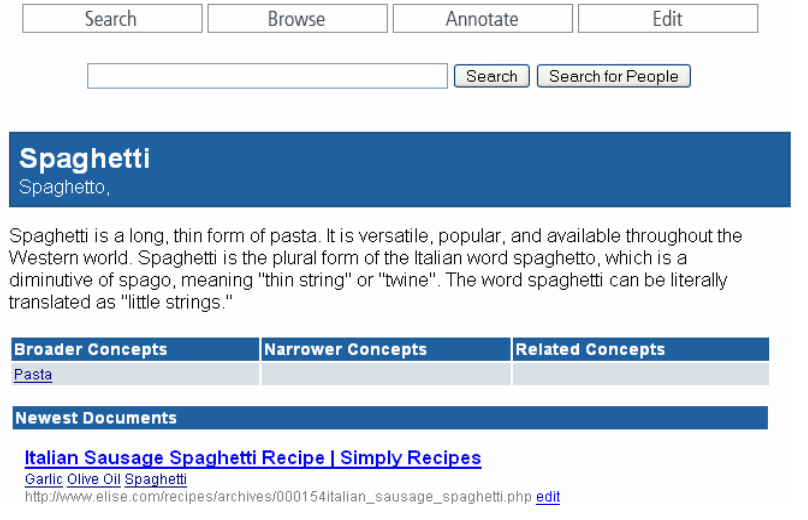


Fig. 3. Browsing interface for navigating to concepts and annotated resources

3.3 Browsing the Repository

Browsing the repository is the most common approach to retrieving information from a shared repository. With a browsing interface users can navigate to the concepts they are interested in and see the resources annotated with these. The browser interface also gives the chance to change any of the annotations. In SOBOLEO and social semantic bookmarking the user can also see the ontology and use the taxonomic structure for navigation. Fig. 3 shows the browsing interface for the new 'Spaghetti' concept. The interface shows the concept name, its labels and its description. Also shown are the most recently annotated documents (with links to change the annotation) and the relations to other concepts allowing for navigating there.

3.4 Semantic Search

In addition to the browse interface the ontology is also used to enable semantic search. The semantic search in SOBOLEO combines semantic search utilizing the concept labels and their broader-narrower relations with a full text search over all annotated resources. The semantic search engine also offers query refinement and relaxation functionality.

Fig. 4. Result of the semantic search

In the example, a different user is interested in finding a recipe including noodles, garlic and basil and enters these words as search term. The semantic search recognizes that 'noodles' is a synonym for pasta and that spaghetti is a special kind of pasta. The search engine further finds that garlic refers to another concept and then that the annotation described earlier combines not only spaghetti and pasta as annotation but also includes basil in the sites content – hence this page is returned as a first result. The result is shown in Fig. 4. Please note that neither a full text engine (because 'noodles' is not written on the page), nor a social tagging system (because neither noodles nor basil is a tag), nor a pure semantic search engine (because basil is not annotated) could make a comparable ranking of the result.

4 Related Work

There are a number of other approaches often presented as (social) semantic tagging. These include Bibsonomy, Int.ere.st, GroupMe!, Fuzzy, and Annotea, which we will describe and compare in the following:

- **BibSonomy:** Bibsonomy [10] is a system for the management of bookmarks of internet resources and publication entries. Bibsonomy offers functionality similar to that of well-known social bookmarking services but specifically tailored towards academics – e.g., it offers sophisticated support for uploading and exporting bibliographic information. At its core, Bibsonomy has a functionality very similar to social bookmarking services, but additionally offers users the possibility to create broader/narrower relations between tags. However, tag relationships are only local, i.e., each user can (and has to) maintain its own relationships and cannot profit from others' contributions in that respect.
- **Int.ere.st:** Int.ere.st [11] is a system concentrating on the transferability of tags and tagged resources between systems. Int.ere.st is created by the Digital Enterprise Research Institute, Galway and the Biomedical Knowledge Engineering of Seoul National University, Korea. Its functionality centers on making uploading and exporting tagging data simple and to allow for creating relations between tags (potentially coming from different systems).
- **GroupMe!:** GroupMe [1] attempts to bridge the gap between the Semantic Web and Web2.0 with an RDF based social bookmarking application. GroupMe! is developed by the Semantic Web Group at the University of Hannover in Germany. The main unique functionality of GroupMe! is the extension of the tagging idea with the concept of 'groups': all annotated resources can be organized into groups and these form another level of information that can be used for browsing and search.
- **Fuzzy:** Fuzzy [13] is a system for managing bookmarks of internet resources and ISBN numbers. Fuzzy is developed within the PhD project of Roy Lachica at the University of Oslo. It is based on Topic Maps technology and besides parent/child and horizontal tag relations the users can choose of 22 specific predefined association types to link tags. Another main concept is voting for gardening and maintenance: the users can vote on bookmarks, tags a bookmark is annotated with, relations between tags, and users.
- **Annotea:** Annotea [12] is a metadata standard for semantic web annotations, it is implemented in a number of tagging tools and server applications. Annotea and its implementations have been developed by the W3C. Annotea differs from other approaches to social tagging in its emphasis on standards on decentrality, that it has sharing of bookmarks among services build in from ground up.

4.1 Comparison

To give a comprehensive overview of the respective strength and weaknesses of the approaches shortly introduced above, Tab. 1 details the main discriminating features among the applications, including SOBOLEO. The features used for the comparison are the following:

Table 1. Comparison of social semantic bookmarking tools

	Public	Full Text Indexing	Import/ Export Formats	Synonyms	Other Relations	Shared Relation Editing	Open Source
Bibsonomy	Yes	No	XML, RSS, BURST, SWRC, Bibtex	No	Broader/ Narrower	No	No
Int.ere.st	No	No	SCOT, SIOC, FOAF	Yes	Identical	No	No
GroupMe!	Yes	No	RSS, DC, FOAF	No	Group	Yes	No
Fuzzzy	Yes	Yes	XTM, RSS	Yes	Broader/ Narrower, Specific association types	Yes	No
Annotea	No	No	Annotea	Yes	Broader/ Narrower	No	yes
SOBOLEO	No	yes	SKOS, RSS	Yes	Broader/ Narrower, Related	Yes	No

- **Public:** Whether the application has a public installation that can be used by any user.
- **Full Text Indexing:** Whether the application stores the text content of the annotated resources and uses it to facilitate search.
- **Import/Export Formats:** All tools discussed have some means to import or export the bookmarks, this row details which formats are used.
- **Synonyms:** Whether the application supports a notion of two natural language terms representing the same thing.
- **Other Relations:** The relations between tags/concepts that are supported by the applications, other than synonyms.
- **Shared Relation Editing:** Whether relations between tags exist only for one user or whether they are shared, i.e. in some systems the relation between tags is only visible to one user. Other users would need to create the same relation again.
- **Open Source:** Whether the source code of the applications is available as open source.

As a general conclusion, there is a big interest to extend social bookmarking in the direction of more semantics and in particular to tackle the problem how tagging data can be exchanged between systems, however, at the same time the table shows that

there still is considerable disagreement about what are the most important features and – even more crucially – what are suitable formats to exchange the tagging data. Without an agreement in this domain, the promise of exchanging tagging data can obviously not be achieved. It is also interesting to see that the majority of the approaches still restricts the editing of relations between tags to only the private space and/or do not allow for a real community driven evolution of the semantic model.

5 Conclusion

Social Semantic Bookmarking allows a group of users to collaboratively create and evolve an index of resources together with the powerful semantic vocabulary used to organize it. Social Semantic Bookmarking promises better retrieval, better use of annotation, better integration of the repository with semantic web infrastructure etc. while avoiding the problems commonly associated with semantic annotation approaches – such as a high initial cost to build ontologies.

Parts of the vision of Social Semantic Bookmarking are already realized and used today, and evaluation studies like [3] confirm that users appreciate the new paradigm. In three user studies with 4, 24, and 33 participants we found that users liked the ease of use of the ontology editing (in comparison to other, more heavy-weight applications) and particular enjoyed the simple way of annotating resources with concepts or tags. Some users had initial problems, due to their very basic knowledge about ontologies, but all were able to obtain the necessary skills within a very short time.

Social Semantic Bookmarking applications promise a huge potential for future development as part of the developments towards a Web 3.0 as a user-centered semantic web. However, to realize this potential we also need a better understanding of the emergence and evolution of ontologies as part of everyday collaborative activities and appropriate models and support mechanisms. Promising research approaches include the ontology maturing process [4], which is further explored as part of the Integrating Project MATURE[3].

References

1. Abel, F., Henze, F.M., Krause, D., Plappert, D., Siehndel, P.: Group Me! Where Semantic Web meets Web 2.0. In: Proc. of the 6th Int. Semantic Web Conf. (2007)
2. Barker, K., Chaudhri, V.K., Char, S.Y., Clark, P., Fan, J., Israel, D., Mishra, S., Porter, B.W., Romero, P., Tecuci, D., Yeh, P.: A question-answering system for AP Chemistry: Assessing KR&R technologies. In: Proc. of the 9th Int. Conf. on Principles of Knowledge Representation and Reasoning, pp. 488–497 (2004)
3. Braun, S., Schmidt, A., Walter, A., Nagypal, G., Zacharias, V.: The Ontology Maturing Approach to Collaborative and Work-Integrated Ontology Development: Evaluation Results and Future Directions. In: Proc. of the ESOE-Workshop at ISWC 2007, CEUR Workshop, vol. 292, pp. 5–18 (2007)
4. Braun, S., Schmidt, A., Walter, A., Nagypal, G., Zacharias, V.: Ontology Maturing: a Collaborative Web 2.0 Approach to Ontology Engineering. In: Proc.of the Workshop on Social and Collaborative Construction of Structured Knowledge at WWW 2007, CEUR Workshop, vol. 273 (2007)

[3] http://mature-ip.eu

5. Dean, M., Schreiber, G.: OWL Web Ontology Language Reference. W3C Recommendation (February 10, 2004)
6. Golder, S., Huberman, B.A.: The Structure of Collaborative Tagging Systems. J. of Inf. Sc. 32(2), 198–208 (2006)
7. Guy, M., Tonkin, E.: Folksonomies: Tidying Up Tags? D-Lib Magazine 12(1) (2006)
8. Gruber, T.R.: Towards Principles for the Design of Ontologies Used for Knowledge Sharing. Int. J. of Human-Computer Studies 43, 907–928 (1995)
9. Hepp, M.: Possible Ontologies: How reality constraints building relevant ontologies. IEEE Internet Computing 11(1), 90–96 (2007)
10. Hotho, A., Jäschke, R., Schmitz, C., Stumme, G.: BibSonomy: A Social Bookmark and Publication Sharing System. In: CS-TIW 2006. Aalborg University Press, Aalborg (2006)
11. Kim, H.L., Yang, S.-K., Song, S.-J., Breslin, J.G., Kim, H.-G.: Tag Mediated Society with SCOT Ontology. In: Proc. of the 5th Semantic Web Challenge at ISWC 2007 (2007)
12. Koivunen, M.-R.: Semantic Authoring By Tagging with Annotea Social Bookmarks and Topics. In: Proc. of the Semantic Authoring and Annotation Workshop at ISWC 2006 (2006)
13. Lachica, R., Karabeg, D.: Metadata creation in socio-semantic tagging systems: Towards holistic knowledge creation and interchange. In: Scaling Topic Maps. Topic Maps Research and Applications 2007. Springer, Heidelberg (2007)
14. Manola, F., Miller, E.: RDF Primer. W3C Recommendation (February 10, 2004)
15. Miles, A., Bechhofer, S.: SKOS Simple Knowledge Organization System Reference. W3C Working Draft (January 25, 2008)
16. Tanaka, J.W., Taylor, M.: Object categories and expertise: Is the basic level in the eye of the beholder? Cognitve Psychology 23(3), 457–482 (1991)
17. Zacharias, V., Braun, S.: SOBOLEO – Social Bookmarking and Lightweight Ontology Engineering. In: Proc. of the Workshop on Social and Collaborative Construction of Structured Knowledge at WWW 2007, CEUR Workshop Pro., vol. 273 (2007)

Closing Information Gaps with Inverse Search

Hans-Jörg Happel

FZI Research Center for Information Technologies
Karlsruhe, Germany
happel@fzi.de

Abstract. In this paper, we propose to improve the overall process of information retrieval by explicitly addressing information provision from private spaces of individual users into the public information space of an organization. Therefore, we present our approach of *inverse search*, which aims to stimulate the diffusion of documents from these private spaces. We introduce the notion of an organizational information need (OIN) based on query logs and further usage statistics of our system. This information is used to recommend people to share private documents containing relevant information. Our main contributions are describing means to identify documents that should be shared and a framework to foster the diffusion of such documents. We also describe the implementation and results from initial evaluation studies[1].

1 Introduction

Current web and enterprise search engines allow their users to query for relevant documents among the total set of documents available. Since only those documents can be returned which have been analyzed before the time of the query, this model is also called *retrospective search* [1].

However, in the web as well as in the enterprise, new documents which might be relevant for user queries are created continuously. In order to address this issue, the paradigm of *prospective search* [1,2] has been conceptualized. Prospective search systems allow users to store their queries and notify them, as soon as new results arrive. Popular examples for such systems are Google Alerts or Windows Live Alerts[2]. While prospective search acknowledges the dynamic nature of document creation, it does not address *how* and *why* new information is fed into the information retrieval process. In this paper, we propose a novel approach called *inverse search* which addresses these issues. While both retrospective and prospective search mainly consider an information seeker, her queries and a corpus of documents, we introduce information providers and their private information space as additional elements in the information retrieval process.

We believe that the role of information providers is especially important in an enterprise search context, since many people nowadays have to collaborate in

[1] This work has been supported in part by the TEAM project, which is funded by the EU-IST programme under grant FP6-35111 and the BMBF-funded project WAVES.

[2] http://www.google.com/alerts, http://alerts.live.com/Alerts/Default.aspx

T. Yamaguchi (Ed.): PAKM 2008, LNAI 5345, pp. 74–85, 2008.

ad hoc or distant project teams, introducing the need to exchange information and knowledge [3].

A common IT-based solution is the introduction of knowledge management systems (KMS, see e.g. [4,5]) such as central information repositories or groupware with shared folders. However, their successful introduction and adoption is influenced by a number of motivational, organizational and technical barriers [6,7] which make workers in an organizational setting avoid to share information (see section 2.3).

To the best of our knowledge, there exist no widespread technical approach to foster the sharing of private information held by individuals. With respect to document sharing, standard approaches are based on indexing all information – public and private – and exposing them to searchers. However, this requires an information exposure by the information provider, which might harm the adoption of those systems. Our main contributions are therefore to describe a method how to identify information that should be shared and to develop a framework to foster the diffusion of such information. We additionally describe the implementation and results from initial evaluation studies.

The structure of the paper is as follows: We first describe some basic notions which are referenced throughout the paper. Afterwards, we specify requirements which drive the technical underpinnings for our inverse search approach. We describe this approach in more detail, provide results from initital evaluation studies and summarize related work.

2 Foundations

2.1 Information Need

The information need of a user is a primary subject of investigation in information retrieval (IR). The main purpose of IR systems is to help users satisfying their information needs by providing a set of relevant documents. A *personal information need* can be defined as information which a user requires to complete a specific task [8]. To use an IR system, the user typically has to express this information need in terms of the query language which can be interpreted by the search system. In most systems, this is a textual, "keyword-based" representation of the information need.

Based on this definition of personal information need, we conceptualize *organizational information need* (OIN) as an aggregate of the personal information needs of members in an organization. Thus, the organizational information need denotes the overall amount of information, which the members in an organization require to complete their particular tasks. Concerning keyword searches, an OIN should thus represent the most frequent queries that have been executed throughout the organization.

2.2 Information Gap

The notion of a personal information need already implies a distinction between either having some information at hand or not. Talking about search systems,

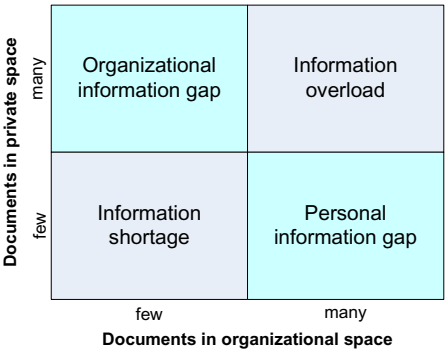


Fig. 1. Availability of documents containing certain terms in the private vs. organizational information space [9]

information needs can typically be satisfied either from the own private information space or from information available in the public space. The private spaces of other users are usually not accessible, although they might contain relevant information. As we have shown in previous work an unequal distribution of information exists across the overall information space due to a specialization of expertise and interests [9].

When comparing this distribution between the private space of a user and the public space of the organization, four typical situations can be distinguished as depicted in Figure 1. These situations and their implications for organizational knowledge management are discussed in [9]. In this paper, we are particularly interested in the sitution of *organizational information gap* which means that a user has lots of information on a topic, while there is few information publicly available. This means that other users, searching for that information might not be able to satisfy their information need, although there is information in the private space of at least one user in the organization.

From an organizational knowledge sharing perspective, this raises the issues 1) how to determine which information from private spaces should be made available to the organization ("need to share") and 2) how this diffusion can be achieved. Methods and barriers for knowledge sharing are discussed in the following section.

2.3 Knowledge Sharing

The topic of knowledge sharing is widely investigated in the fields of organizational studies and information systems. It can be defined as a "dual problem of searching for (looking for and identifying) and transferring (moving and incorporating) knowledge across organization subunits" [10].

Personalization and codification are typically described to be the two core strategies for knowledge sharing [11]. The personalization strategy primarily relies on personal communication to share implicit knowledge, which only exists

in the "heads" of individuals (see e.g. [11]). In turn, the codification strategy targets explicit knowledge which is captured in documents. Here information technology is an important enabler, especially in cases where regular personal communication is not feasible, such as in large organizations or distributed settings [4,5].

However, the success of KMS largely depends on individuals contributing their knowledge. One of the major challenges of KMS is to overcome barriers that prohibit people from doing so. The most important barriers are:

Motivation. People usually have a low motivation to contribute knowledge to public repositories. Reasons are a lack of personal benefit [12,13,14] and privacy, since people do not like to expose their information and expertise to others [15,16].

Effort. Computer-supported knowledge sharing initiatives require effort for creating and maintaining central knowledge repositories. This includes the cost of knowledge capturing, categorization and setting access rights for documents [16,17].

In this paper, we claim that both issues can be addressed by informing potential information providers about actual information needs in an organization. In the following section we describe the architecture and algorithms of our approach called *inverse search*.

3 Inverse Search

"Conventional" search systems, such as retroactive and prospective search, are focussed on information seekers and a public index of documents (see e.g. [8]). Although private information spaces might contain additional relevant documents, information providers are not part of the standard search model.

We conceptualize *inverse search* as information providers, matching their corpus against a given set of queries – in opposite to *conventional search*, where information seekers match queries against a given corpus of documents. While users "import" public documents into their private space in conventional search, inverse search helps to move documents from the private space to the public space, where they might satisfy the information need of other users.

With inverse search, we aim to conceptualize *how* and *why* documents move from private to public spaces. In the remainder of this section, we further elaborate on our approach for inverse search. We therefore describe requirements, architecture and algorithms to realize the envisioned knowledge sharing process.

3.1 Requirements

As described in section 2.2, our overall functional requirement is to foster the sharing of existing information and knowledge in organizations. This is achieved by enabling people to access relevant information, wherever it may exist in the organization. As lined out in section 2.3, state-of-the art KMS fall short on

a number issues, which actually limits their practical use. We address these limitations by a set of non-functional requirements for our envisioned system:

R1. Retain privacy. An information provider must not expose information to others by default. KMS implementations often lack acceptance, since contributing information to the public space means losing control about it. However, many information providers want to retain such control, since information might be premature or sensitive [18].

R2. Minimize effort. The effort for both, information providers and information seekers should be minimized. In contrast to existing approaches for prospective search [1,2] which recommend a set of documents to each query of a user, we need to restrict to a smaller set of documents. Since covering all information needs would require the sharing of too many documents, we aim to restrict sharing recommendations to a small set of documents, which cover a large organizational information need.

R3. Motivate to share. Information providers should be motivated to share relevant information with information seekers. Traditional KMS and knowledge sharing practices often require to share information without signaling any benefit to the provider. Thus, those practices are often perceived as self-purpose with an unclear value for the organization. In opposite to this, we want to give the potential information provider more concrete information that can help to estimate the benefit of sharing certain documents within the organization.

As already stated, these requirements are not satisfied by state-of-the-art information retrieval or knowledge management systems. However, as we will show in the following paragraphs, only slight modifications to existing architectures are required in order to realize the potential benefits for knowledge sharing.

3.2 Architecture

In order to differentiate between documents in public and in private spaces and to fulfill requirement R1, our system requires a public index for the documents in the public space ($Index_{Public}$) and a private index of the private documents of each user (e.g. $Index_{Alice}$ and $Index_{Bob}$). This private index is not accessible to any other user. The existence of a private index seems to be a viable precondition, since many users are already using desktop search engines nowadays.

Queries to the public index are automatically saved to a public *query log*. Both, the public index and the query logs can be retrieved by any user. In order to retain privacy (R1), queries may be anonymous and must not contain information about the querying user. However, if the user likes to receive automatic notifications when new information arrives, they might need to reveal their identity.

On her local machine, each user runs a *SearchApplication* which allow to query both the local index and the public index. Additionally, each user is running a *SharingEngine*, which periodically compares the local index with the global index and the query log.

Thus, the sharing engine can provide an estimation of how useful it would be to share a certain document. This helps to satisfy requirement R3, since the user is guided in her decision which documents are worth sharing. In order to minimize the effort of sharing (requirement R2), several ways are possible to suggest sharing certain documents to the user. This might either happen by enriching existing interfaces (e.g. by decorating existing document icons with information about its value) or by periodically presenting a ranked list of few documents, which the user should share within the organization.

3.3 Algorithms

Based on this infrastructure, the sharing engine of a user is able to identify interesting documents in her local space, which can be useful for other users in the organization. From an implementation perspective, this requires two steps, very similar to prospective search:

- Identify the existing information need, as represented by queries.
- Match these queries against the document corpus and identify appropriate results for recommendation.

In standard prospective search, these steps have a straightforward implementation: the user queries are given explicitly (e.g. defined "alerts" in Google Alerts) and documents are matched by simply querying the search engine. For the first step, Yang et al. [2] define a notion of "standing interest", which they automatically derive from query logs. After identifying a standing interest, a "typical" query for that interest is selected out of the queries of a user. Thus, this approach does not require users to explicitly register "alerts". In the second step, prospective search engines typically provide a list of all results for the registered query, which appeared since the previous notification. Yang et al. [2] discuss a number of heuristics in order to restrict this to "interesting results".

While we require the same two steps in our approach to identify appropriate results for sharing, we need a different implementation of these steps. There are two main reasons for this. First, unlike in prospective search, we do not want to recommend sharing documents related to queries of *single users*. Instead, we need to aggregate the information need of the users in the organization in order to identify the most pressing organizational information needs.

Secondly, we do not want to identify *any new documents* which could be shared. Instead we want to limit the number of sharing recommendations to a small set of documents (requirement R2). Furthermore, this small set should not just satisfy some particular queries, but a large organizational information need.

Thus, instead of using the concrete queries to represent an information need, we use the terms extracted from queries as an approximation for the aggregated organizational information need. Economically spoken, our goal is to determine those terms, which have the highest value for the organization from the perspective of the particular user of our system. Candidate documents shall be selected for recommendation based on this information. Therefore, we need to perform two steps:

- Select from the *terms* in the local documents those with the (from a subjective perspective) highest value for the overall organization.
- Based on this, select those currently private *documents* with the highest value for the overall organization.

Selection and Valuation of Terms. We compute the "value" of a term for the overall organization from two different concepts:

- First, those terms have to be selected which appear relatively more often in the private index than in the public index and thus have a large "organizational information gap" (OIG).
- In the second step, information from the query logs is used to compute the aggregate "organizational information need" (OIN).

Organizational information gap. The computation of interesting terms is quite straightforward. We are interested in those terms, which appear very infrequently or even not at all in the public index, but in the private index (denoted "organizational information gap" in section 2.2). Thus, we need both the global and the local document frequency for each term, normalize it by the total number of documents in each corpus, and rank the results by their difference. We call this the normalized document frequency of a term ($NDF(t)$). Its exact computation is described in [9]. However, the resulting value alone does not tell us so much, as lots of information about a special-interest topic in the private space of a user must not necessarily mean that this information is needed for the whole organization.

Organizational information need. Thus, we introduce the notion of organizational information need (see also section 2.1), which estimates the relevance of terms for the organization. In our framework, OIN is based on the user queries, which can be retrieved from the central query log (see also section 3.2). Those queries give us an impression of how much certain terms are sought. Besides the actual query strings, the query log stores the querying user, the NDF for the query terms in her local corpus, and a timestamp.

Our basic rationale for computing the OIN is that it is higher, 1) the more often and more recently a term has been part of a query, 2) the more different users used the term in a query and 3) the more seldom a term is in the local index of the users. Based on this, we define the OIN as the weighted sum of individual information needs of the querying users. Accordingly, we propose the following four measures as signals for an organizational information need:

Frequency. We assume that the OIN regarding a term is the higher, the more often it has been part of a query.

Freshness. Since the OIN regarding some term is a dynamic value, we also assume that the OIN is higher, if the term has been queried more recently. This allows recent organizational information needs to score a high value, if they are part of the set of recent queries Q_{recent} which contains all queries within the last 30 days and is a subset of the set of all queries Q.

Experience. We assume that the OIN is higher, if the querying user has a low experience. For measuring the experience, we propose the value of $NDF(t)_{user}$, since it tells us, how many documents match the term in the private space of the querying user. We assume a user is less interested in additional results, if she already has a number of documents in the private space.

Universality. We define that an OIN is the higher, the more different users queried for a term. The rationale behind this is that an information provider may only receive a limited set of sharing recommendations (see section 3.1). Thus, in order to maximize the overall benefit for the organization, such terms should be boosted, which are relevant for a large number of different information seekers.

In order to formally define the OIN, we group the first three signals into a personal information need. Thus, personal information need for a user consists of the frequency and freshness of queries and on the personal experience regarding a certain term. Since we want to penalize high experience, we include the inverted value of $NDF(t)$:

$$PIN(t, user) = \frac{1}{NDF(t)_{user}} \cdot (\frac{Q_{t,user}}{Q} \cdot (1 + \frac{Q_{t,user} recent}{Q_{recent}})) \tag{1}$$

Accordingly, the OIN is the sum of the values for PIN, normalized by the total amount of querying users:

$$OIN(t) = \frac{Users_t}{Users} \cdot \sum_{user_t} PIN(t, user) \tag{2}$$

Finally, the value of a term can be computed as the product of the organizational information gap, as based on the comparison of the public space with the private space of the user, and the organizational information need, as based on the analysis of the public query logs:

$$val(t, user) = OIG(t, user) \cdot OIN(t) \tag{3}$$

Since we do not assume any entity with a global view on all private information, $val(t, user)$ is always a subjective value from the perspective of a certain user. However, this is sufficient for our purpose of proposing a user to share certain documents.

Selection and Valuation of Documents. After identifying the most valuable terms in the private space of the user, we need to derive concrete documents that can be recommended to be shared. A straightforward solution would be to query the local index with the most valuable terms and suggest the top-ranked documents to the user for sharing. However, there are some obvious limitations with this approach, since the top results might cover certain terms redundantly. Due to space limitations we omit the presentation of a more appropriate algorithm at this place.

4 Evaluation

Evaluating our approach involves two steps – evaluating the existence of organizational *information gaps* and the existence of respective *information needs*. While previous work of ours has shown the existence of organizational information gaps – i.e. a significant number of terms for which information is hidden in the private spaces of users (see [9]), we concentrate on information needs in this paper. The study which we describe in the following was a direct follow-up of our previous study on information gaps.

Setup. As described in section 3.3, our approach for calculating organizational information needs requires query logs from an organization. Since such query logs were not available in our organization during the time of our study, we decided to estimate the OIN for a number of terms by using questionnaires.

Therefore, the author collected five meaningful terms from the top-ranked terms (high OIN) of the private indices analyzed in our previous work[3]. As *meaningful* we defined terms which are related to entities such as projects, person names, scientific topics or software tools.

The resulting 50 terms (5 terms from 10 indices) were compiled into a questionnaire, which was then distributed to members of our organization. This set of members was disjoint with the set of users of our analysis described in our previous work. For each term, the questionnaire provided four options to choose from: actively interested (user actively sought within the last six months or plans to seek information related to the term), passively interested (user is interested in information related to this term, without actively seeking for it), not interested (user is not interested in the term) and don't know (user does not know/understand the term or is not sure about interest).

Results. A total number of 12 users completed the survey. In average, across all 50 terms, 30% of the users selected "don't know", 37% selected "not interested", 20% selected "passively interested" and 13% selected "actively interested". Out of the 50 terms, 40 terms were marked "actively interested" by at least one user. 10 terms were selected "actively interested" and additional 16 terms "passively interested" by at least four respondents.

These results show that there is a significant interest for terms, which are more frequent in private information spaces than in the public information space of the organization. While our questionnaire did not ask if the respondents were able to satisfy their active information needs, we assume that they would be interested in results provided by their colleagues.

Clearly, our study suffers from the impreciseness and possible ambiguity of single terms. However, we believe that our selection of entity-related terms allows at least for a rough approximation of real information needs. Also, due to feasibility reasons, our questionnaire was based on a small and subjective selection of 50 terms. While we think that the analysis shows already some promising

[3] Out of the 13 indices analyzed in [9], we excluded three (UserA, UserF and UserJ), since their analysis did not yield a sufficient number of such meaningful terms.

opportunities for our solution, an implementation of the algorithms as described in this paper could discover many additional terms of active interest beyond those 50 terms.

5 Related Work

Knowledge management A large number of knowledge management or organizational memory systems (KMS/OMS) has been developed to support various aspects of knowledge management resp. knowledge sharing. Examples are groupware systems, document repositories, enterprise search applications, expert databases and more recent approaches such as Wikis (see e.g. [4,5]).

The main difference between those systems and our approach, is that *information providers* need to reveal their information first. In our approach, the *information seeker* must reveal her information need (the query) and potential knowledge providers can decide, based on concrete information about demand (OIN), if they want to share their information.

The Answer Garden system [19] is an exception, since potential contributors are triggered based on user information needs. However, Answer Garden does not aggregate those information needs and does not recommend to share actual documents (i.e. has no notion of private information spaces), but requires contributors to capture implicit knowledge.

Information retrieval. There is a number of works dealing with the notification of users when new information w.r.t. a certain query is available. A popular example for such a system is Google Alerts. In literature this is called *prospective search* [1], *continous querying* [20] or *retroactive answering of queries* [2]. These papers mainly concentrate on selecting relevant documents to existing queries. Thus, research in this area is synergetic to our approach, since we strive to diffuse information from private to public spaces. Once this diffusion happened, strategies from prospective search can be employed to notify interested users.

Reference [2] is particularly related to our paper. Similar to our notion of organizational information need, it defines *standing interest* as a measure for queries for which a user would be interested in future results. The authors discuss a number of indicators retrieved from query logs which partly overlaps with our calculation of OIN. However, they focus on deriving standing interest for a specific user, while our notion of OIN relates to a groups of users (c.f. 3.3).

P2P-based information retrieval approaches challenge traditional centralized search engines which do not scale naturally, hold centralized control and fail to provide access to the whole web. Systems such as Minerva or P-Grid[4] assume that peers carry out crawling and indexing tasks, maintain a local index, share parts of it, and provide services for other peers and searchers. Similar to our approach, peers publish statistical metadata (to the peer network).

However, all these approaches are based on the assumption of shared information, even if it does not reside on central machines. While this must not

[4] http://www.mpi-inf.mpg.de/d5/software/minerva/, http://www.p-grid.org/

mean that users have to hand out their results by default, they must provide metadata (i.e. index terms) to have the system working. In contrast to this, our approach does not require information providers to reveal any content they have but information seekers to reveal their information need.

6 Conclusion

In this paper, we presented *inverse search* – a novel approach that aims to foster knowledge sharing in organizations. It is designed to stimulate the diffusion of relevant documents from private information spaces of particular users to the public information space of an organization. Based on the notion of organizational information need (OIN), our approach recommends people to share private documents containing information relevant for other members of their organization.

Our paper makes two major contributions. First, we provide a mechanism for identifying documents that should be shared by deriving an organizational information need (OIN) from query logs and further usage statistics of our system. Second, our system provides means to foster the diffusion of information by recommending users to share private documents that cover such organizational information needs.

Based on the initial evaluations in section 4, we think that our approach can help to accelerate the diffusion of information in organizations. In particular, our approach tackles the requirements mentioned in section 3.1 as follows:

R1. Retain privacy. In our model, not the information provider, but the information seeker must expose her information. The content of documents in the private information space stays under full control of its owner – it is not part of a central index.

R2. Minimize effort. By relying on queries from a search system, there is almost no initial effort for the information provider. Only the later sharing itself might require action. However, this can be supported by restricting recommendations for sharing to a small number of highly relevant documents.

R3. Motivate to share. Our model of calculating the value of private, local information gives the user feedback about the relevance of her information. She can decide herself to share this information. Additionally, the system can be complemented by mechanisms to reward sharing information, which is heavily sought in the network, but scarcely available.

References

1. Irmak, U., Mihaylov, S., Suel, T., Ganguly, S., Izmailov, R.: Efficient query subscription processing for prospective search engines. In: WWW 2006: Proceedings of the 15th international conference on World Wide Web, pp. 1037–1038 (2006)
2. Yang, B., Jeh, G.: Retroactive answering of search queries. In: WWW 2006: Proceedings of the 15th international conference on World Wide Web, pp. 457–466. ACM Press, New York (2006)

3. Olson, G.M., Olson, J.S.: Distance matters. Human-Computer Interaction 15(2/3), 139–178 (2000)
4. Alavi, M., Leidner, D.E.: Review: Knowledge management and knowledge management systems: Conceptual foundations and research issues. MIS Quarterly 25(1), 107–136 (2001)
5. Maier, R.: Knowledge Management Systems. Springer, Heidelberg (2003)
6. Small, C.T., Sage, A.P.: Knowledge management and knowledge sharing: A review. Information, Knowledge, Systems Management 5(3), 153–169 (2006)
7. Mooradian, T., Renzl, B., Matzler, K.: Who trusts? personality, trust and knowledge sharing. Management Learning 37(4), 523–540 (2006)
8. Baeza-Yates, R., Riberio-Neto, B.: Modern IR. ACM Press, New York (1999)
9. Happel, H.J., Stojanovic, L.: Analyzing organizational information gaps. In: Proceedings of the 8th Int. Conference on Knowledge Management, pp. 28–36 (2008)
10. Hansen, M.T.: The search-transfer problem: The role of weak ties in sharing knowledge across organization subunits. Adm. Sci. Quarterly 44, 82–111 (1999)
11. Davenport, T.H., Prusak, L.: Working Knowledge. HBS Press (1998)
12. Cress, U., Hesse, F.W.: Knowledge sharing in groups: experimental findings of how to overcome a social dilemma. In: ICLS 2004: Proceedings of the 6th international conference on Learning sciences, pp. 150–157 (2004)
13. Cabrera, A., Cabrera, E.F.: Knowledge-sharing dilemmas. Organization Studies 23, 687–710 (2002)
14. Wasko, M.M., Faraj, S.: Why should i share? examining social capital and knowledge contribution in electronic networks. MIS Quarterly 29(1), 35–57 (2005)
15. Ardichvili, A., Page, V., Wentling, T.: Motivation and barriers to participation in virtual knowledge-sharing communities of practice. Journal of Knowledge Management 7(1), 64–77 (2003)
16. Desouza, K.C.: Barriers to effective use of knowledge management systems in software engineering. Commun. ACM 46(1), 99–101 (2003)
17. Desouza, K.C., Evaristo, J.R.: Managing knowledge in distributed projects. Commun. ACM 47(4), 87–91 (2004)
18. Orlikowski, W.J.: Learning from notes: organizational issues in groupware implementation. In: CSCW 1992: Proceedings of the 1992 ACM conference on Computer-supported cooperative work, pp. 362–369. ACM Press, New York (1992)
19. Ackerman, M.S., Malone, T.W.: Answer garden: a tool for growing organizational memory. In: Proceedings of the ACM SIGOIS and IEEE CS TC-OA conference on Office information systems, pp. 31–39. ACM, New York (1990)
20. Kukulenz, D., Ntoulas, A.: Answering bounded continuous search queries in the world wide web. In: WWW 2007: Proceedings of the 16th international conference on World Wide Web, pp. 551–560. ACM Press, New York (2007)

Representing and Retrieving Knowledge Artifacts

Rosina Weber, Sid Gunawardena, and George Abraham

The iSchool at Drexel, College of Information Science & Technology, Drexel University
{rweber,sidath.gunawardena,george.abraham}@ischool.drexel.edu

Abstract. This paper recommends a structure to represent and a method to retrieve knowledge artifacts for repository-based knowledge management systems. We describe the representational structure and explain how it can be adopted. The structure includes a temporal dimension, which encourages knowledge sharing during knowledge creation. The retrieval method we present is designed to benefit from the representational structure and provides guidance to users on how many terms to enter when creating a query to search for knowledge artifacts. The combination of the structure and the retrieval method produces an adequate strategy for knowledge sharing that guides targeted users toward best results.

1 Introduction

Repository-based knowledge management systems (KMS) are a typical knowledge management (KM) initiative employed in today's organizations to promote knowledge sharing between their members. The knowledge to be shared is retained in knowledge artifacts [1], such as lessons-learned, alerts, or best practices. These systems became popularized with the implementation of text databases [2]. Text databases allow contributors to enter knowledge artifacts in a free-text form and can be searched by members seeking knowledge. Unfortunately, such organizations usually lack proper understanding about implementing a strategy for knowledge sharing. Simply making searchable text databases available is not enough to foster knowledge sharing [3].

Consider an example where a scientist has learned that in a commercial airplane, pathogens spread towards the sides and the back–suggesting that passengers in aisle seats in the front will be less likely to become in contact with pathogens spread during a flight. In order to share this knowledge in a KM context, a contributor needs to be guided with answers to questions such as, "Is it the kind of knowledge my organization wants me to share?, "Is it in the right format?", Is it complete or is there anything missing?, "Is it at the right level of specificity?", etc. Providing those answers to KMS contributors has been discussed as a managerial responsibility that organizations have to enforce when implementing a knowledge sharing strategy [4]. Further examples in support of knowledge sharing can be found in, e.g., [5][6][7][8].

Following recommendations in the literature, the contributor of the knowledge in the above example would, for instance, be asked to include the process in which such knowledge becomes relevant, i.e., when booking air tickets. Another structural element to promote sharing would be to include a justification or description of how

T. Yamaguchi (Ed.): PAKM 2008, LNAI 5345, pp. 86–97, 2008.

such knowledge was learned, so that other members would be able to determine its validity and decide whether or not reusing it.

This paper advocates the principle that part of this guidance on what and how to submit knowledge artifacts can be embedded in a structure to represent them. The use of the structure is augmented by a reviewing step that provides feedback to contributors helping to educate them about the potential benefits of following proper guidance. We define a structure for artifacts and describe such a structure on a conceptual level that is sufficiently general to be adopted by different target communities in different domains. We also introduce a method that complements the structure to retrieve these knowledge artifacts. We show results from the usage of the representation and a study of the retrieval method.

2 Background and Motivation

This paper extends work that originally focused on lessons-learned, whose concepts are also useful to describe other categories of artifacts, leading to an approach that is general to all. A definition of a lesson-learned was proposed by Secchi, Ciaschi, and Spence [9], which revealed its core concepts. Their definition includes that it is knowledge or understanding that is learned, that it must be validated, and that it has to identify a process or decision where it can be applied [9].

The discussions at the 2000 AAAI Workshop on Intelligent Lessons Learned Systems [10] (see for example, [11] and [12]) focused on the use of intelligent methods to reason with and to retrieve knowledge artifacts. The survey in [3] revealed that the majority of systems used free-text fields to represent knowledge artifacts. It also uncovered the need to include and emphasize the strategy to be learned in fields labeled with names such as *lesson* or *recommendation*. Nevertheless, such a labeled field is usually used as a reminder that a recommendation is expected but there is no reviewing process in place to verify that the artifact meets any necessary requirements.

A preliminary version of the structure presented in this paper for knowledge artifacts was introduced by Weber and Aha [13]. The purpose of that structure was to represent lessons-learned, and to reason with and retrieve them using case-based reasoning. In its essence, it is the same structure that we present in this paper. The difference is that now we have a deeper understanding of its generality and can recommend how it can be used by and adapted to communities from multiple domains. Most importantly, we understand how the formatting of the contributions entered using the structure should be controlled to be complete, effective, and to avoid long texts, which are hard to review, read, and interpret. In this sense, it is controlled because constraints are imposed to the format of contributed artifacts. As a result, the concepts we describe now are more general and more complete. Furthermore, we now have experiences with the use of the structure and have devised a method to retrieve knowledge artifacts represented with it.

The result of the excessive use of free-text forms without guidance in repository-based KMS is collections of knowledge artifacts that are seldom reused. These artifacts may lack vital contents and their unstructured nature can result in difficulties in comprehension. There are many reasons indicating that free-text artifacts are responsible for failure in KMS (e.g., [3], [6]). Nonetheless, we advocate that the most

important aspect is how KMS ignore organizational responsibilities for knowledge capture, as discussed by Marshall, Prusak and Shpilberg [4], which are crucial for knowledge sharing. The structure proposed in this paper is aimed at enforcing organizational responsibilities such as guiding contributors on what is to be shared to benefit knowledge sharing.

It may seem that proposing constraints to control the format of artifacts could cause a burden on the contributors whose freedom would be limited. However, our experience has shown that the majority of users are relieved by our concern with not retaining artifacts of poor quality. This resonates with one of the appointed causes of failure in KM approaches that claims that users are not motivated to contribute to a system where they see no value [7]. As a result, our concern with quality of artifacts, even if through some form of control, could potentially increase the confidence users have in the usefulness of the final collection.

3 Structure to Represent Knowledge Artifacts

The time taken by each community member to produce knowledge can vary. Recognizing this variation, we propose a representation for knowledge artifacts that incorporates a temporal dimension associated with the creation of knowledge. We recommend that members share their efforts that are in progress, rather than wait until they are completed and new knowledge is learned. This is particularly useful for communities with tasks spanning long periods.

We first explain concepts in the structures for artifacts that capture artifacts describing learned knowledge, i.e., completed tasks. Next we describe the structure for artifacts that capture the knowledge generation, what we label *in progress* artifacts, i.e., in progress tasks. We also describe how to control their formatting. Then, we present some evidence indicating the usability of the structure, illustrating its benefits for KM tasks.

Table 1. Representation structure for learned knowledge artifacts

Description	Labels	Purpose
It must declare what it teaches	Contribution	Reusable elements
It must state how it was learned	Rationale	
It must explain its usefulness in general terms	Applicable task	Indexing elements
It must explain its usefulness in specific terms	Contexts	

3.1 Elements of the Controlled Structure for Learned Artifacts

The proposed structure for knowledge artifacts consists of four core fields (Table 1). These fields were developed by taking into account the main purpose of knowledge artifacts–knowledge sharing. Consequently, the first field we discuss is the one designated to contain the knowledge to be shared: Contribution.

Contribution. This is the strategy or lesson to be shared, what a contributor learned and believes may be useful to other members of the community. It is a strategy to be either reused or it may even be on to avoid. Part of the concept of the Contribution

field is the concept of singleness of an artifact, where each artifact's scope is limited to presenting one strategy. Contribution can be used as a reference to the scope of a knowledge artifact and can also help identify its most suitable level of specificity. The contribution is to be singular in nature and hence it should be communicated by a single statement. Exceptions are granted when a contribution entails technical specifications that may require additional sentences. One rule-of-thumb to identify a contribution is to think of mentioning it and acknowledging its authorship. For example, for the contribution, "*White light is heterogeneous and composed of colors that can be considered primary*" is usually followed or preceded by a statement that it was discovered by Newton. Another form of recognizing the unity in a knowledge artifact is to think of its rationale. Generally, a scientific experiment is designed to demonstrate one hypothesis; on occasion, a single scientific experiment will produce a set of contributions with multiple results. If results had to be repeated to break down the contribution, then this contribution should include all this set. Finally, the unity of a contribution may also be guided by its applicability, which should also be singular. Strategies in contributions are meant to be applicable in an activity or process. Therefore, there should be one single process where it should be applicable. For example, *installing speakers* is a process where a contribution can be applied. Although we would not want to state a contribution that may be useful for more than one process like *installing and selling speakers*, because this would not correspond to a single contribution and would require more than one explanation.

The input into this field should be controlled in length. Though a crisp bound is not defined, the goal is to stick to the contents covered by the guidelines described above. Our experience reveals that 30 or 40 words represent the average, though in a few exceptional cases the text contained close to 300 words. We recommend that humans review these cases and accept exceptions with caution. Long texts are difficult to interpret and may hinder sharing. In case contributors want to include details or background, they can use an additional field that is not a core item in the structure. Nonetheless, excessive background is not necessary because knowledge artifacts are to be communicated to members of the same community, who have some knowledge of the domain.

Rationale. This component of the representation provides an explanation that addresses one of the concerns posed by Szulanski [5] about users being unable to reuse an artifact because they ignore if it is valid. The rationale varies depending on the nature of the contribution. For example, in the Navy Lessons Learned System, one artifact describes a contribution that was learned as advice received from someone else. Advice would be then a type of rationale. Alternative types would be failure or success, meaning that a strategy could have been learned because it was attempted (or not) and the result was success (or failure). A scientific community will obtain results either through quantitative or qualitative methods to support a contribution. A combination of arguments may be used in philosophical contributions, just as in argumentative text.

Control of the Rationale field should follow the same principles adopted to review the Contribution field. The event or source that supports the contribution should correspond to statements in the Contribution field in that the contents of the Rationale field gives the basis for the Contribution field. Descriptions of methodology and experimental design, although relevant, are not meant to be described in Rationale; but results of an event or experiment that substantiate the contribution to be shared.

The two fields Contribution and Rationale belong to the category of reusable elements [13]. They are necessary because they inform the users of the KM system of knowledge they may want to reuse. These fields also provide evidence explaining how the contribution was learned, and help users decide whether to reuse it or not. For effective retrieval of knowledge artifacts, the structure also includes two fields that are indexing elements: Applicable Task and Contexts.

Applicable Task. This is the general activity that one needs to be engaged in for the contribution of a knowledge artifact to become applicable, e.g., installing speakers. While submitting new artifacts, contributors should try to visualize themselves in a position where they could benefit from learning the knowledge they want to convey. Having learned such knowledge warrants them enough understanding to conjecture on what tasks their contribution is applicable.

The Applicable Task is stated in general terms. It is broad enough that it is likely to repeat in the same collection many times. To facilitate its identification, it is limited to only two expressions. The easiest way to control input to the Applicable Task is by identifying verbs and complements that are typical of a community and using drop down lists for user selection. Although the labels of the two dimensions of the Applicable Task will vary depending on the domain, we recommend that the first be a verb and the second be a complement. Of course, not all verbs would be adequate. According to Levin's categorization of verbs [14], we recommend verbs of neutral assessment (e.g., analyze, evaluate) because they have the connotations of exercising actions that are likely to produce a process. These verbs are obviously transitive, so the complement will define the domain the action will impact. The second element of Applicable Task, the complement, is usually a domain-specific term. In the Applicable Task *installing speakers*, *speakers* are associated to the domain because they represent a product sold and installed by a specific organization. It is where the verb's action will be done.

Contexts. The Contexts field is responsible for giving specificity to the Applicable Task. The Applicable Task field presents the general task or process that will match many instances of situations in which community members will be engaged. In contrast, the Contexts field narrows down that generality to help match a description to the usefulness and applicability of the strategy in the Contribution field and exclude contexts that are not applicable. In this sense, Contexts can be filled in with a list of terms that would be potentially used in different situations where the strategy would be applicable. Another way of interpreting Contexts is as a list of state variables. Consider the Applicable Task above *Installing speakers* and think of the contexts of that task and potential variables that could be assigned values. For example, *location* would be a variable and could be assigned values such as home, office, or car; a variable *method* could be assigned values manual or automatic; whereas *type* could be tweeter, back, side, or subwoofer. The ideal list of contexts we are interested in is one that has at least one corresponding value to each variable. The specificity of the artifact would be characterized, for example, by knowledge applicable to manual installation and not applicable to automated installation of speakers; which is useful for home installation and not for in-car installation.

The control of this field is minimal. We recommend keeping a list of terms or expressions rather than full sentences. There are certain situations in which some variables could be treated differently. An example would be to recommend members of a microbial project to enter values for agents as *agent is norovirus*, rather than the word *norovirus*.

Table 2. Representation structure for artifacts in progress

Description	Labels	Purpose
Declare what one is trying to learn	What do you expect to learn?	Reusable elements
State what will be done to learn it	How do you plan to learn it?	
Explain its usefulness in general terms	Applicable task	Indexing elements
…and specific terms	Contexts	

3.2 Elements of the Controlled Structure for Artifacts in Progress

The structure of *in progress* artifacts is laid out in Table 2. The purpose of artifacts that are in progress is to anticipate the timing of contribution, thus increasing the opportunity of sharing.

As presented in Table 2, only the Reusable Elements change by attempting to capture knowledge generation. Rather than asking what someone learned we ask what they expect to learn, like a hypothesis. Similarly, we ask how they plan to learn it instead of asking for results. The way to control them is also analogous to the completed forms.

The decision of the suitability of sharing knowledge artifacts that are in progress should be a community wide decision, based on its potential benefits. For example, in a domain in which this representation is applied, a distributed community of scientists, we learned that sharing in progress knowledge artifacts has many benefits. Interests of members are made available to the community earlier than it would if everyone waited for a completed research, as one may take several months, or even years before it produces an innovation and can be considered completed. Sharing in progress work may help identify collaborations that would otherwise be missed.

Another important benefit to scientific communities is in the question of how members plan to learn a contribution. The benefit is that sharing *in progress* artifacts gives an opportunity to describe the experimental design (or methods) used to reach results that substantiate a contribution; particularly because a substantial portion of the knowledge learned by scientists is embedded in the refining of hypotheses and their respective experimental designs. This may seem to be missing from the completed artifacts structure, but we find that it would make that representation too long to enter and read. This approach that combines this temporal dimension allows us to meet all our objectives in support of knowledge tasks.

3.3 Results from Usage

The most comprehensive application of the proposed structure to represent knowledge artifacts is with a community of scientists, described in PAKM 2006 [15]. This section presents analyses of results from this implementation and attempts to relate it to some of our expected results.

Sufficiency. One metric of quality relates to the structure's ability to represent contents that are in the minds of contributors. Once they understand the concepts of the representation, it is our expectation that they will find the structure sufficient to capture all contents they have in mind. In order to assess how well the proposed structure is meeting this expectation, we computed the number of domain specific expressions in 177 knowledge artifacts (the first 2 years worth of submissions) for the system described in [15]. We then compared that number with the number of domain-specific terms that appeared in the four core fields and the title. The title was used because it is not a core field for the structure but it was requested by users given that it is a habit that may help them organize their ideas. Our assumption is that whenever domain-specific terms show up in a title and not in the core fields, the structure fails to provide sufficient concepts for contributors to communicate knowledge artifacts. For the artifacts contributed, we found that 99.5% of terms in the artifacts can be found within the core fields of the representation; implying that contributors did not find a proper field for only 0.5% of the terms. We will observe this metric in next years to determine consistency.

Temporality. The expectation of adopting knowledge artifacts that vary along the temporal dimension is that it allows knowledge to be shared before it would otherwise be available to the community. More specifically, our assumption is that in progress artifacts would make knowledge available for sharing before completed artifacts. In the first two years, while 22% of the contributed artifacts were completed, 64.4% were in progress. Adopting this temporal dimension provided for roughly three times more knowledge for sharing than it would have been possible without it.

Knowledge Sharing. At the end of the submission of a knowledge artifact, the contributor is asked to search existing knowledge artifacts and indicate, by creating an association between two knowledge artifacts, relationships that only experts can recognize. Our assumption is that, in a community of scientists, a contributor has to understand an existing artifact to be able to recognize an association. When this happens, we are comfortable to claim that knowledge in the existing artifact was shared with this contributor. Consequently, we interpret associations between artifacts contributed by different authors as evidence of knowledge sharing. In the first 2 years, 93 associations were made, 71 between units entered by different authors, for a total of 177 artifacts. This represents that there was knowledge sharing activity that is equivalent to about 40% of the effort in contributing artifacts.

4 Retrieving Knowledge Artifacts

Information retrieval (IR) methods are used to select a subset of records from a collection that are relevant to a given query. Poor IR performance in repository-based KMS is an impediment to knowledge sharing. It may be caused by a) lack of knowledge about the format of the records, b) lack of knowledge of the domain of records, and c) poor query construction. The adoption of the structure presented in Section 3 addresses the lack of knowledge about the format by defining it. It addresses lack of domain knowledge by keeping domain-specific expressions in the fields Applicable Task and Contexts. This allows the construction of a domain-specific taxonomy,

which can be used, e.g., to resolve ambiguities in query expansion. Finally, it addresses poor query construction by influencing its length with a recommended cardinality factor: RCF [16].

4.1 Recommended Cardinality Factor: RCF

The distinguishing characteristic of the retrieval method is the RCF. The RCF becomes necessary due to the variable number of terms that may be included in the artifacts. This can be problematic for retrieval quality; therefore, we counteract it by computing the ideal length of a query that leads to a better retrieval performance, which is found from the cardinality of individual artifacts in the repository. In [16] we demonstrated the performance of Equation (1) to compute the RCF from the averaged number of Terms per Artifact (TpA) in a repository,

$$RCF = \mathrm{e}^{(0.833 \times \ln(TpA))} \quad . \tag{1}$$

When RCF is used in search, it bounds the comparison between a query and each artifact. The actual comparison between terms may be carried out as described in [16], or via cosine, n-grams, etc. This parameter limits the number of comparison to improve retrieval quality. In other words, if the artifacts have, on average, a large number of terms, submitting a query with few terms will not produce a retrieval of the same quality as a query with a greater number of terms. Next we describe a comparison study with a comparative IR method that does not utilize the RCF.

4.2 Comparison Study: IR vs. RCF

This study utilizes a dataset used in a survey of users of the KMS discussed in [15]. The survey consisted of 6 queries with hypothetical results for the users to score whether they considered them *relevant*, *somewhat relevant* or *not relevant*. Queries and results were designed from actual knowledge artifacts. Each query is created from the Contexts field of an artifact, for example: "modeling, aerosol dispersion, indoor air," and the query result would be the contents of a different artifact. The survey results produced 16 *query result pairs*. For this study, we use only the pairs that were consistently assessed as relevant and somewhat relevant. The users consistently labeled only one query result as *not relevant*, so it was excluded from the analysis.

This study hypothesizes that a search method that adopts the bounding parameter RCF will produce better quality in retrieval than an alternative IR method. We adopt an average performance computed as follows. For each *query result pair*, consisting of a result to a query and a score, we compute the proportion of results that are selected by each method to be in the top n results of the retrieved set. We compute n from 1 to 10, and then average each of those results, for both the following methods: RCF and IR.

For RCF, each value of the query is compared against each value in the knowledge artifacts, assigning 1 for matching values or 0 otherwise. The results are added until either (a) the number of matches meets or exceeds the RCF or (b) there are no more terms to match in either the query or the knowledge artifact. In (a), a score of 1.0 is assigned; in (b) the score is given by the number of matches divided by the RCF. In other words, the RCF is the number of terms that have to match in order to consider a knowledge artifact to completely match a query.

Table 3. For n =1 to 10, the proportion in which a result appeared in the retrieval set

Results rated *relevant* or *somewhat relevant* in survey											
n	1	2	3	4	5	6	7	8	9	10	AVE
IR	0.38	0.62	0.62	0.69	0.77	0.92	1.00	1.00	1.00	1.00	0.80
RCF	0.67	0.80	0.87	0.93	0.93	0.93	0.93	1.00	1.00	1.00	0.91

Or the IR, we used the Indri search engine of the Lemur Toolkit a language modeling IR tool freely available on the web [18]. Indri represents documents as multiple binary feature vectors.

Results and Discussion. Table 3 shows the average and the individual results for each value of n. Thus, for instance, when considering the top 5 terms (n = 5), the table shows what proportion of the test results were presented by each method (i.e., higher is better). The results are more distinguished at the levels where fewer results considered, i.e., 1 to 5. These are also more strict assessments, making the better values when the RCF is used, more valuable.

A two tailed t-test performed on the paired rankings of the query results for both methods showed that there was a significant difference between the IR and RCF methods at $p < 0.1$ (0.072). Moreover, the RCF provides guidance to users. Most users of KMS draw their mental models of search from their interactions with web search engines. Spink et al. [17] show that the average number of terms in web search queries is around 2.5 terms. Knowing the optimal number of terms to include in a search query can help users create better queries and thus give them a better chance at finding relevant results.

5 Related Work

In this section we leverage discussions from design, human-computer interaction, and KM to argue for the persistence and contribution of structure behind knowledge artifacts. Some broad goals of such structures, in these different disciplines, are to promote communication and comprehension. There appears to be considerable overlap between the goals mentioned in other disciplines and what a knowledge management system attempts.

The pattern concept has received much attention from various design communities (e.g., software engineering, HCI, education), and each have adapted and repurposed it for suiting what they do. At some level, these different design disciplines agree on patterns as a representation or structure to share best practices. Alexander et al. [19], who are attributed for introducing the pattern concept in architecture, argue patterns may help designers and non-designers communicate. Another shared, or implicitly agreed upon property of a pattern, is structuring design guidance or best practices as a *... three part rule, which expresses a relation between a certain context, a problem and a solution*. ([20]; p. 247). In addition to these three elements, like our structure, a pattern description also contains a rationale that argues why it is a good solution in the given context.

The EUREKA system at Xerox is often described a win for KM [21]. Bobrow and Whalen point out that their system emerged by learning how Xerox technicians actually share experiences. The experiences were not presented as knowledge per se, but according to Orr's [22] ethnographic study of technicians, as war-stories. Bobrow and Whalen explain that the experiential knowledge, that is embedded in practice, was being exchanged through such rich stories. Each time the story was narrated it was enhanced or contextualized based on the given situation [22]. Understanding the nature and role played by these war-stories showed Xerox a way to structure the experiential knowledge in the EUREKA system. This structure or tips contained a symptom, cause, test and action.

The proposed structure has similarities with structures used in other disciplines. We assert that the elements of the structure described in Section 3 can be mapped, with little effort, into a problem (i.e., Applicable Task and Contexts) and solution (i.e., Contribution and Rationale). Nonetheless, it is novel in how it is presented here; particularly on the use of two fields with different levels of specificity to help discriminate a problem. When we argue that our structure is generalizable to other similar KM efforts, we are referring to our structure in principle. We encourage readers to customize this structure for their specific audience as it is difficult to argue for a one-size-fits-all approach.

6 Conclusions

This paper discusses a structure to represent and a method to retrieve knowledge artifacts in repository-based KMS. Section 3 describes the concepts of the structure for adoption in multiple domains. The recent implementation of this structure indicates that it is sufficient to capture contents that contributors want to share. The approach recommends the incorporation of a temporal dimension to the structure. Its use suggests that it makes knowledge available for sharing sooner than it would without such a dimension. The ease of interpreting a knowledge artifact encourages contributors to make associations between new and existing artifacts. Those associations convey evidence of knowledge sharing.

Section 4 presents the method to retrieve knowledge artifacts that adopts such a structure. The method is characterized by the use of a parameter, RCF, which determines the number of terms that are supposed to match for a knowledge artifact to be considered relevant to a query. A study demonstrates the superiority of this method when compared to an IR method without it.

An essential part of any KM system is the contributors and users of these knowledge artifacts; most importantly whether the users accept it [2]. When approaching KM from the users', or even use, perspective, questions about comprehension and usability of the structure and the knowledge artifacts should receive high priority. We believe that our proposed structure has the potential to answer some of these questions.

Acknowledgements

First and second authors are supported in part by the U.S. EPA-Science to Achieve Results (STAR) Program and the U.S. Department of Homeland Security Programs, Grant # R83236201. The testing of the search is being conducted under IRB protocol #16449.

References

[1] Holsapple, C.W., Joshi, K.D.: Organizational Knowledge Resources. Decision Support Systems 31(1), 39–54 (2001)
[2] Davenport, T.H., Prusak, L.: Working knowledge: How Organizations Manage what they Know. Harvard Business School Press, Boston (1998)
[3] Weber, R., Aha, D.W., Becerra-Fernandez, I.: Intelligent Lessons Learned Systems. Expert Systems with Applications 20(1), 17–34 (2001)
[4] Marshall, C., Prusak, L., Shpilberg, D.: Financial Risk and the Need for Superior Knowledge Management. California Management Review 38(3), 77–101 (1996)
[5] Szulanski, G.: Exploring Internal Stickiness: Impediments to the Transfer of Best Practice within Firms. Strategic Management Journal 17, 27–44 (special issue, 1996) (Winter)
[6] Atwood, M.E.: Organizational Memory Systems: Challenges for Information Technology. In: 35th Annual Hawaii International Conference on System Sciences. IEEE Press, Los Alamitos (2002)
[7] Disterer, G.: Individual and Social Barriers to Knowledge Transfer. In: 34th Annual Hawaii International Conference on System Sciences. IEEE Press, Los Alamitos (2001)
[8] Weber, R.O.: Addressing Failure Factors in Knowledge Management. Electronic Journal of Knowledge Management 5(3), 333–346 (2007)
[9] Secchi, P., Ciaschi, R., Spence, D.: A Concept for an ESA Lessons Learned System. In: Secchi, P. (ed.) Alerts and LL: An Effective way to prevent failures and problems (Technical Report WPP-167), ESTEC, Noordwijk, pp. 57–61 (1999)
[10] Aha, D.W., Weber, R. (eds.): Intelligent Lessons Learned Systems: Papers from the AAAI Workshop (Technical Report WS-00-03). AAAI Press, Menlo Park (2000)
[11] Ashley, K.D.: Applying Textual Case-based Reasoning and Information Extraction in Lessons Learned Systems. In: Aha, D.W., Weber, R. (eds.) Intelligent Lessons Learned Systems: Papers from the 2000 Workshop (Technical Report WS-00-03). AAAI Press, Menlo Park (2000)
[12] Everett, J.O., Bobrow, D.G.: Resolving Redundancy: A Recurring Problem in a Lessons Learned System. In: Aha, D.W., Weber, R. (eds.) Intelligent Lessons Learned Systems: Papers from the AAAI Workshop (Technical Report WS-00-03), pp. 12–16. AAAI Press, Menlo Park (2000)
[13] Weber, R.O., Aha, D.W.: Intelligent Delivery of Military Lessons Learned. Decision Support Systems 34(3), 287–304 (2003)
[14] Levin, B.: English Verb Classes and Alternations: A Preliminary Investigation. University of Chicago Press, Chicago (1993)
[15] Weber, R.O., Morelli, M.L., Atwood, M.E., Proctor, J.M.: Designing a Knowledge Management Approach for the CAMRA Community of Science. In: Reimer, U., Karagiannis, D. (eds.) PAKM 2006. LNCS (LNAI), vol. 4333, pp. 315–325. Springer, Heidelberg (2006)
[16] Weber, R.O., Gunawardena, S., MacDonald, C.: Horizontal Case Representation. In: Althoff, K.-D., Bergmann, R., Minor, M., Hanft, A. (eds.) ECCBR 2008. LNCS (LNAI), vol. 5239, pp. 548–561. Springer, Heidelberg (2008)
[17] Spink, A., Wolfram, D., Jansen, M., Saracevic, T.: Searching the Web: The Public and their Queries. Journal of the American Society for Information Science and Technology 52, 226–234 (2001)
[18] Lemur Project. Lemur Language Modeling Toolkit. University of Massachusetts and Carnegie Mellon University, `http://www.lemurproject.org/`

[19] Alexander, C., Ishikawa, S., Silverstein, M., Jacobson, M., Fiksdahl-King, I., Angel, S.: A Pattern Language: Towns, Buildings, Construction. Oxford University Press, New York (1977)
[20] Alexander, C.: The Timeless Way of Building. Oxford University Press, New York (1979)
[21] Bobrow, D.G., Whalen, J.: Community Knowledge Sharing in Practice: The Eureka Story. Journal of the Society for Organizational Learning 4(2), 47–59 (2002)
[22] Orr, J.E.: Talking About Machines: An Ethnography of a Modern Job. ILR Press, Ithaca (1996)

Extracting Advantage Phrases That Hint at a New Technology's Potentials

Risa Nishiyama, Hironori Takeuchi, Tetsuya Nasukawa, and Hideo Watanabe

Tokyo Research Laboratory, IBM Research
1623-14 Shimo-tsuruma, Yamato-shi, Kanagawa, 242-8502 Japan
{lisa,hironori,nasukawa,hiwat}@jp.ibm.com

Abstract. Technical literature such as patents, research papers, whitepapers, and technology news articles are widely recognized as important information sources for people seeking broad knowledge in technology fields. However, it is generally a labor intensive task to survey these resources to track major advances in a broad range of technical areas. To alleviate this problem, we propose a novel survey assistance tool that focuses on a novel semantic class for phrases, advantage phrases, which mention strong, advantageous points of technologies or products. The advantage phrases such as "reduce cost," "improve PC performance," and "provide early warning of a future failure" can help users to grasp the capabilities of a new technology and to come up with innovative solutions with large business values for themselves and their clients. The proposed tool automatically extracts and lists up those advantage phrases from large technical documents, and places the phrases that mention novel technology applications high on the output list. The developed prototype of the tool is now available for consultants analyzing patent disclosures. In this paper, a method to identify advantage phrases in technical documents and a scoring function to give a higher score to novel applications of a technology are proposed and evaluated.

1 Introduction

It is important for people who are responsible for the technical strategy of a company or a research organization to grasp an emerging technology area. Applications of new technologies have capabilities of enabling innovative products and services, and understanding potential capabilities of those new technologies is valuable for preparing available technology seeds to enhance their business. In order to obtain such knowledge, they need to survey technical literature such as research papers, whitepapers, and technology news articles. Since the technical literature is becoming easily accessible for people due to the spread of the Web, it is natural and actually a common method to examine those resources related to a technology area of interest.

One of the most important requirement of those surveys is time efficiency. This is particularly important for consultants who are required to have knowledge of a broad range of technology areas to match the business needs of clients to the available technical capabilities. However, search engines, which are the primary information retrieval tool for the huge amount of information on the Web, are

T. Yamaguchi (Ed.): PAKM 2008, LNAI 5345, pp. 98–110, 2008.

imperfect for this type of survey. The intrinsic function of a search engine is to accept a query and to return a list of webpages that consists of the titles and snippets of each webpage. In a case of the survey where a consultant seeks to know the capabilities in a new technology area, the consultant inputs keywords that represent the technology field (e.g. "Web 2.0") to a search engine. From the titles and the snippets on the obtained lists, the users can estimate which webpages they should read for the survey, but they still need to read the contents of those webpages to grasp the wide range of usage of the technology. Such a survey process using a general search engine is clearly time consuming.

Based on the analysis of around 300,000 patent disclosures by those consultants, we propose a novel information retrieval tool to support a technical document survey that aims at obtaining broad knowledge in targeted technology fields [1]. This tool identifies and lists up a novel semantic class of phrases called advantage phrases. This class includes phrases like "reduce cost," "improve PC performance," "provide early warning of a future failure" and other phrases mentioning a strong point of a new technology or a product. These phrases are key phrases of the original technical documents and give the users an idea of how they can actually apply the technology for their own needs.

To organize the list of advantage phrases, we also propose a scoring function for the extracted advantage phrases. Since various types of advantages are described for each patent or product, many advantage phrases are extracted from each patent disclosure or product release. This means that size of the output advantage phrase list tends to be larger than the size of the document set and this requires the tool to have some metrics to organize the advantage phrase list so that the most informative phrases appear at the top of the list. For the consultants, who are the main users of this tool, the most informative phrases are the ones mentioning novel applications of the input technology. Listing the novel applications of the focused technology is useful for brainstorming new applications with large business values.

The prototype of the proposed tool, called CAPHMIT (CApability PHrase MIning Tool), is now available for consultants. In the prototype, three natural language processing and information retrieval technologies are employed:

1. Advantage phrase extraction from the original technical documents
2. Document retrieval to select documents related to the input technology field
3. Scoring the extracted phrases to generate the output advantage list

This paper mainly describes the advantage phrase extraction and the advantage phrase scoring techniques. The document retrieval is done with string search in this study. The tool accepts a string query that represents an arbitrary technology field, and then selects documents that include the string as being related to the input technology field[1].

This paper first gives a detailed system overview, then defines the advantage phrases that we focus on. Next, the proposed method to identify the advantage

[1] When using patent disclosures as technical documents, another promising approach would be to use the International Patent Classification provided by the patent office.

phrases from the technical documents is described along with the scoring function to find phrases mentioning novel applications of the technology. The effectiveness of the phrase extraction methods and scoring functions are shown in discussions of the results. Finally, this paper concludes by introducing related work.

2 System Overview

Figure 1 shows an overview of CAPHMIT. Technical documents on the Web or in a local database are first processed by the Advantage Phrase Annotator. This module automatically identifies and annotates advantage phrases in the original patent disclosures, product releases and various other types of technical documents. The annotated documents are then stored in a database for run-time processing.

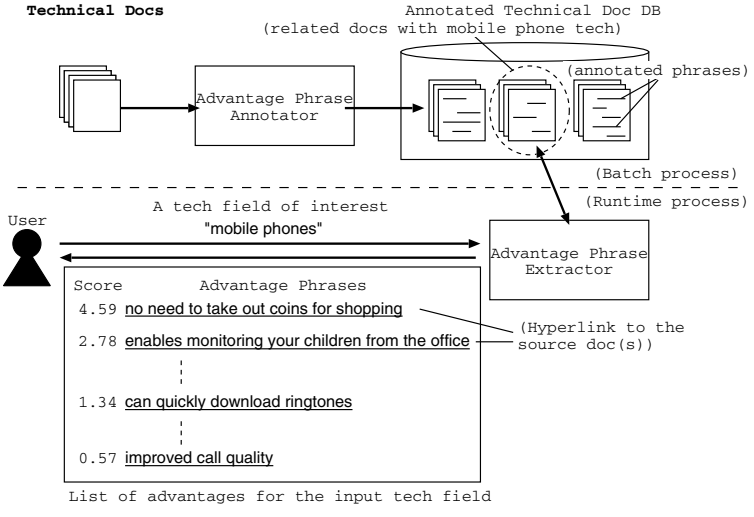


Fig. 1. CAPHMIT System Overview

In each run-time session, users input a query that expresses a technology field that they are interested in for the recent trends (such as "mobile phones"). The Advantage Phrase Extractor accepts the input query and determines documents related to the input technology field in the annotated technical documents database. From the selected documents, the Advantage Phrase Extractor extracts and lists the annotated advantage phrases. This list generated by the Advantage Phrase Extractor allows the user to grasp what has become possible with mobile phone technologies or what types of features are the focus of the newest mobile phones.

The output advantage list provided by the proposed tool is not a list of document titles and snippets as provided by conventional search engines, but a list of phrases allowing users to grasp the input technology's capabilities as a whole.

There is another important point that is different from the conventional search engines: the extracted advantage phrases are scored by the proposed scoring function to find novel applications of the technology. The tool also has the same function as with the conventional search engines, the listed advantage phrases work as hyperlinks to the original documents to start the more detailed survey.

3 Definition and Extraction of Advantage Phrase

We define an advantage phrase as *a phrase describing a favorable characteristic of a technology or a product from technical viewpoint.*

An advantage phrase can be classified into one of two types:

Enhancement Class	Advantages that are provided by enhancing intrinsically favorable points. This class includes advantages achieved by enabling preferable functions and effects of the technology or product.
Ameliorate Class	Advantages that are provided by repressing intrinsically unfavorable points. This class includes advantages achieved by inhibiting unfavorable side effects or limitations in conventional technologies.

To identify advantage phrases in the technical documents, we utilize predicates that are frequently used to express those classes of advantages. In technical documents, we often observe that verbs such as "increase" and "enhance" are used to express favorable advantages of the new technology in technical documents. Simultaneously, phrases such as "does not have to" also hints at existence of advantages in the subsequent text. Those verbs and phrases are employed as clue words to find advantage phrases from the original text.

All clue words employed in this study are listed in Table 1. These words are mainly selected from frequent verbs in Japanese patent disclosures. Even those few clue words appear in 79% of patent disclosures. In the advantage phrase annotation, those clue words are first identified in the technical documents, then an identified clue word and its modifiers are annotated as an advantage phrase.

It also should be noted that this approach utilizing clue words is expected to be applicable in other languages, as we can observe many of the direct translations of clue words in the table are employed in technical documents written in English. In the case of English documents, advantage phrases to be extracted are identified clue words and other words modified by the clue word; e.g. "enables monitoring your children from the office."

As explained in Section 2, those identified advantage phrases are annotated in the document that is stored in the database for later run-time processing. In the run-time processing, the input technology field is converted to a search query $\boldsymbol{q}$ to select documents that are related to the technology field. From the set of documents, the advantage phrases are extracted to form an output list. At the same time, those extracted phrases are scored from a perspective of how novel the application mentioned in the phrase is. In the next section, the proposed scoring function is explained.

Table 1. The clue words used to extract advantage phrases from Japanese technical documents and examples of the extracted phrases

Clue Words†	Advantage Phrase Examples
Enhancement Class	
... $*_{adp}$ koujou-$suru_{v}$ (improves ...)	shinrai-sei *wo koujou-suru* (*improves* reliability)
... $*_{adp}$ $takameru_{v}$ (enhances ...)	hikari no riyoukouritsu *wo takameru* (*enhances* light use efficiency)
... $*_{adp}$ $sugureru_{v}$ (exceed in ...)	keitai-sei *ni sugureru* (*exceed* in mobility)
... $kanou_{adjv}$ $*_{adp}$ $naru_{v}$ (enables ...)	kyoudo wo kakuho suru koto wo *kanou ni suru* (*enables* maintaining strength)
... $*_{v}$ $dekiru_{v}$ (can [verb] ...)	kairo-kiban no furyou wo *kensyutu dekiru* (*can detect* a defective component on a circuit board)
... $*_{adp}$ jitsugen-$suru_{v}$ (actualizes ...)	kairo no antei-dousa *wo jitsugen-suru* (*actualizes* stabilized circuit operation)
... $*_{*}$ $dekiru_{v}$ (allows ...)	seikaku na kyaribure-syon wo okonau koto *ga dekiru* (*allows* accurate calibration)
Ameliorate Class	
... $*_{adp}$ boushi-$suru_{v}$ (prevents ...)	gazou no rekka *wo boushi-suru* (*prevents* image quality degradation)
... $*_{adp}$ yokusei-$suru_{v}$ (controls ...)	hendou ni yoru eikyou *wo yokusei-suru* (*controls* fluctuation effects)
... $*_{adp}$ teigen-$suru_{v}$ (reduces ...)	shouhi-denryoku *wo teigen-suru* (*reduce* power consumption)
... $fuyou_{n}$ $*_{adp}$ $naru_{v}$ (become unnecessary ...)	sai-kyouiku ga *fuyou to naru* (*become unnecessary* to retrain)
... $hitsuyou_{n}$ $*_{adp}$ nai_{adj} (does not have to ...)	nan-mai mo no kurejitto ka-do wo mochiaruku *hitsuyou ga nai* (*does not have to* bring so many credit cards)
... $koto_{n}$ $*_{adp}$ nai_{v} (never ...)	tentou-suru you na *koto ga nai* (*never* stumbles)

† The subscript of a Japanese word in clue words indicate a part-of-speech (POS) tag of the word: adp indicates an adposition, v indicates a verb, adj indicates an adjective, adjv indicates an adjective verb, and n indicates a noun. Also, an asterisk in patterns means that any word or POS tag is allowed at the position.

4 Scoring Function for Advantage Phrase

To find phrases mentioning novel applications of the input technology, we focus on nouns, which are the important keywords in the extracted advantage phrases, and then the unexpectedness in the field and the generality of the nouns are defined and exploited.

It is clear that overly general nouns in the technical document corpus, such as 'invention' and 'device' in the case of patent disclosures, do not really contribute to finding novel applications of the technologies. To address this problem, the k most frequent nouns in the technical document corpus are not considered in the score calculation for phrases. The nouns in a stop list provided to the tool are also discarded. After removing those frequent nouns and stop words, the remaining nouns N appearing in a phrase p are used to score the phrase p.

The unexpectedness of a noun n in the input technology field is calculated by utilizing the correlation between the noun n and the query $\boldsymbol{q}$ used to obtain a set of documents related to the technology field. In the field of robotics, for example, nouns such as 'arms' and 'pet' are highly correlated with the field. Those nouns have low values of unexpectedness. On the other hand, nouns that rarely appear

in robotics receive high values for unexpectedness. This principle is implemented in the unexpectedness metric defined in Eq. (1). The argument takes a large value for the nouns highly correlated with the query $\boldsymbol{q}$, and therefore the value of unexpectedness(n) for those nouns has a small value. The unexpectedness(n) has a value between 0 and 1.

$$\text{unexpectedness}(\boldsymbol{q}, n) = \exp(-P(n|\boldsymbol{q})/P(n)) \tag{1}$$

To derive the generality of a noun, we assume that frequent nouns in newspapers, general webpages, and other documents that are common for ordinary people rather than professionals can be considered as general words. We took Japanese newspaper articles as the general document corpus in this study[2]. In this case, nouns that frequently appear in the newspaper are considered as general. This assumption is similar to document frequency (DF), which is a commonly utilized metric in IR, but for this noun generality metric, the DF of each noun is normalized by using Eq. (2) so that generality(n) takes a value between 0 and 1 similar to the unexpectedness($\boldsymbol{q}, n$).

$$\text{generality}(n) = \frac{1}{\log(\text{DF}_{\text{news}}(\text{ALL}+1))}\log(\text{DF}_{\text{news}}(n)+1) \tag{2}$$

Here, $\text{DF}_{\text{news}}(n)$ is the number of news articles that contain a noun n and $\text{DF}_{\text{news}}(\text{ALL})$ is the total number of news articles in the news corpus.

The generality and the unexpectedness of nouns are then combined as Eq. (3) to calculate a score for an advantage phrase p, where α is a coefficient to determine the effect of the unexpectedness ($0 \leq \alpha \leq 1$).

$$\text{score}(\boldsymbol{q}, p) = \sum_{n \in N} \alpha\text{unexpectedness}(\boldsymbol{q}, n) + (1-\alpha)\text{generality}(n) \tag{3}$$

5 Evaluation

This section consists of two experiments. One is about the advantage phrase extraction task and aims at understanding the difficulties of this new task on different types of corpora. The other is about the task of scoring the advantage phrases and aims at comparing the proposed scoring function with the manual labeling of phrases.

5.1 Experiment 1: Advantage Phrase Extraction

In this experiment, Japanese patent disclosures and new product announcements were employed as the technical document corpus. As the patent data, 50 Japanese patent disclosures submitted to the Japanese Patent Office were used. These 50 disclosures were randomly chosen from the disclosures submitted in

[2] The news articles in Sankei News for 2000 to 2005, which is a newspaper published in Japan, were used in this study. This corpus contains 214,867 news articles.

2006. A patent disclosure consists of several pre-defined sections. In the following experiments, the section describing the summary and the section describing the uses of a patent were used. In contrast, the product announcement data, the product and service announcements of IBM Japan published from 2002 to 2008 were used. The 50 announcements used in the experiment contain several types of topics: New software and hardware product announcements, new service announcements, new sales promotion announcements, and new product development announcements.

From the 50 patent documents and 50 announcements, two annotators manually extracted the phrases that would be taken as advantage phrases. The phrases extracted by the two annotators were then compared and their shared word sequences were used as the correct answer data. For instance, if Annotator A extracted the phrase *'yori chisana chikara de buhin no torihazushi wo okonau koto ga dekiru'* (can remove attachments with less force), and Annotator B extracted the phrase *'buhin no torihazushi wo okonau koto ga dekiru'* (can remove attachments) from the same sentence, the overlapped phrase *'buhin no torihazushi wo okonau koto ga dekiru'* (can remove attachments) was taken as the correct advantage phrase. Note that multiple advantage phrases can be extracted from one sentence. For the subjectivity of the advantage phrase extraction task, one often used indicator of inter-annotator agreement is the kappa coefficient, but this metric is not applicable for this task, since it is a task that extracts multiple phrases of various lengths from documents, while the kappa coefficient is designed for the task classifying an arbitrary number of samples. Therefore, the inter-annotator agreement in this phrase extraction task was defined as Eq. (4).

$$\textit{inter-annotator agreement} = \frac{2 \times n_{ab}(D)}{n_a(D) + n_b(D)} \tag{4}$$

D is the set of documents that advantage phrases were extracted from, $n_a(D)$ (or $n_b(D)$) is the number of advantage phrases that Annotator A (or B) extracted from a document set D, and $n_{ab}(D)$ is the number of correct advantage phrases after merging the advantage phrases found by the annotators.

Meanwhile, the versatility of clue words for multiple types of technical documents was investigated by comparing the precision (Eq. (5)), recall (Eq. (6)) and F-measure (Eq. (7)) for each document set.

$$\textit{precision} = \frac{TP(D)}{TP(D) + FP(D)} \tag{5}$$

$$\textit{recall} = \frac{TP(D)}{TP(D) + FN(D)} \tag{6}$$

$$\textit{F-measure} = \frac{2 \times \textit{precision} \times \textit{recall}}{\textit{precision} + \textit{recall}} \tag{7}$$

For this experiment, $TP(D)$ is the number of times that an correct advantage phrase in a document set D matches at least one word with an advantage phrase extracted by the proposed method. $FP(D)$ is the number of extracted phrases

that do not match with any of the correct phrases, and $FN(D)$ is the number of answer phrases that do not overlap with any of the extracted phrases.

Manual advantage phrase extraction found 259 phrases from the patent document set and 82 phrases from the announcement set as correct results. In testing, the proposed extraction method obtained 227 phrases from the patent document set and 90 phrases from the announcement set.

The inter-annotator agreement for the patent document set was 87.9[%], while the inter-annotator agreement for the announcement set was 43.5[%], which is roughly half of the agreement for the patent document set. This result implies that advantage phrase extraction task from free-style text such as product announcements is a difficult task, even for humans, considering the low inter-annotator agreement.

The advantage phrases that are extracted by at least one of the two annotators from the announcement set include phrases such as *'tokubetsu kakaku-de heisya seihin wo gokounyuu itadakeru'* (can buy our products at a special price) and *'mushou de shin-ba-jyon no sapo-to wo ukerareru'* (can obtain free support for the new version) that do not express advantages provided by technologies but that are due to special offers and services. These phrases are clearly not the advantage phrases that we aim at extracting from the technical documents. Extraction of these phrases should have been avoided by training the annotators so that only phrases mentioning technical advantages were extracted.

Another types of advantage phrase that lowered the inter-annotator agreement for the announcement set were phrases such as *'DVD R/W no sapo-to'* (Supports DVD R/W), *'TCP/IP soketto suu no zouka'* (more TCP/IP sockets) that mention technical advantages, but that require prior knowledge in the technology field for the annotators to understand that the phrase shows a strong point of the technology or the product. To obtain highly accurate data, those types of phrases should also be avoided.

Table 2 shows the precision, recall, and F-measure values for the patent document set (Patents) and the announcement set (Announcements). To compare the system performance with the human performance, F-measure values calculated by considering performance of one human annotator as system performance and performance of the other as a gold standard are also put in the table. These F-measure values obtained by the human-human comparison can be considered as the upper bound of the F-measure obtained by comparing the performance of the system with the manually-annotated answer data. The F-measure value of system-human comparison for the announcement set is around 0.5, though the F-measure value for the patent set is over 0.7. In both document sets, many false-positive advantage phrases were extracted by one of the two annotators.

Phrases that could not be extracted by the proposed method included *'shori no kousoku-ka'* (accelerating the process), *'dokuei wa-kufuro- no kouritsu-ka'* (streamlining of a X-ray interpretation work flow), and *'tei-kosuto'* (low cost). These phrases do not contain general predicates such that are utilized in our phrase extraction approach, but noun phrases and verbs that mention advantages in some specific domains. However, even these types of phrases often appear with the proposed phrasal patterns, such as *'kouritsu-ka wo kanou ni suru'*

Table 2. Accuracy of advantage phrase extraction from each document set

Document set	Precision	Recall	F-measure	F-measure (Human-Human Comparison)
Patents	0.797	0.700	0.745	0.872
Announcements	0.433	0.476	0.453	0.510

(enables streamlining), *'kou-kinou na sistemu wo jitugen-suru'* (actualizes a highly functional system). This means that some of these types of advantage phrases can be extracted by using phrasal patterns.

5.2 Experiment 2: Scoring Advantage Phrases

In this experiment, the validity of the proposed scoring function for the extracted advantage phrases was investigated.

For this experiment, three consultants who are potential users of the proposed tool were asked to evaluate the advantage phrases extracted for two different technology fields. From the technology fields suggested by the test subjects for investigation with this proposed tool, Data Mining and Video and Image Processing were chosen for this experiment. These technology fields were converted into queries as "data mining" and "video AND image processing." Documents that contain the query words were extracted from 290,889 patents and used for manual evaluation.

Each test subject gave one of the scores below for each phrase.

1: Expected to be a novel application in the query technology field
0: Appears to be a traditional application

In addition to these two scores, test subjects were allowed to give another score if desired. They could score -1 for a phrase that cannot be evaluated due to insufficient context. This score was often given for excessively short phrases. Scores 1 and 0 were averaged for the three test subjects. Phrases having a converted

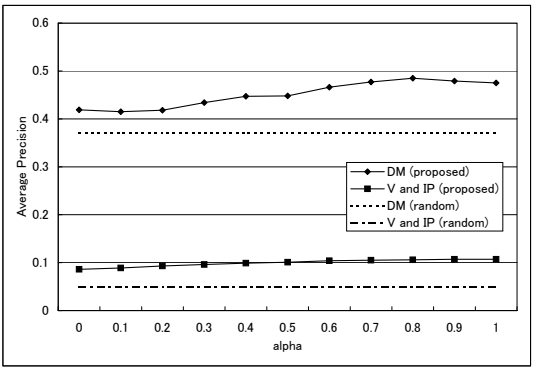


Fig. 2. Average Precision for each α setting

Table 3. A list of advantage phrases for 'Data Mining' query (top 15 phrases)

Advantage Phrase (English translation)	Score	Scored Nouns	
improve usability of handheld devices with low typing capacity	2.81	capacity, typing, usability, handheld device	*
can prevent reduction of information retrieval accuracy	2.01	reduction, information retrieval, accuracy	*
improve usability of taking various types of information services	1.81	various†,usability, information service	*
can be modified during the computation	1.80	during†, computation	
can derive potential targets	1.80	target, derive†, potential†	
can request delivery arrangement	1.73	request†, target, arrangement	
can appropriately grasp relationships between nurses or environments and (something)	1.63	environment, relationship, nurse	
can achieve transaction reduction	1.58	reduction, transaction	
can create process-quality model	1.53	quality, process, model	*
can generate graphic image	1.41	image, graphic	*
can be utilized in prediction of focused goods quality	1.30	quality, prediction	
can prevent forgetting to send	1.29	forget†, send†	*
can develop correspondent business scheme in real-time	1.17	business, real-time, scheme	*
can provide profitable enhancement for information processing capability	1.16	enhancement, information processing	
can observe detailed status of the phenomenon change	1.16	change, phenomenon, status	

* indicates the advantage phrases labeled as a novel application
†While these words are not nouns in English, they were POS tagged as nouns in the original Japanese corpus.

Table 4. A list of advantage phrases for 'Data Mining' query (bottom 15 phrases)

Advantage Phrase (English translation)	Score	Scored Nouns	
can leverage as a tool	0.115	tool	
can intuitively understand a social phenomenon	0.078	social phenomenon	*
can obtain insight	0.067	insight	
can efficiently run data analytics	0.037	data analytics	
can efficiently provide information obtained by data mining	0.034	mining†	
can achieve effective information system	0.034	information system	
can prevent system from insufficient operation	0		
can realize the system	0		
can realize how new (something) is	0		*
can suggest (something) to users	0		*
can realize the device	0		
can visualize	0		*
can infer something	0		
can get to know something	0		
can provide a way to analyze	0		

* indicates the advantage phrases labeled as a novel application
†While these words are not nouns in English, they were POS tagged as nouns in the original Japanese corpus.

and averaged score over 0.5 were labeled as novel applications in the technology field. Phrases given a score of 0 by more than two out of the three test subjects were not given an answer label.

Figure 2 shows average precisions for ranking by varying α value. This shows that average precisions are improved by increasing α, that is, by increasing the effect of unexpectedness. Average precisions for random ranking were 0.37 for the Data Mining field and 0.050 for the Video and image processing field. This shows

that the proposed scoring function worked much better than random ranking in both of the fields.

Tables 3 and 4 show the top 15 and bottom 15 phrases in the list ranked by the best parameter setting ($\alpha = 0.8$). Phrases appearing with an asterisk are phrases that were labeled as novel applications in the technology field, and phrases that were given 0 points (insufficient for evaluation) were also included in the lists. As we can see, 7 of the 19 phrases labeled as a novel application are in the top 15 phrases in the list, while only 4 of these 19 phrases appear in the bottom 15 phrases. This result implies that the proposed scoring function tends to agree with the decisions made by the test subjects who are the actual target users of the tool.

6 Related Work

Several tools have been proposed to support technical document mining and to give insights for decision making in management [2,3,4]. In particular, to leverage patents to understand technology trends, patent mining tools have been proposed and used [5,6,7,8]. The major function common in these tools is extracting keywords appearing in the documents and visualizing them to imply relationships between those keywords and the patent inventors and/or the publication years. There are prior no tool to grasp a wide range of potential capabilities and advantages of new technologies and products.

There have also been conventional studies that aim at discovering unexpected information from webpages of competitor companies [9,10] and scientific research papers [11,12]. They modeled target documents (webpages or papers) using the bag-of-words model scored with TF-IDF based scores, and compared them with the prior knowledge modeled with the user's own webpages or papers selected by the user. This research supported finding documents containing unexpected terms. However, our research aims at grasping advantages and potential capabilities of a technology, and our proposed scoring function aims at finding phrases mentioning novel applications of technologies. These differences in the purposes of the conventional studies and our study required this new survey assistance tool focusing on a new semantic class of phrases and a scoring function.

There have been many studies that focus on analysis of customers' voices such as sentiment analysis and opinion mining. In particular, several studies are focuses on demands of customers [13] and troubles they had [14] related to given types of products or topics. Extracted customer demands would be complementary relationship with advantage phrases in enhancement class, and customer troubles would be complement with advantage phrases in ameliorate class.

7 Conclusion

In order to support quick surveys by consultants, we proposed a novel information retrieval tool that lists up advantage phrases describing a favorable characteristic of a technology or a product from technical viewpoints. The results

of the experiments showed that the proposed phrase extraction method using phrasal patterns is able to identify most of the advantage phrases annotated by humans. We found that this task is highly subjective for some types of documents and inter-human agreement in this task for such domains was less than 50%. Nevertheless, the proposed phrase extraction method was able to identify nearly half of the phrases that were commonly selected in the manual extraction. Evaluation of the proposed scoring method showed that advantage phrases that can be expected to provide high business impact are highly ranked by focusing on the generality and the unexpectedness of the advantage phrases.

Future work for this study includes employing supervised learning methods to obtain the clue words, and to evaluate effectiveness of the classifications of the advantage phrases. The proposed phrase extraction approach should also be tested with documents in other languages.

Acknowledgement. The authors would like to thank Junji Maeda, Wataru Sasamoto, Toshiyuki Kuramochi, and Akihiro Kuroda in IBM Business Consulting Services for their valuable comments on our research.

References

1. Nishiyama, R., Takeuchi, H., Watanabe, H.: Towards Future Technology Projection: A Method for Extracting Capability Phrases from Documents. In: Corruble, V., Takeda, M., Suzuki, E. (eds.) DS 2007. LNCS (LNAI), vol. 4755, pp. 270–274. Springer, Heidelberg (2007)
2. Lent, B., Agrawal, R., Srikant, R.: Discovering Trends in Text Databases. In: KDD 1997, pp. 227–230 (1997)
3. Losiewicz, P., Oard, D., Kostoff, R.: Textual Data Mining to Support Science and Technology Management. Journal of Intelligent Information Systems 15(2), 99–119 (2000)
4. Qazvinian, V., Radev, D.R.: Scientific Paper Summarization Using Citation Summary Networks. In: COLING 2008, pp. 689–696 (2008)
5. Porter, A.L., Newman, N.C.: Patent Profiling for Competitive Advantage: Deducing Who Is Doing What, Where, and When. In: Handbook of Quantitative Science and Technology Research, pp. 587–612. Kluwer Academic Publishers, Dordrecht (2005)
6. Bragge, J., Relander, S., Sunikka, A., Mannonen, P.: Enriching Literature Reviews with Computer-Assisted Research Mining. Case: Profiling Group Support Systems Research. In: HICSS 2007, p. 243a (2007)
7. Tseng, Y.H., Lin, C.J., Lin, Y.I.: Text Mining Techniques for Patent Analysis. Information Processing and Management 43(5), 1216–1247 (2007)
8. Jin, B., Teng, H., Shi, Y., Qu, F.: Chinese Patent Mining Based on Sememe Statistics and Key-Phrase Extraction. In: Alhajj, R., Gao, H., Li, X., Li, J., Zaïane, O.R. (eds.) ADMA 2007. LNCS (LNAI), vol. 4632, pp. 516–523. Springer, Heidelberg (2007)
9. Liu, B., Ma, Y., Yu, P.S.: Discovering Unexpected Information from Your Competitors Web Sites. In: KDD 2001, pp. 144–153 (2001)
10. Chen, X., Wu, Y.: Web Mining from Competitors Websites. In: KDD 2005, pp. 550–555 (2005)

11. Jacquenet, F., Largeron, C.: Discovering Unexpected Information for Technology Watch. In: Boulicaut, J.-F., Esposito, F., Giannotti, F., Pedreschi, D. (eds.) PKDD 2004. LNCS (LNAI), vol. 3202, pp. 219–230. Springer, Heidelberg (2004)
12. Jacquenet, F., Largeron, C.: Using the Structure of Documents to Improve the Discovery of Unexpected Information. In: SAC 2006, pp. 1036–1042 (2007)
13. Kanayama, H., Nasukawa, T.: Textual Demand Analysis: Detection of Users' Wants and Needs from Opinions. In: COLING 2008, pp. 409–416 (2008)
14. De Saeger, S., Torisawa, K., Kazama, J.: Looking for trouble. In: COLING 2008, pp. 185–192 (2008)

Extracting Causal Knowledge Using Clue Phrases and Syntactic Patterns

Hiroki Sakaji[1], Satoshi Sekine[2], and Shigeru Masuyama[1]

[1] Toyohashi University of Technology, 1-1 Hibarigaoka, Tempaku-cho, Toyohashi-shi, Aichi 441-8580, Japan
sakaji@smlab.tutkie.tut.ac.jp, masuyama@tutkie.tut.ac.jp
[2] NewYork University, 715 Broadway, 7th floor, New York, NY 10003 USA
sekine@cs.nyu.edu

Abstract. This paper proposes a method to extract causal knowledge (cause and effect relations) using clue phrases and syntactic patterns from Japanese newspaper articles concerning economic trends. For example, a sentence fragment "World economy recession due to the subprime loan crisis ..." contains causal knowledge in which "World economy recession" is an effect phrase and "the subprime loan crisis" is its cause phrase. These relations are found by clue phrases, such as "ため (*tame*: because)" and "により(*niyori*: due to)". We, first, investigated newspaper corpus by annotating causal knowledge and clue phrases. We found that some specific syntactic patterns are useful to improve accuracy to extract causal knowledge. Finally, we developed our system using the clue phrases and the syntactic patterns and showed the evaluation results on a large corpus.

1 Introduction

A large amount of machine-readable textual documents including Web pages and newspaper articles are now available. We can find a lot of valuable information for many real applications in the documents by text mining. One of such information is "causal knowledge". We expect causal knowledge in the economic domain would be useful to forecast economic trends and prevent loss of business opportunities. For example, if we get causal knowledge where *cause* is "the Year 2000 problem" and *effect* is "the decline of the sales of hotels", we infer that problems like "the Year 2000 problem" may cause the decline of the sales of hotels. (Note: Hotels heavily rely on IT technologies. However, hotel managers are very busy handling daily business, and the hotel industry was not ready for such a problem until the last minutes.) Nevertheless, extracting such knowledge costs prohibitively high and time-consuming. Thus, there are some attempts to extract causal knowledge from textual documents automatically [1,2,3,4].

We propose a method to extract causal knowledge from Japanese newspaper articles concerning economic trends using syntactic patterns. For example, "World economy recession due to the subprime loan crisis ...", here, "World economy recession" is an effect phrase, "the subprime loan crisis" is its cause phrase and "due

T. Yamaguchi (Ed.): PAKM 2008, LNAI 5345, pp. 111–122, 2008.

to" is a clue phrase, respectively. These cause and effect relations are explicitly expressed by the clue phrase "による (*niyori*: due to)". However, if we just use the clue phrase to extract causal knowledge, it extracts a lot of noises. By investigating the corpus, we found that there are some syntactic patterns in cause-effect phrases and these patterns are useful to improve the accuracy of the results. Furthermore, causal knowledge is sometimes expressed in more than one sentence. Our method can also handle causal knowledge stridden over two consecutive sentences. Finally, we developed our system to extract causal knowledge using the clue phrases and the syntactic patterns and evaluated on a large corpus.

2 Related Work

A lot of work has been done on causal information extraction from a large corpus. Inui et al. proposed a method for acquiring causal relations (*cause*, *effect*, *precond* and *means*) from a complex sentence containing a Japanese resultative connective "ため (*tame*: because)" [1]. The Japanese resultative connective "ため (*tame*: because)" is a strong clue for causal information. In their research, clue phrases other than "ため (*tame*: because)" are not used, therefore, their method can not extract causal relations expressed by other clue phrases. In contrast, our method can extract many causal relations expressed by 36 clue phrases.

Khoo et al. proposed a method for extracting cause-effect information from newspaper articles by applying patterns made manually[2]. Furthermore, they proposed a method for extracting causal knowledge from a medical database by applying graphical patterns[3]. However, in their research, both *cause* and *effect* need to be contained together in the same sentence. Thus, these methods are not able to extract causal knowledge stridden over two sentences, while, our method can extract causal knowledge in such a case.

Chang et al. proposed a method for extracting causal relations that exist between noun phrases using clue phrases and word pair probabilities[5]. This probability was defined as the probability of a causal noun phrase pair. They used a bootstrapping method for learning Naive Bayes causality classifier. Girju proposed a method for automatic detection and extraction of causal relations based on clue phrases[4]. In their paper, the causal relation is expressed by a pair of noun phrases. Girju used WordNet as semantic constraints for selecting candidate pairs. Hence, her method can not extract unknown phrases that are not in WordNet. In contrast, our method deals with causal knowledge expressed by not only noun phrases but also verb phrases and sentences. In addition, our method can extract unknown phrases.

Sakai et al. proposed a method for extracting cause information from Japanese financial articles concerning business performance[6]. Their method only extracts cause phrases. On the other hand, our method can also extract effect phrases.

3 Investigation of Clue Phrases

We investigate 300 newspaper articles concerning economic trends using clues for extracting appropriate causal knowledge. We expect that newspaper articles

concerning economic trends contain a large amount of causal knowledge. Newspaper articles concerning economic trends are acquired by Sakaji's[7] method from the Nikkei newspaper published from 1990 to 2005.

3.1 Tagging Rules

Newspaper articles concerning economic trends are annotated with the following tags (in Table 1) to investigate how causal knowledge and the clue phrases are expressed. A human annotates tags to causal knowledge that is expressed explicitly by a fact and its account in a sentence or two adjacent sentences. We have two notes about the definition. In Japanese, the subject and the predicate can be expressed separately in a sentence. In such a case, the subject of an effect phrase is annotated with "EFFECT_SBJ" and the predicate of an effect phrase is annotated with "EFFECT_PRED". Otherwise, the effect phrase is annotated with "EFFECT".

Table 1. List of tags

Tag	Description	Examples
CLUE	clue phrase	ため (*tame*: because), から (*kara*: from)
CAUSE_VP	verb cause phrase	株式市場が下落した (*kabushiki sijyou ga gerakushita*: stock market declined)
CAUSE_NP	noun cause phrase	景気の回復 (*keiki no kaihuku*: recovery of economy)
EFFECT	effect phrase	世界不況 (*sekai hukyou*: world economy recession)
EFFECT_SBJ	subject of effect phrase	農産物価格は、(*nousanbutsu kakaku ha*: agricultural price)
EFFECT_PRED	predicate of effect phrase	下落した (*geraku shita*: fell off)
INV	word which changes verb/noun of cause phrase	の (*no*: of), こと (*koto*: thing)

We found that specific clue phrases are used depending on the type of the cause phrase. For example, clue phrase "ため (*tame*: because)" is used only when the cause phrase is a verb phrase. However, if the cause is a predicate, a special word, such as "の (*no*: of)" or "こと (*koto*: thing)" is used to convert from a verb to a noun. We annotated "INV" tag to such words.

3.2 Results of Tagging

Inui et al. also reported an investigation of clue phrases[8] on the social domain of a general newspaper, the Mainichi newspaper. Table 2 shows the number of investigated articles, the total number of clue phrases and the set of clue phrases of the investigation by Inui's and ours. It shows that our investigation found more clue phrases per article than Inui's. We believe that the articles concerning economic trends have more causal knowledge than that of the social

Table 2. The number of investigated articles, the total number of clue phrases, the set of clue phrases

	Num. of articles	Total num. of clue phrases	Set of clue phrases
inui's investigation	750	219	34
our investigation	300	695	154

Table 3. The number of phrases used as clue phrases(Frequency), the total number of phrases(Sum), the ratio of phrases used as clue phrases(Rate)

Clue phrase	Frequency	Sum	Rate
で (*de*: by)	155	2385	0.065
による (*ni yoru*: by)	48	244	0.197
で、(*de*: by)	48	340	0.141
から (*kara*: from)	46	646	0.071
ため、(*tame*: because)	30	55	0.545
を背景に (*wo haikei ni*: behind)	24	26	0.923
から、(*kara*: from)	17	46	0.367

domain, though minor differences of definitions may cause some differences. Clue phrases with the seven highest frequencies found by our investigation are shown in Table 3.

In Table 3, clue phrases such as "で (*de*: by)", "による (*ni yoru*: by)" and "から (*kara*: from)" frequently appear in articles. However, these clue phrases' ratio between the frequency that the phrase is used as a clue phrase compared with the total frequency of the phrase in the corpus is very low. Therefore, it is not appropriate to use them in our automatic causal knowledge extraction system. We define the following score using the frequency and the ratio, in order to accurately extract causal knowledge.

$$Score(t_i) = \log(TF(t_i)) \times R(t_i) \tag{1}$$

Here, $TF(t_i)$ is frequency that phrase t_i is used as a clue phrase. $R(t_i)$ is a ratio that phrase t_i is used as a clue phrase. Clue phrases with the ten highest *score* are shown in Table 4.

4 Extraction of Causal Knowledge

From our investigation, we found that most of the cause phrases appear before clue phrases in a sentence. On the other hand, effect phrases appear in various locations in a sentence. Furthermore, we found that clue phrases have four types of syntactic roles in indicating causal knowledge. Considering these observations we propose a method based on syntactic information. That is, we categorize patterns of cause-effect phrases by the location of the phrases and syntactic roles of clue phrases into four patterns as follows. These four patterns cover 86% of the cause-effect phrases in the corpus that we examined.

Table 4. Score of clue phrases

Clue phrase	Frequency	Sum	Rate	Score
を背景に (*wo haikei ni*: behind)	24	26	0.923	1.274
を背景に、(*wo haikei ni*: behind)	10	10	1.000	1.000
を受け、(*wo uke*: under)	12	14	0.857	0.925
を挙げる (*wo ageru*: quote)	10	12	0.833	0.833
ため、(*tame*: because)	30	55	0.545	0.806
に伴う (*ni tomonau*: with)	14	23	0.609	0.698
に伴い、(*ni tomonau*: with)	6	7	0.857	0.667
を反映して (*wo haneishite*: reflect)	6	7	0.857	0.667
に加え、(*ni kuwae*: besides)	9	13	0.692	0.661

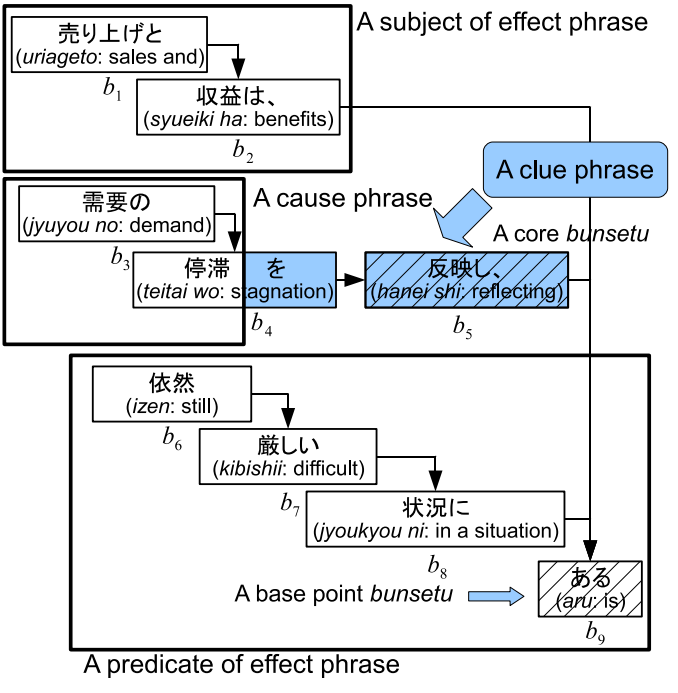


Fig. 1. An example of Pattern A

Pattern A: both a predicate and a subject are effect phrases in a sentence (see Fig. 1). The clue phrase has a role of connecting a predicate of an effect phrase and a subject of an effect phrase with a cause phrase.

Pattern B: an effect phrase appears after a cause phrase in a sentence (see Fig. 2). The clue phrase has a role of connecting an effect phrase with a cause phrase.

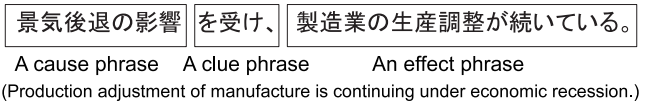


Fig. 2. An example of Pattern B

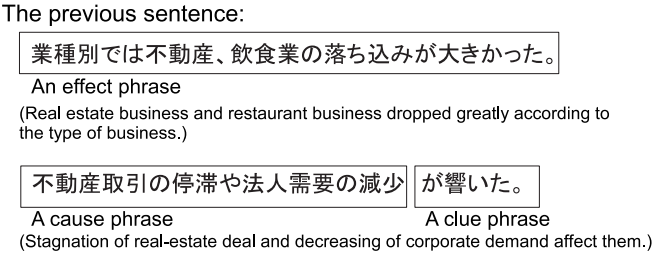


Fig. 3. An example of Pattern C

Pattern C: an effect phrase is the sentence just before a sentence including a clue phrase (see Fig. 3). The clue phrase has a role of indicating a previous sentence that is a cause phrase.

Pattern D: an effect phrase appears before a cause phrase in a sentence (see Fig. 4). The clue phrase has a role that is modified by an effect phrase and a cause phrase.

Patterns A, B and D extract causal knowledge from one sentence. On the other hand, Pattern D extracts causal knowledge from two adjacent sentences.

In the following three subsections, we will explain how we identify the patterns and how to extract cause-effect phrases using syntactic information.

4.1 Parser

The sentences that include clue phrases are parsed by a Japanese dependency analyzer, Cabocha[1]. For example,

サブプライムローンの危機により、世界不況が起こった。
(World economy recession was caused by the subprime loan crisis.)

The parsed sentence is shown in Fig. 5. Here, a *bunsetu* is a basic block in Japanese composed of several words. The sequence of *bunsetu*s in a sentence are expressed as $(b_1, b_2, ..., b_n)$, a subscript of b is a *bunsetsu*'s number. *Bunsetu*s are assigned consecutive numbers, beginning from one, in ascending order from the beginning of a sentence. For example, in Fig. 5, "サブプライムローンの (*sabupuraimuro-n no*: the subprime loan)" has a number 1.

[1] http://chasen.org/~taku/software/cabocha/

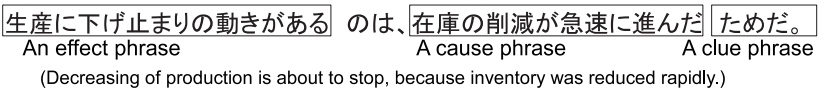


Fig. 4. An example of Pattern D

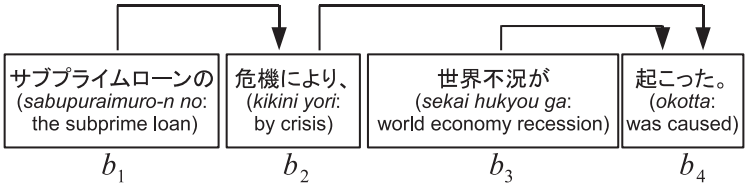


Fig. 5. An example of a parsed sentence

4.2 Pattern Identification

Here, we define the *core bunsetu* as the *bunsetu* that is rearmost *bunsetu*s composing a clue phrase. In addition, we define the *base point bunsetu* as the *bunsetu* modified by the core *bunsetu* (see Fig. 5). Our method searches causal knowledge and identifies an appropriate syntactic pattern.

Step 1: Search a sentence including clue phrases.
Step 2: If the clue phrase includes a punctuation "。" or a punctuation "、" appears after the clue phrase, go to Step 4. Otherwise, go to Step 3.
Step 3: If the base point *bunsetu* is a verb phrase and a *bunsetu* that modifies the base point *bunsetu* includes a particle, Pattern A is chosen. Otherwise, Pattern B is chosen. Go to step 5.
Step 4: If a *bunsetu* that modifies a core *bunsetu* includes a particle, Pattern D is chosen. Otherwise, Pattern C is chosen.
Step 5: Stop.

4.3 Cause Effect Phrase Extraction

Each pattern consists of a "cause phrase extraction process" and an "effect phrase extraction process". Each of these processes has a string variable *CAN* (an abbreviation of a candidate) and a numeric variable M. *CAN* is assigned "null" and M is assigned 0 as an initial value, respectively. First, we will describe the cause extraction process for all patterns, then we will describe the effect extraction process.

Cause Extraction

Patterns A and B

Step 1: Search a core *bunsetu* in the parsed sentence. The core *bunsetu*'s number is assigned to M.
Step 2: If *CAN* is "null", *CAN* is assigned b_{M-1} except for a particle, else connect b_{M-1} with *CAN*, and assign it to *CAN*. M is assigned $M-1$.

Step 3: Repeat Step 2 until $M-1$ becomes 0 or b_{M-1} modifies a phrase that has a number larger than the core *bunsetu*'s number.
Step 4: Acquire *CAN* as a cause phrase. □

Patterns C and D

Step 1: Search a core *bunsetu* in the parsed sentence. The core *bunsetu*'s number is assigned to M.
Step 2: If *CAN* is "null", *CAN* is assigned b_{M-1} except for a particle, else connect b_{M-1} with *CAN*, and assign it to *CAN*. M is assigned $M-1$.
Step 3: Repeat Step 2 until $M-1$ becomes 0 or b_{M-1} modifies the core *bunsetu* and b_{M-1} includes a syndetic particle(*kakari joshi*) or a conjunctive particle(*setsuzoku joshi*).
Step 4: Acquire *CAN* as a cause phrase. □

Effect Extraction

Pattern A
The procedure to extract the predicate of an effect phrase:

Step 1: Search a base point *bunsetu* from the parsed sentence and assign it to *CAN*. The base point *bunsetu*'s number is assigned to M.
Step 2: Connect b_{M-1} with *CAN*, and assign it to *CAN*. M is assigned $M-1$.
Step 3: Repeat Step 2 until $M-1$ becomes 0 or b_{M-1} becomes a core *bunsetu*.
Step 4: Acquire *CAN* as a predicate of an effect phrase. □

The procedure to extract the subject of an effect phrase:

Step 1: Search a *bunsetu* that modifies a base point *bunsetu* and includes a syndetic particle(*kakari joshi*) or a case particle(*kaku joshi*) from the parsed sentence. Assign this *bunsetu* to *CAN*. The *bunsetu*'s number is assigned to M.
Step 2: If b_{M-1} modifies b_M, connect b_{M-1} with *CAN*, and assign it to *CAN*, else go to Step 4. M is assigned $M-1$.
Step 3: Repeat Step 2 until $M-1$ becomes 0.
Step 4: Acquire *CAN* as a subject of an effect phrase. □

Fig. 1 illustrates an example of extracting causal knowledge by Pattern A. When extracting a predicate of an effect phrase, first, *CAN* is assigned a base point *bunsetu* "ある (*aru*: is)". M is assigned a base point *bunsetu*'s number 9. Next, connect b_8 with *CAN*, and assign "状況にある (*jyoukyou ni aru*: is in a situation)" to *CAN*. M is assigned 8. Repeat this process, until b_{M-1} becomes a core *bunsetu* "反映し、(*han ei shi*: reflecting)". Then, *CAN* "依然厳しい状況にある (*izen kibishii jyoukyou ni aru*: is still in a difficult situation)" is extracted as a predicate of an effect phrase.

When extracting a subject of an effect phrase, first, *CAN* is assigned a *bunsetu* "収益は、(*syuueiki ha*: benefits)" that modifies a base point *bunsetu* "ある (*aru*: is)" and includes a syndetic particle(*kakari joshi*) "は (*ha*)". M is assigned this *bunsetu*'s number 2. Next, connect b_1 with *CAN* "収益は、(*syuueiki*

ha: benefits)", and assign "売り上げと収益は、(*uriage to syuueki ha*: sales and benefits)" to *CAN*. Because b_1 modifies b_2. M is assigned 1. Then, $M-1$ becomes 0. Therefore, *CAN* "売り上げと収益は、(*uriage to syuueki ha*: sales and benefits)" is extracted as a subject of an effect phrase.

When extracting a cause phrase, first, M is assigned a core *bunsetu*'s number 5. Assign b_4 except for particles "停滞 (*teitai*: stagnation)" to *CAN* because *CAN* is "null". M is assigned 3. Next, connect b_3 "需要の (*jyuyou no*: demand)" with *CAN* "停滞 (*teitai*: stagnation)", and assign "需要の停滞 (*jyuyou no teitai*: stagnation of demand)" to *CAN*. M is assigned 2. *CAN* "需要の停滞 (*jyuyou no teitai*: stagnation of demand)" is acquired as a cause phrase, because b_2 modifies "ある (*aru*: is)" that has a number larger than the core *bunsetu*'s number.

Pattern B

Pattern B extracts effect phrases in a manner similar to Pattern A's extraction of predicates of effect phrases, except the base point *bunsetu* is a noun phrase. If a base point *bunsetu* is a noun phrase, effect phrases are extracted in a manner described below:

Step 1: Search a base point *bunsetu* from the parsed sentence and assign it to *CAN*. The base point *bunsetu*'s number is assigned to M.
Step 2: If b_{M+1} is modified by a *bunsetu* that has a number smaller than a core *bunsetu*'s number, connect *CAN* with b_{M+1}, and assign it to *CAN*, else go to Step 4. M is assigned $M+1$.
Step 3: Repeat Step 2 until b_M becomes the rearmost *bunsetu* in the sentence.
Step 4: Acquire *CAN* except for a terminal particle as an effect phrase. □

Pattern C

The sentence just before the sentence including a clue phrase is extracted as an effect phrase.

Pattern D

The effect phrases are extracted in a manner described below:

Step 1: Search a *bunsetu* that modifies a core *bunsetu* and includes a syndetic particle(*kakari joshi*). Assign this *bunsetu* except for a particle and punctuation to *CAN*. The *bunsetu*'s number is assigned to M.
Step 2: If b_{M-1} modifies b_M, connect b_{M-1} with *CAN*, and assign it to *CAN*, else go to Step 4. M is assigned $M-1$.
Step 3: Repeat Step 2 until $M-1$ becomes 0.
Step 4: Acquire *CAN* as an effect phrase. □

5 Evaluation

In this section, we evaluate our method. A set of 200 new articles concerning economic trends are used for the evaluation. The clue phrases we used are the highest scored 40 clue phrases in Section 3.2 except for "で、(*de*: by)", "ほか、(*hoka*: except)", "ため (*tame*: because)" and "による (*ni yoru*: by)". Then we run the system.

Table 5. Extraction results

	Recall	Precision	F-Measure
Cause phrase	0.917	0.757	0.829
Effect phrase	0.809	0.526	0.638
Both	0.770	0.449	0.567

Table 6. Precision for each cause and effect phrases

Clue phrase	num. of phrases	Precision	
		Cause phrase	Effect phrase
を背景に (*wo haikei ni*: behind)(B)	14	0.923	0.923
ため、(*tame*: because)(A)	6	0.500	0.200
ため、(*tame*: because)(B)	24	0.708	0.664
に伴う (*ni tomonau*: with)(B)	20	0.750	0.500
に加え、(*ni kuwae*: besides)(B)	10	0.400	0
ためだ。(*tameda*: because)(C)	9	0.889	0.889
ためだ。(*tameda*: because)(D)	6	1.000	0.833
により (*ni yori*: by)(A)	9	0.778	0.667
により (*ni yori*: by)(B)	11	0.455	0.273
の効果が (*no kouka ga*: effect)(B)	8	0.750	0
の影響が (*no eikyou ga*: influenced by)(B)	15	0.733	0
により、(*ni yori*: by)(B)	14	0.929	0.786
によって (*ni yotte*: by)(B)	10	0.500	0.500
から、(*kara*: from)(A)	12	0.833	0.667
から、(*kara*: from)(B)	30	0.700	0.567

First, the system output is evaluated automatically by matching the output with the human made answers. However, minor differences are sometimes acceptable as correct, a human evaluates again when there is a difference. The total evaluation results for cause and effect are presented in Table 5. Table 6 shows the accuracy for each clue phrases with more than 5 frequency. Recall is the percentage of correctly extracted phrases out of the phrases found by the annotator. Precision is the percentage of the correctly extracted phrases out of all phrases found by the system. F-Measure is a combination of recall and precision equally weighted, and is calculated by the following formula 2.

$$F\text{-}Measure = 2 * precision * recall/(precision + recall) \quad (2)$$

F-Measure attains a high value only when both recall and precision are high.

As shown in Table 6, for clue phrases "に加え、(*ni kuwae*: besides)", "の効果が (*no kouka ga*: effect)" and "の影響が (*no eikyou ga*: influenced by)" no effect phrases are extracted. The F-measure for the cause and effect phrases are 0.829 and 0.639, respectively. We believe that the result of cause phrase extraction is good for a causal knowledge extraction task. Therefore, we consider that cause

Table 7. The number of false positives and false negatives

	num. of errors	num. of false positives	num. of false negatives
Cause phrase	71	51	20
Effect phrase	139	103	36

Table 8. Each precision of pattern

		Precision	
	num. of phrases	Cause phrase	Effect phrase
Pattern A	38	0.978	0.816
Pattern B	113	0.901	0.761
Pattern C	13	0.867	0.917
Pattern D	6	1.000	1.000

phrase extraction is sufficient for a real-life application. However, the result of effect phrase extraction is not satisfactory.

6 Error Analysis

We found two kinds of errors, "false positive" and "false negative". False positive is the one where the extracted phrase is not a causal phrase. False negative is the one where the causal phrase is not extracted. We investigated the number of false positives and false negatives, shown in Table 7. We can see that the number of errors of cause phrases are twice the number of errors of effect phrases. Also, the number of false positives are more than the number of false negatives in both cause and effect phrase extraction. Therefore, we consider that a method for determining the presence or absence of causal knowledge is necessary.

Next, we investigated precision of each pattern. In the investigation, we exclude false positive examples in order to examine performance of syntactic patterns in detail. The results are shown in Table 8. A number of Pattern B's faults was the most highest in the four patterns. Most of Pattern B's faults were the parser errors. The rest of them were algorithmic design errors. We expect that errors are decreased by improving an algorithm's conditional expression.

7 Conclusion

We proposed a method that extracts causal knowledge using clue phrases and syntactic patterns from newspaper articles concerning economic trends. We, first, investigated our newspaper corpus by annotating causal knowledge and clue phrases. Then, we found that some specific syntactic patterns are useful to improve accuracy to extract causal knowledge. Finally, we developed our system to extract causal knowledge using the clue phrases and the syntactic patterns and show the evaluation results on a large corpus.

References

1. Inui, T., Inui, K., Matsumoto, Y.: Acquiring causal knowledge from text using the connective marker *tame*. Journal of Information Processing Society of Japan 45(3), 919–933 (2004)
2. Khoo, C.S., Kornfilt, J., Oddy, R.N., Myaeng, S.H.: Automatic extraction of cause-effect information from newspaper text without knowledge-based inferencing. Literary and Linguistic Computing 13(4), 177–186 (1998)
3. Khoo, C.S., Chan, S., Niu, Y.: Extracting causal knowledge from a medical database using graphical patterns. In: Proceedings of the 38th ACL, pp. 336–343 (2000)
4. Girju, R.: Automatic detection of causal relations for question answering. In: ACL Workshop on Multilingual Summarization and Question Answering, pp. 76–83 (2003)
5. Chang, D.S., Choi, K.S.: Incremental cue phrase learning and bootstrapping method for causality extraction using cue phrase and word pair probabilities. Information Processing and Management 42(3), 662–678 (2006)
6. Sakai, H., Masuyama, S.: Cause information extraction from financial articles concerning business performance, ieice trans. IEICE Trans. Information and Systems E91-D(4), 959–968 (2008)
7. Sakaji, H., Sakai, H., Masuyama, S.: Automatic extraction of basis expressions that indicate economic trends. In: Washio, T., Suzuki, E., Ting, K.M., Inokuchi, A. (eds.) PAKDD 2008. LNCS (LNAI), vol. 5012, pp. 977–984. Springer, Heidelberg (2008)
8. Inui, T., Okumura, M.: Investigating the characteristics of causal relations in japanese text. In: The 43rd Annual Meeting of the Association for Computational Linguistics, Workshop on Frontiers in Corpus Annotation II: Pie in the Sky (2005)

Context-Based Text Mining for Insights in Long Documents

Hironori Takeuchi[1], Shiho Ogino[1], Hideo Watanabe[1], and Yoshiko Shirata[2]

[1] IBM Research, Tokyo Research Laboratory, IBM Japan, Ltd.
Shimotsuruma 1623-14 Yamato-shi, Kanagawa, Japan
[2] Graduate School of Business Science, University of Tsukuba
Ohtsuka 3-29-1 Bunkyo-ku, Tokyo, Japan

Abstract. In this paper, we consider long documents and try to find differences between document collections. In the analysis of document collections such as project status reports or annual reports, each document and each sentence tend to be relatively long. Therefore, it can be difficult to derive insights by looking only for representative concepts in the selected document collection based on a divergence metric. In this paper, we propose an analysis approach based on contextual information. By extracting pairs of a topic word and a keyword and assessing their representativeness in the selected document collection, we are developing a method to extract insights from these long documents. Applying the proposed method for the analysis between the annual reports of bankrupt companies and those of sound companies, we were able to derive insights that could not be extracted with the conventional methods.

1 Introduction

Recently large amounts of text data are available for computational analysis within companies. In contact centers, gigabytes of customer contact records are produced every day in the form of call summaries. Analysts try to gain insights for improving business processes from such stored customer contact data. Through analysis by experts results in insights that are very deep and useful, such analysis usually covers only a very small fraction of the total data volume and still requires major efforts. Therefore, the demands to gain insights from entire text data collections by using text mining technologies are increasing rapidly [1].

One of the major text mining analysis methods is to extract trends and information from document collections and derive useful insights. In some domains, useful concepts are defined in advance. In the life science area, for example, many scientists want to know the relationships among genes or among proteins for their research from the biomedical documents. For such purposes, information extraction and visualizing tools [2] work well. However, in many cases, we don't know in advance which concepts are useful for the analysis and finding the concepts leading to insights is also one of the objectives in such text mining analysis.

In this paper, we try to analyze long documents and discover the differences between various document collections. Difference analysis is very useful for getting insights in the text mining analysis. For example, given a collection of contact records at the

T. Yamaguchi (Ed.): PAKM 2008, LNAI 5345, pp. 123–134, 2008.

contact center of a manufacturer, difference analysis between documents mentioning different products often leads to business value by identifying specific problems in a specific product. In such a difference analysis, when viewpoints such as product names and compliant expressions for the difference analyses are defined in advance, we can use them and find concepts that are correlated to one of the document collections.

In this predefined-concept approach, the user has to define viewpoints for the analysis and store the concepts for each viewpoint as a dictionary. This is an important step and a failure here often leads to a failure in the analysis. Also, it has been an artistic task that requires highly experienced analysts who have learned by trial and error. In our target document collection, each document is very long. It is, therefore, very difficult to identify appropriate viewpoints and concepts manually from the data.

There are also approaches that automatically extract representative concepts in each document collection. In this concept-seeking approaches, for each document collection some keywords are automatically extracted as representative concepts. It often happens that many of the extracted keywords that are highly representative are well known concepts and do not lead to any useful insights. When each document is very long, it is sometimes difficult to interpret the extracted keywords to gain insights without reading the corresponding documents directly.

To address these problems, in this paper we propose a method to identify differences between document collections using context information. Using topic words that describe the contexts and their regional information, we calculate a "representativeness" metric for of the keywords in the document collection for each context. By calculating the representativeness values of the keywords in each context, we identify the topics that contain many highly representative keywords and extract the keywords that have different representativeness values in different contexts. Such keywords are not well-known facts compared with document-level representative keywords and by looking at the context-dependent keywords with their topics we can easily derive insights and hypotheses as analytic results. In the experiment, we show the effectiveness of our method through the analysis between the annual reports of bankrupt companies and those of going concerns.

Organization of the Paper: We start by describing the conventional difference analysis between document collections based on the representative expressions in the documents and consider the limitations of this approach when analyzing large documents or documents where the sentences are relatively long. Section 3 describes our analysis method for identifying differences between document collections using context information. Section 4 describes a financial account report analysis based on the proposed method and Section 5 provides the analytic results. We discuss the results in Section 6 and conclude the paper in Section 7.

2 Analysis of Document Collection Based on the Representative Expressions

Here, we consider the difference analysis between document collections A and B, and use kwd for some keyword expression appearing in the document collection. In many

text mining applications, $P(A|kwd)$ and $P(B|kwd)$, the probabilities of A and B when kwd appears in the document are used as metrics to identify key expressions in each text collection. When $P(A|kwd)$ is reasonably large compared with $P(A)$, we can feel justified in saying that kwd is correlated with A.

For example, in [3], keywords and expressions related to a defined viewpoint are registered in a dictionary. By using a two-dimensional table whose cells represent the numbers of documents containing corresponding keywords in two selected viewpoints, users can find strong relationships among keywords and derive insights.

There are some approaches that automatically extract representative keywords in a selected document collection. In [4], the most representative keywords are extracted from a selected document collection and they are presented in a network view. In the document categorization, some metrics like mutual information and the χ^2 statistic are used to identify keywords and expressions that are representative in each document collection and contribute to the categorization performance [5]. In the document clustering, extracting useful keywords from each cluster is important for the cluster labeling [6].

In these conventional approaches, representative keywords and expressions in a selected document collection are extracted, but many of the ones with high representativeness are well known concepts in the selected collection. We therefore tend to overlook concepts that lead to useful insights. It also sometimes happens that we cannot easily gain any insight just by looking at the extracted concepts without reading the corresponding sentences or parts in the document collection. In the document clustering, we also cannot summarize each cluster from only the labeled keywords and have to refer to some documents for a deep understanding of the cluster.

3 Difference Analysis between Document Collections Using Context Information

To address the problems described in the previous section, we propose a method for the difference analysis of long documents using context information.

Given context information that consists of a topic word and its region, our proposed method consists of these steps:

- Calculate the representativeness values of the keywords in each context based on a divergence metric.
- Identify the topic words that contain many representative keywords in their region in the selected document collection.
- Find keywords whose representativeness values change when the contexts are different.

In calculating the representativeness values of the keywords, we calculate the difference between $P(A|kwd)$ and $P(A)$. As mentioned in the previous section, for the difference metric, the Kullback-Leibler divergence (KL-divergence), mutual information, and the χ^2 statistic have been proposed. However these scores depend on $P(A)$, and it is difficult to define proper thresholds for identifying correlated concepts from these scores though we can easily interpret $P(A|kwd) - P(A)$. When A is binary, the KL-divergence between $P(A|kwd)$ and $P(A)$ is calculated as follows.

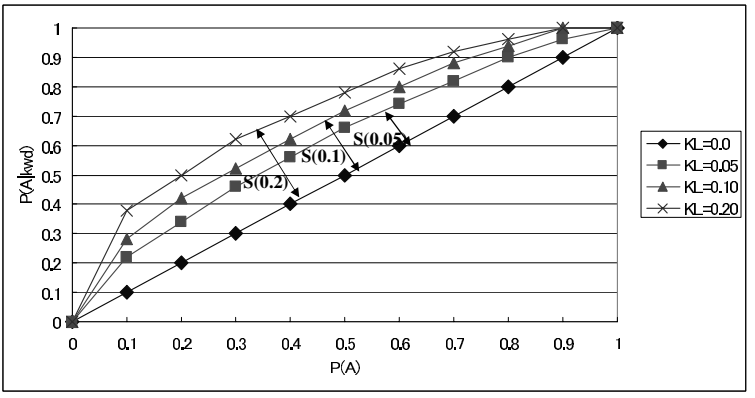


Fig. 1. Relationships between $P(A)$ and $P(A|kwd)$

$$KL(P(A|kwd), P(A)) = P(A|kwd) \log \frac{P(A|kwd)}{P(A)} + (1 - P(A|kwd)) \log \frac{1 - P(A|kwd)}{1 - P(A)}$$

Figure 1 shows the relationship between $P(A)$ and $P(A|kwd)$ for each KL-divergence value.

In Figure 1, $S(a)$ is defined as the area between the curve when the KL-divergence is a and the baseline, which is the curve when the KL-divergence is 0. We can regard $S(a)$ as the gain in the ROC curve [7]. When we decide on a, we can derive $P(A|kwd)$ for each $P(A)$ and can calculate $S(a)$.

From the value of $S(a)$, we define two parameters τ_1 and τ_2 ($0 \leq \tau_1 < \tau_2 \leq 0.5$) that represent the representativeness thresholds. Using τ_1 and τ_2, we assess the following representativeness classes for each kwd.

- kwd is not representative in A when $S(a) < \tau_1$.
- kwd is slightly representative in A when $\tau_1 \leq S(a) \leq \tau_2$.
- kwd is highly representative in A when $S(a) > \tau_2$.

When we decide on τ_i, we can calculate a representative score a that satisfies $\tau_i = S(a)$. From a for the KL-divergence, we can determine whether kwd is representative by investigating whether the difference metric between $P(A|kwd)$ and $P(A)$ is larger than a. Table 1 shows $P(A|kwd)$ for each τ and $P(A)$ when we use the KL-divergence as a difference metric. In this table, when we decide on τ_1 and τ_2, we can find the thresholds of slight and high representative probabfilities for each $P(A)$.

Now given the topic keywords and their dominating region, we can estimate a representativeness class for each keyword or expression. As results, we get the pairs of the topic word and the keyword in its context with a representativeness class. From these results, we can derive the trends using context information. For example, we can find topic words that have many representative keywords. From these topic words, we can derive which topics are representative in the selected document collection. We can also

Table 1. $P(A|kwd)$ for each τ and $P(A)$

		$P(A)$			
τ	KL	0.2	0.4	0.6	0.8
0.05	0.010	0.26	0.47	0.67	0.85
0.1	0.030	0.31	0.53	0.72	0.89
0.2	0.15	0.46	0.67	0.84	0.96
0.3	0.40	0.64	0.80	0.92	1.0

find keywords that have different representativeness classes in different contexts. We can conclude that these keywords have different roles in the different contexts. Furthermore, compared with extracting only representative keywords, we can easily interpret the representative pair of the topic word and the keyword and can derive insights on the selected document collection.

4 Analysis Experiment

In this section, we describe the experiment where we use our proposed analysis approach.

4.1 Data and Purpose of Analysis

We selected 90 annual reports of bankrupt Japanese companies and 90 reports of sound Japanese companies for the experiment. The bankrupt samples include all companies that went bankrupt from 1999 to 2005, because bankrupt companies are far scarcer than soundF ones. We used the final annual reports of these bankrupt companies. To prepare the 90 samples of sound companies, we sorted the all of the sound companies listed on the Tokyo Stock Exchange in 2005 by using the SAF (Simple Analysis of Failure) model [8]. The SAF model is a bankruptcy prediction model and provides a score value calculated from the corporation's four variables selected by the CART (Classification and Regression Tree) analysis of the financial data of companies [9]. From the ranked companies list, systematic extraction was performed at equal intervals so that the total came to 90. It is known that the SAF value is related to the results of the company rating. We can consider that these 90 companies is a small subset of all of the sound companies.

The distribution of the 90 sound companies was similar to the distribution of all of the companies listed in the actual stock market. From these two document collections, we are trying to find distinctive expressions peculiar to the annual reports of bankrupt companies and continuing companies by using the proposed method. Corporate annual reports consist of various sections such as numerical tables of financial accounting information, company histories, audit certifications, and so on, and therefore have some tens of pages. For this experiment, we focused on the sections about dividend policies, because such sections are included in all corporate annual reports and seem appropriate for investigating the differences between bankrupt and continuing companies, and for extracting expressions peculiar to each group.

4.2 Text Mining System

For the analysis, we constructed a text mining system for the difference analysis of "bankrupt company" versus "sound company". The experimental system consists of two parts, an information extraction part and a text mining part.

In the information extraction part we prepared dictionaries and patterns. The domain experts assign semantic categories to the words that frequently appear and prepare the dictionary. This dictionary consists of entries with surface representations, parts of speech (PoS), canonical representations, and semantic categories. In the input document, terms appearing in the dictionary are automatically mapped to the corresponding canonical form with semantic categories. By using patterns we can identify the word-to-word dependencies and the specific word sequences in the text. These processed are applied to the resutls of the PoS tagging and the dependenency analysis.

Once appropriate concepts have been extracted from the documents, we can apply various statistical analysis methods in data mining to the set of concepts as well as to the structured data. As a result, even a simple function that examines the increase and decrease of occurrences of each concept in a certain period may allow us to analyze trends. In the text mining part we selected appropriate topics and keywords to find interesting associations and insights between concepts and the company type (bankrupt or sound) leading to useful insights. When we set up the viewpoints for the analysis, we need to develop a dictionary and information extraction rules for identifying mentions of each item in the text. For example, we can define a dictionary whose contents are investiment related words such as capital investiment, plant and equiequipment and technology acquisition. As a result, we can fill in each cell in a two-dimensional table as in Table 2 by counting the number of texts that contain both the column and row labels[3].

Table 2. Two Dimensional Association Analysis

		company type	
		Sound Company Group	Bankrupt Company Group
Investiment relating terms	capital investiment		
	research and development		
	plant and equipment		
	technology acquisition		
	new business		

However, because of the differences in recall and precision for information extraction for each concept, the absolute numbers may not be reliable. Still, if we can assume that the recall and precision for extracting each concept are coherent over the whole data set, we can calculate indices showing the strengths of the associations for each cell compared to the other associations in the table.

For the concepts extracted in the information extraction phase, we make indexes and get the number of documents that satisfy the condition defined by users' query. In the analysis, therefore we can set various viewpoints and get analysis results such as a two-dimentional table mentioned abobe dynamically.

4.3 Information Extraction Using Context Information

As shown in Figure 2, in the official Japanese documents such as annual reports, the topic words are usually presented at the start of the sentence in a form such as "<topic word> *ni tsukimashite ha* ..." or "<topic word> *ha* ..." (As regards <topic word> ...). In this experiment, we assume that the region that the topic word is dominating is from the word after the topic words to the end of the sentence. We define this region as a topic region.

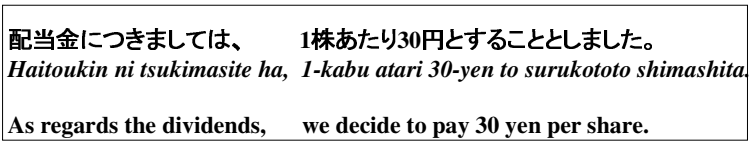


Fig. 2. Example sentence in an annual report

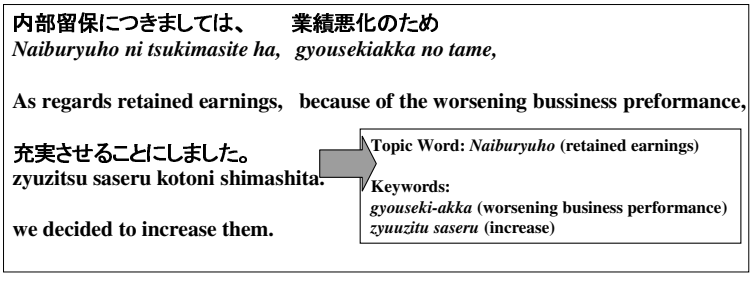


Fig. 3. Example of the information extraction using context information

The features extracted in the information extraction part of the experimental text mining system consist of nouns, compound nouns, specified noun phrases (e.g. numeral + noun), and verbs. Also, we extracted topic words using word sequence pattern matching and estimated their regional information. From this derived context information, we extracted the pairs of a topic word and a keyword for each topic appearance and estimated the representativeness class shown in Section 3 for the keyword in the context. Figure 3 shows an example of this information extraction process.

5 Analytic Results

We extracted topic words that appeared in more than 3 documents. The following expressions are examples of the features extracted as topic words.

haitou kin(dividends), *shikin*(capital),
rieki haibun(profit sharing), *naibu ryuuho*(retained earnings),
haitou seikou(payout ratio), *tyukan haitou*(interim didends)

For each topic, some features in the topic region are extracted. We first extracted sound-company-related features. We set $\tau_1 = 0.1$ and $\tau_2 = 0.2$ and decided on the

Table 3. Representativeness of features in sound companies in each topic

		Representativeness			
Topic Word	Features	No	Slight	High	Other
haitou kin (dividends)	10	1	3	6	0
shikin (capital)	13	4	1	6	2
rieki haibun (profit sharing)	10	6	1	2	1
naibu ryuuho (retained earnings)	41	10	8	10	9
haitou seikou (payout ratio)	3	0	2	0	1
tyukan haitou (interinm dividends)	26	4	19	3	0

representativeness types of the extracted features in each context. Table 3 shows the distribution of the representativeness types of the features in each topic. In this table, "Features" denotes the number of features in each context. In each context, we extracted features that appeared in more than 10 documents and assessed their representativeness in the sound company group. "Other" denotes the number of features that are correlated with bankrupt companies. We followed the same procedure for the bankrupt-company-related features.

In these results, in the '*naibu ryuho*'(retained earnings) topic work, some keywords are correlated to sound companies and other keywords are correlated with bankrupt companies. Following are examples of extracted expressions correlated with each group of companies in this retained earnings context.

- **Sound company group:** *zyuutou suru* (allot), *seichou* (grow), *setsubi toushi* (capital investment), *kenkyuu kaihatsu* (research and development), *kyousou* (competency), *gourika* (rationalization), *kigyou kachi* (corporate value), *seisan setsubi* (plant and equipment), *shinki zigyou* (new business)
- **Bankrupt company group:** *kihon* (basic), *zyuuzitsu suru* (enrich), *ouziru* (react), *antei* (stable), *rieki kangen* (profit return), *reimbursement*, *zyoukyou* (status)

It was found that the '*shikin*'(capital) topic word has a similar trend of representativeness classes.

Table 4 shows the representative keywords in each document collection that have high conditional probabilities ($P(\mathrm{sound}|kwd)$ and $P(\mathrm{bankrupt}|kwd)$).

From this table, we can see that '*kenkyuu kaihatsu*'(research and development) and '*kigyou kachi*'(corporate value) are representative in the sound company group. However '*naibu ryuho*'(retained earnings) appears similarly in both document collections and is not a representative keyword. Similarly '*shikin toushi*'(capital investment) and *seisan setsubi*(plant and equipment) are not representative in the sound group in Table 4.

Compared with this conventional keyword based approach, from our results we can see that these two keywords ('*shikin toushi*' and '*seisan setsubi*') are highly representative in the context of '*naibu ryuho*'(retained earnings). This result means that every company refers to retained earnings in the annual reports but the keywords in this topic are different in the two document collections (sound and bankrupt companies).

Table 4. Representative keywords in each document collection

Sound group	Bankrupt group
nenkan (annual)	*ikannagara* (I am sorry to say)
kenkyuu kaihatsu (research and development)	*hikitsuzuki* (continuously)
zyuutou suru (allot)	*sonshitsu* (loss)
renketsu gyouseki (consolidated financial settlement)	*kaifuku* (recovery)
tyuukan (interim)	*ohabana* (large-scale)
kigyou kachi (corporate value)	*zenryokude* (with all our might)
syutoku (acquire)	*hayaku* (soon)
ziko kabushiki (treasury stock)	*miokuru* (stop)
kaisai (hold)	*itaru* (come to)
tyuukan haitou (interim dividends)	*shitagatte* (therefore)

Compared with the results in Table 4, from the extracted pairs of topics and representative keywords, we can easily derive the insight that the sound companies used their **retained earnings** for **research and development** to improve their **corporate value** without evening referring to the corresponding documents.

There are two hypotheses on the relationships between the retained earnings and the capital investment for new business or research and development. In the business administration area, it appears that by investing their money in new business or research and development companies can then increase their retained earnings. In the financial accounting area, in contrast, it appears that companies can invest their money in new business or research and development only after increasing their retained earnings. Our analysis results, therefore, support that hypothesis for the financial accounting approach.

Now we consider the extracted features. From Table 4, we can see that '*tyuukan haitou*'(interim dividends) is associated with the sound group and apologetic expressions are associated with the bankrupt group. From the view of the financial data analysis, this result is clear because sound companies can pay dividends but imminently bankrupt companies cannot pay dividends as their business results become critical. This shows that we derive only well-known insights from the document-level representative keywords. In the information extraction phase, we extracted expressions referring to "the amount of money" using word sequence patterns such as "<numeral> + yen". This was not a representative expression in either document collection. It was, however, found that "amount of money" is highly representative in the sound companies when considering the '*haitoukin*'(dividends) context and is highly representative in the bankrupt companies when considering the '*gyouseki*'(business results) context. In these results, "amount of money" is a dividend in the dividends context and is a business loss in the business results context. This means that some keywords have different roles in different contexts and that looking only at the representativeness keywords leads to incorrect insights. Such keywords are interesting and lead to useful insights that the analysts had not noticed.

6 Discussion

In this paper, we are proposing a method to extract the representative pairs of a topic word and a keyword in the topic region in each document collection to derive insights into the differences between collections. In text mining analysis, looking only at the extracted keywords is not sufficient to derive insights. Therefore, word-to-word dependency information is frequently used [3].

For example, in the contact center of a manufacturer, extracting noun-to-verb direct dependencies such as "product AAA . . . broken" and "require . . . software patch BBB" is very useful to identify problems in a specific product and users' requirements. Here are some frequent noun-verb dependencies containing '*haitoukin*'(dividends) or '*naibu ryuho*'(retained earnings).

- **haitoukin**: *haitoukin . . . tsukimashiteha* (talk about . . . dividends), *haitoukin . . . naru* (dividends . . . become), *haitoukin . . . ketteisuru* (decide . . . dividends).
- **naibu ryuho**: *naibu ryuho . . . hitsuyouda* (need . . . retained earnings), *naibu ryuho . . . kakuhosuru* (preserve . . . retained earnings), *naibu ryuho . . . tutomeru* (try . . . retained earnings), *naibu ryuho . . . zyuuzitsusuru* (enrich . . . retained earnings)

This extracted dependency information is too general to lead to insights. The contact records in the call centers are usually summaries of the calls. In the records, the customer's problems and the requirements are described very tersely such as "The HDD on PC-ABC is broken" or "I need a fix". From such short sentences, only the important dependency information can be extracted. In contrast, the sentences in official documents tend to be long and complicated. In such sentences, topic words frequently do not have direct dependencies on important or meaningful concepts. To extract suitable dependency information, additional work on constructing extraction rules and methods for each data type is needed. Compared to this extracted dependency information, with our approach, meaningful results are robustly extracted from the documents that consist of long and complicated sentences.

In our approach, we extract representative pairs of topics and features by using topic words and their regional information. By using this context information, we can find representative topics in the selected document collection and also find keywords that have different roles in different contexts. There are various methods to show concepts correlated with each other in the document collections. In [10] and [11], the correlated concepts are connected by edges, and the widths of the edges represent the strengths of the correlation. In these approaches, we cannot determine which keyword is a main concept and cannot easily interpret the correlated keywords. In our approach, we can determine whether the topic keyword is a main concept in the pair of the topic and the keyword and can easily derive some insights on the difference between document collection.

In the experiment with Japanese annual reports, we extract topic words and their regional information by using the characteristic of Japanese that in official documents the topic words tend to be referred to at the beginning of the sentences. It was found that this simple heuristic approach works well in the analysis of Japanese annual reports and it can be expected that we can apply such context information extraction approaches to

the analysis of other types of official Japanese documents. To analyze other types of documents, we need methods to identify topics and their dominating regions. As future work, we want to consider methods to extract context information from the data without relying on heuristics.

There are many projects on financial data mining [12]. In these projects, in most cases the numeral financial results such as revenue or profit are used for the corporate analyses. Investigating relationships between the text mining analysis results and the financial data analysis results will also be part of our future work.

7 Conclusion

In this paper, we proposed a method to extract representative pairs of a topic word and a keyword for the analysis of long documents. We applied our context-information-based analytic approach to the analysis of Japanese annual reports and tried to find differences between bankrupt and sound companies. As a result, we confirmed that we could easily derive meaningful insights that could not be derived from the conventional keyword-based approaches. As future work, we want to consider non-heuristic approaches to extract context information from the data and integration of the text mining analysis results with the financial data analysis results.

References

1. Feldman, R., Sanger, J.: The Text Mining Handbook: Advanced Approaches in Analyzing Unstructured Data. Cambridge University Press, New York (2007)
2. Rzhetsky, A., Iossifov, I., Koike, T., Krauthammer, M., Kra, P., Morris, M., Yu, H., Duboue, P.A., Weng, W., Wilbur, J.W., Hatzuvassuloglou, V., Friedman, C.: Geneways: A system for extracting, analyzing, visualizing, and integrating molecular pathway data. Journal of Biomedical Informatics 37, 43–53 (2004)
3. Nasukawa, T., Nagano, T.: Text analysis and knowledge mining system. IBM Systems Journal, 967–984 (2001)
4. Hisamitsu, T., Niwa, Y.: A measure of term representativeness based on the number of co-occurring sailent words. In: Proceedings of the 19th International Conference on Computational Linguistics (COLING), pp. 1–7 (2002)
5. Yang, Y., Pedersen, J.O.: A comparative study on feature selection in text categorization. In: Proceedings of the 14th International Conference on Machine Learning (ICML), pp. 412–420 (1997)
6. Beil, F., Ester, M., Xu, X.: Frequent term-based text clustering. In: Proceedings of the ACM Conference on Knowledge Discovery and Data Mining (KDD), pp. 436–442 (2002)
7. Manning, C.D., Raghavan, P., Schütze, H.: Introduction to Information Retrieval. Cambridge University Press, New York (2007)
8. Shirata, C.Y.: Bankruptcy Prediction Model. Chuokeizai-Sha, Tokyo (2003) (in Japanese)
9. Shirata, C.Y., Terano, T.: Extracting predictors of corporate bankruptcy: Empirical study of data mining method. In: Terano, T., Liu, H., Chen, A.L.P. (eds.) PAKDD 2000. LNCS (LNAI), vol. 1805, pp. 204–207. Springer, Heidelberg (2000)

10. Ohsawa, Y., Benson, N.E., Yachida, M.: Keygraph: Automatic indexing by co-occurrence graph based on building construction metaphor. In: Proceedings of IEEE International Forum on Research and Technology Advances in Digital Libraries (ADL), pp. 12–18 (1998)
11. Aumann, Y., Feldman, R., Yehuda, Y., Landau, D., Liphstat, O., Schler, Y.: Circle graphs: New visualization tools fo text-mining. In: Proceedings of the 3rd European Conference on Principles and Practice of Knowledge Discovery in Databases (PKDD), pp. 277–282 (1999)
12. Vityaev, E., Kovalerchuk, B.: Data mining fo financial applications. In: Data Mining and Knowledge Discovery Handbook: A Complete Guide for Practioners and Researchers, pp. 1203–1224. Springer, Heidelberg (2005)

A Knowledge Management Approach for Structural Capital

Dimitris Karagiannis, Florian Waldner, Anita Stoeger, and Martin Nemetz

University of Vienna, Department of Knowledge and Business Engineering,
Bruennerstr. 72, 1210 Vienna, Austria
{dk,fw,as,mn}@dke.univie.ac.at
http://www.dke.univie.ac.at

Abstract. Knowledge represents the most important production factor for huge fields of today's economy. Since more than a decade the research on knowledge management and intellectual capital management has brought insights for the question of how organizations and their employees create value and eventually also profits. As human capital is highly volatile, organizations strive to transfer employee's knowledge into the more institutionalized structural capital. For this transformation as well as to identify and fill knowledge gaps we propose the integration of the ICRB framework and the eduWEAVER approach to a comprehensive IT-based management approach for knowledge management.

Keywords: Knowledge Management, Intellectual Capital, eLearning, Knowledge Gaps.

1 Introduction

In fact, it is an open secret that knowledge represents the most important production factor for huge fields in today's economy [1], [2]. Since more than a decade the research on knowledge management and intellectual capital management has brought insights for the question of how organizations and their employees create value and eventually also profits (see e.g. [3] or [4]). However, although theoretical foundations, such as the resource-based view of the firm [5], [6], the evolution-oriented perspective of the firm [7], or the knowledge-based view of organizations [8], [9], have quantified the importance of knowledge, the systematic integration and efficient usage of knowledge in an organization's daily operations represents still a highly complex field of studies. When reviewing both academic and management literature (see e.g. [10], [11], [12]) of the previous years for approaches and frameworks of how to apply knowledge and/or intellectual capital management in organizations, it seems to be apparent that those frameworks are designed as either iterative or step-oriented procedures or as integration of both.

Organizations strive to transfer employee's knowledge into organizational memory. The reasons for that are the high volatility of human capital and the fact that it is harder to extract value from human capital than from the more

T. Yamaguchi (Ed.): PAKM 2008, LNAI 5345, pp. 135–146, 2008.

institutionalized and conceptualized structural capital [13]. The core human capital of an organization is often made up of a small number of individuals whose input is crucial to the firm's ability to generate either current or future revenues [13]. Therefore organizations have to take steps not only to externalize the knowledge of these individuals but also to share their knowledge among employees.

The paper is organized as follows: Section 2 describes the knowledge management approach according to Probst et al. [14], before in Section 3 the Intellectual Capital Report Benchmarking (ICRB) framework for managing and reporting an organization's capital will be described. Based on the results of Section 3, Section 4 will present the eduWEAVER approach, a method for organizing an organization's training concepts. Section 5 illustrates the advantages of combining the ICRB framework with the eduWEAVER approach and presents a conceptual model for enabling an organization to better fulfill the very important transfer of employees' knowledge into the organizational memory. Conclusions are presented in Section 6.

2 Modules of Knowledge Management

In the context of this article we define knowledge as the collective of skills and abilities that are applied by individuals for problem-solving [14]. Furthermore knowledge is based on information and data, whereas it is always linked to people.[1]

One of the most wide-spread concepts of knowledge management (KM) in organizations is the "Modules of Knowledge Management"[2] according to Probst, Raub, and Romhardt [14]. It serves as the foundation stone of the practical realization of the article's approach. The concept is divided into two parts: Core processes and pragmatic modules. Figure 1 depicts the six core processes and two pragmatic modules of the conception.[3] Measurable KM Goals are defined in the pragmatic modules, whereas the core processes are concerned with the actual execution of the activities.

Core processes:

- *Knowledge Identification:* The identification of knowledge within an organization is a critical task in large organizations. A transparent knowledge landscape should be the result of knowledge identification.

[1] Translated from German from Probst, Raub, and Romhardt 2003, p. 22.:"[Wissen ist] die Gesamtheit der Kenntnisse und Faehigkeiten, die Individuen zur Loesung von Problemen einsetzen. Wissen stuetzt sich auf Daten und Informationen, ist im Gegensatz zu diesen jedoch immer an Personen gebunden".

[2] Translated from German: Bausteine des Wissensmanagements.

[3] This figure is based on Probst, Raub, and Romhardt 2003, p. 32. The termini have all been translated from German.

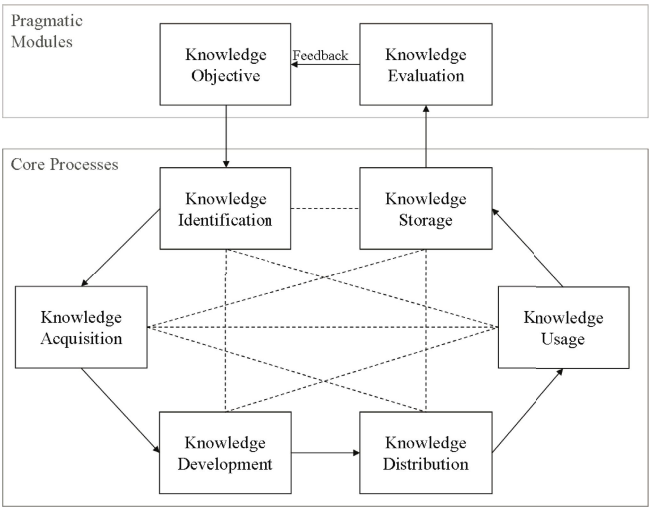


Fig. 1. Modules of Knowledge Management [14]

- *Knowledge Acquisition:* The huge amounts and fragmentation of knowledge makes it impossible to generate all necessary knowledge within the organization's memory. Therefore a demand exists to acquire knowledge (e.g. by buying innovative companies, patents, or other knowledge carriers).
- *Knowledge Development:* Generating new skills, products, ideas, and improved processes lies at the heart of knowledge generation.
- *Knowledge Distribution:* The correct distribution of knowledge influences the effectiveness of KM. The aim is to provide workers with the relevant knowledge at the right time.
- *Knowledge Usage:* As soon as knowledge has been distributed, it is essential that it is also applied in the organization's processes. Characteristics, like knowledge protection (e.g. via patents, trade secrets, or licenses) have to be considered.
- *Knowledge Storage:* Knowledge vanishes quickly and hence, has to be stored properly and accessible when needed.

Pragmatic modules:

- *Knowledge Objectives:* Defines strategic and tactical objectives for ensuring to steer knowledge management activities and their impacts into the right direction.
- *Knowledge Evaluation:* Measures the results of the KM-initiative after having executed the core processes.

In the following paragraphs Probst et al.'s [14] modules of KM will serve as a basis for the integration of two information systems into a comprehensive IT-based management approach for knowledge management:

– Intellectual Capital Report Benchmarking (ICRB) framework ([4], [15], [16])
– eduWEAVER approach ([17], [18], [19]).

3 The Intellectual Capital Report Benchmarking (ICRB) Framework

Intellectual capital is defined as *knowledge converted into profits* [13]. Its management and reporting is currently a major topic in research (as e.g. World Congress on Intellectual Capital 2008, European Congress on Intellectual Capital 2007). Although reporting of an organization's capital is yet not a standardized procedure, the structure of intellectual capital has been defined only lately as a composition of human, structural, and relational capital [20]. Human capital is defined as capabilities, knowledge, and expertise that is literally located in the employees' heads, structural capital is *what is left after the employees have gone for the night* [21] and finally relational capital determines the value brought into an organization by cooperating with external stakeholders, such as customers, suppliers, and others. An overview of the three intellectual capital types is illustrated in Figure 2.[4]

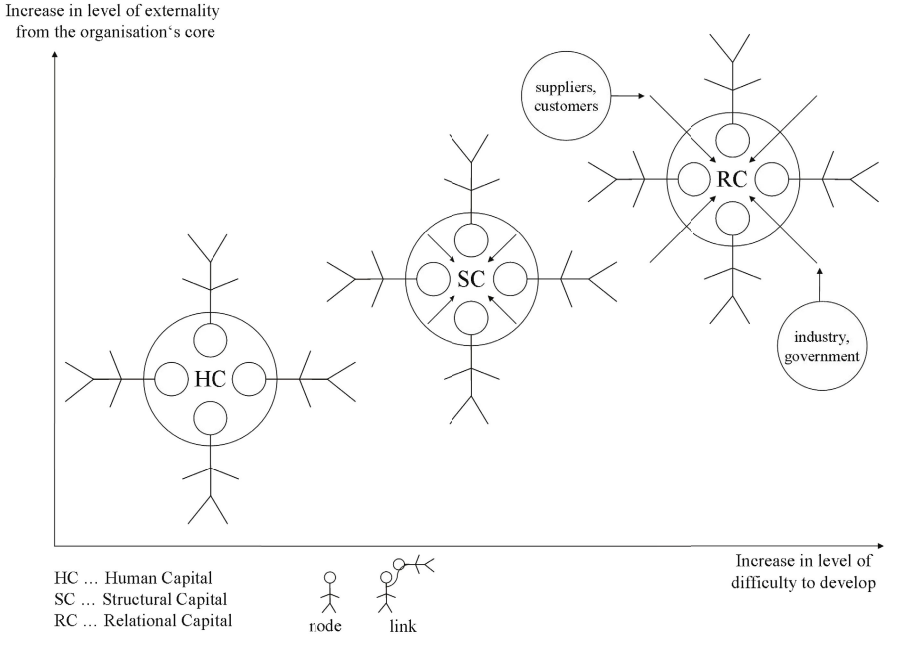


Fig. 2. Transformation of Human Capital to Structural Capital [9]

[4] The figure is based on Bontis 2000, p. 386.

Figure 2 aims to illustrate the important transformation of both relational and human capital into structural capital, as the latter is the only type of intellectual capital that can be owned by organizations. Human capital as well as relational capital may be lost as soon as either the employee or the external stakeholder ends working and/or collaborating for/with the organization [22]. However, this transformation represents a very complex procedure as current approaches for managing and reporting intellectual capital differ dramatically in structure and outcome, whereas the ICRB framework proposes a procedure for generating comprehensive, comparable, and expressive intellectual capital reports [4]. Those reports represent the basis for enhancing the process of the afore-mentioned transformation of human to structural capital by allowing to follow this transfer. The following paragraphs illustrate the conceptual architecture of the ICRB framework.

Based on the structure of intellectual capital (human, structural, and relational capital), the ICRB framework clusters so-called meta-indicators for diverse conceptions of intellectual capital reports. However, the absolutely static approach of reporting is further on combined with the dynamic structure of an organization's daily business. The well-known concept of business processes is nowadays applied in the majority of organizations. By proposing a three stage modeling concept, the ICRB framework combines the static and dynamic parts as it is illustrated in Figure 3.[5]

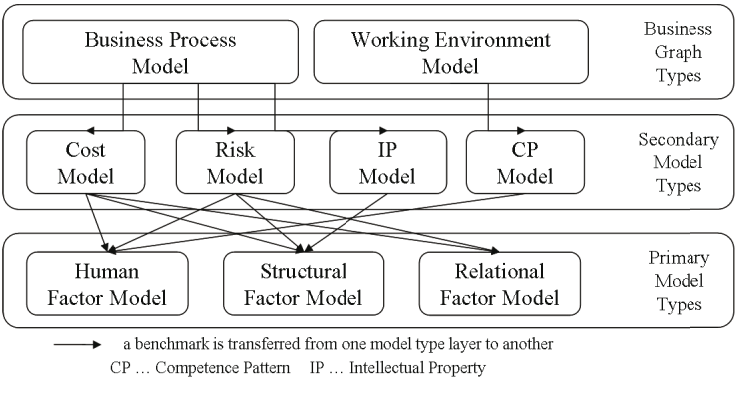


Fig. 3. Transferring Meta-Indicators in ICRB [4]

When starting from the top of Figure 3, business process models and working environment models are representing the dynamic parts of daily business. Based on these models, the second layer - so-called secondary model types - extract required information from the bottom layer by classifying positive and negative aspects of intellectual capital, whereas the former is realized by the intellectual

[5] The figure is based on Nemetz 2008, p. 194.

property (IP) and competence pattern (CP) model and the latter is represented by risk and cost models. Finally, the primary model types including the afore-mentioned human, structural, and relational factor model represent the third and final stage of the ICRB framework architecture. By selecting a desired concept of intellectual capital report, sets of meta-indicators are transformed by mathematical operations to key indicators of the respective intellectual capital report.

On the basis of these intellectual capital reports a diagnosis of lacks of intellectual capital seems to be possible. Thus, appropriate trainings can be designed for achieving a reduction of these knowledge gaps. For ensuring both an appropriate way of planning and executing trainings according to the results of intellectual capital reports, the eduWEAVER approach - a concept for process-based and modular training design - will be presented. The overall goal of combining the ICRB framework and the eduWEAVER approach is the realization of the afore-mentioned transformation of an organization's human capital to structural capital.

4 eduWEAVER – Personalized Education for People

eLearning scenarios that are more than just sequenced content are highly needed nowadays [23]. Both learning objects and course sequences should be enriched with pedagogical and didactical elements. These didactics need to be flexible and shall support easy adaptation for an instructors own purposes. Currently we see two major aspects in dealing with eLearning content:

- Re-usability and interoperability of learning objects
- Didactical and pedagogical enrichment of learning objects

To address these two aspects a web-based courseware modeling tool named eduWEAVER[6] was developed. eduWEAVER is based on the meta-modeling platform ADVISOR®[7] and supports design, development, and dissemination of educational content. Figure 4 shows the interfaces of eduWEAVER and how it interacts with instructors, learning management systems (LMS), and databases for learning materials and course structures.

First of all instructors can store learning materials and/or course sequences in eduWEAVER via the Compilation Interface. The Integration Interface enables the export of these materials and structures to LMS. The Personalized Interface allows the storage of one's own learning material within a personal learning object pool, which can be published through the Public Interface. In accordance with the respective instructor learning material and/or course sequences can be adapted by other instructors. This is possible via the Instructor Interaction Interface.

[6] http://www.eduweaver.net/
[7] BOC Information Technologies Consulting AG, http://www.boc-group.com

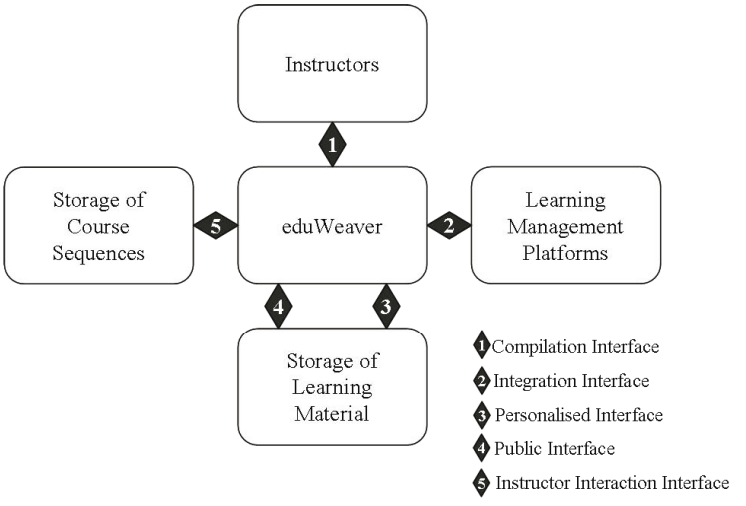


Fig. 4. Interfaces of eduWEAVER with its environment [19]

4.1 How to Create Educational Processes in eduWEAVER

An Instructor or course designer creates learning content by using a visual modeling language. For a comparison of eduWEAVER and other visual instructional design languages we refer to [24]. The learning object pool (LOP) represents a knowledge base where physical learning material like text-files, presentations, websites, etc. can be stored and organized. Course sequences can be build with learning objects from various learning object pools. For creation of learning objects, two principles should be taken into consideration:

- *The principle of re-usability:* to assure a certain quality of content so that the re-usable learning objects (RLOs)[8] need not be changed for different learner groups.
- *The principle of collaboration:* to share RLOs with other instructors to assure a continuously growing pool of learning materials.

4.2 Modeling Courses and Course Sequences

The eduWEAVER modeling method consists of four hierarchical layers [25]. Level 1 - the *Course Map* - provides a graphical overview of all courses of an instructor. Within level 2, the instructor creates *Modules* for all *Courses* of level 1 by splitting them into thematically coherent parts. The *Modules* themselves are broken down into several *Lessons* within level 3. Depending on the requirements

[8] A re-usable learning object (RLO) within eduWEAVER is defined as a learning unit of high cohesion that aims at a certain learning objective. To fulfill this objective it aggregates different information objects like hypertexts, pictures, animations, audio, and video. The optimal learning time of one RLO lies between 10 and 30 minutes.

for content and structure the instructor can create flexible teaching scenarios by including control structures like loops or conditional statements. A *Lesson* - modeled in level 4 - corresponds to a teaching unit and consists of RLOs. All these levels are hierarchically linked to each other by internal references.

4.3 Presenting Designed Courses

An instructor has the possibility to export learning objects or whole course models via an XML interface [17]. The export functionality includes HTML, XML, ADL[9], as well as SCORM 1.2, SCORM 2004[10], and IMS Learning Design Level A[11]. Therefore course sequences modeled in eduWEAVER are platform independent and can be used in all major LMS (e.g. Moodle, Blackboard, ...).

5 The Combination of ICRB and eduWEAVER

In a nutshell the combination of the ICRB framework and the eduWEAVER approach aims at diagnosing lacks of knowledge with intellectual capital reports and then create and distribute adequate learning materials and structures to eliminate these knowledge gaps.

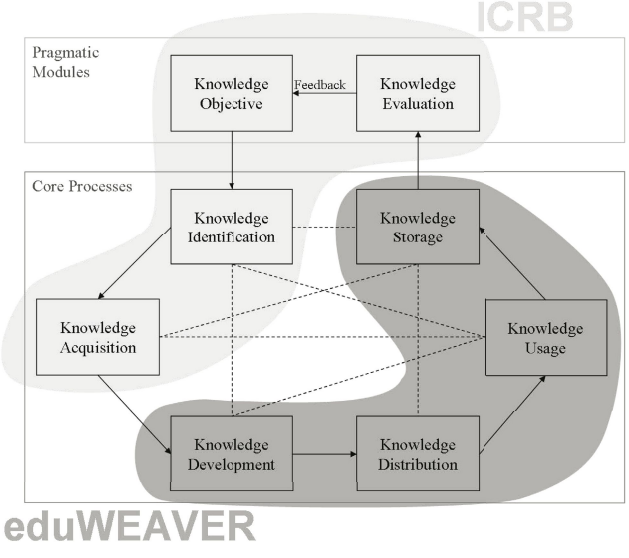


Fig. 5. ICRB and eduWEAVER in the Modules of KM

[9] Adonis Definition Language.

[10] SCORM is a collection of standards and specifications adapted from multiple sources to enable interoperability, accessibility and re-usability in eLearning [26].

[11] IMS Learning Design is a standardized modeling language for representing learning designs as a description of executable teaching and learning processes [27].

Figure 5 demonstrates which Knowledge Modules (cf. Section 2) are supported by the ICRB Framework and which by the eduWEAVER approach. The Pragmatic Modules (Knowledge Evaluation and Objective), as well as Knowledge Identification and Knowledge Acquisition are covered by the ICRB framework. eduWEAVER's central learning object pool (cf. Section 4.1) manages the Knowledge Storage. The export functionality of eduWEAVER (cf. Section 4.3) enables the Knowledge Distribution and Knowledge Usage. Due to eduWEAVER's web-based architecture collaborative creation and administration of learning materials and course structures can be attained, and thus Knowledge Development could be realized.

5.1 Identifying Knowledge Gaps with ICRB

Intellectual capital reports serve as a means for diagnosing lacking intellectual capital as well as for enhancing an organization's intellectual capital according to requirements of daily business. By either comparing intellectual capital reports from diverse years or by comparing intellectual capital reports of diverse organizations, it seems to be possible to derive lacks of knowledge in the organizational memory. Thus, appropriate trainings can be designed for achieving a reduction of these knowledge gaps. After identifying the knowledge gaps appropriate training is developed and executed to fill the gaps.

5.2 Transforming Human into Structural Capital

As mentioned above the human capital of organizations is highly volatile. So organizations try to generate structural capital out of human capital, as explicit codified knowledge (intellectual assets) is often easier to use for developing profits than tacit knowledge [13]. Structural capital is owned by the organization and is easier to be shared and reproduced [28]. Organizations may gain a significant added value by a systematic transformation of human capital into structural capital [29]. We propose a conceptual model for the transformation of human to structural capital that consists of four subsequent steps:

1. **Identify knowledge gaps:** By comparing intellectual capital reports knowledge gaps are identified.
2. **Find an expert:** An expert able to fill the gap identified in step 1 should be assigned.
3. **Turn individual into group knowledge:** The expert identified in step 2 externalizes his knowledge relevant to the gap.
4. **Institutionalize group knowledge:** Finally, the external representations of the expert's knowledge are distributed and used within the organization.

Figure 6 illustrates the four steps mentioned above. The ICRB conceptual framework together with the ICRB method unify existing concepts for intellectual capital management and reporting, and allow the creation of comparable,

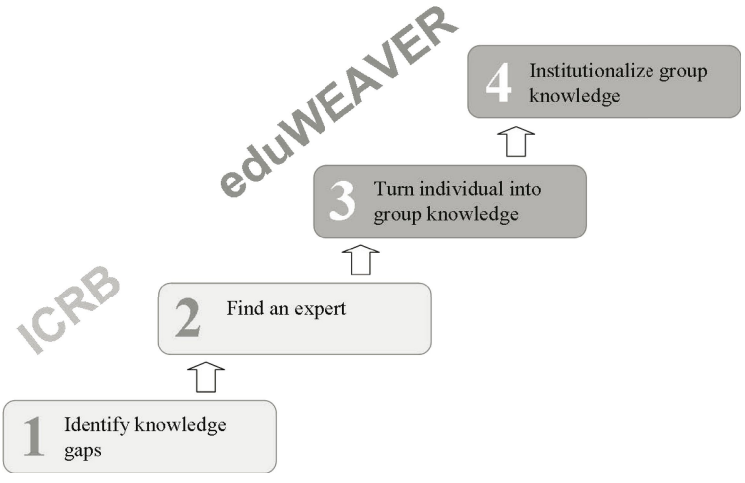


Fig. 6. Transformation of Human to Structural Capital with the help of ICRB and eduWEAVER

expressive, and comprehensive intellectual capital reports [4]. The automatically generated reports serve as a basis to identify knowledge gaps in step 1. In step 2 an expert that is able to provide the knowledge needed to fill up a potential knowledge gap is searched. Knowledge acquaintance within the organization is facilitated by the competence pattern models of the ICRB method. Competence pattern models extend role descriptions in working environment models with intellectual capital-relevant information, such as education, knowledge, experience, and skills [4]. Organizations have to address the knowledge of individuals first, as they are not able to create knowledge on their own without individuals [30]. Mindful work or management means that employees bring with them their visions, their ideas, and their listening to the organization [31]. Nonaka and Takeuchi [30] describe organizational knowledge creation as a process that organizationally amplifies the knowledge created by individuals and crystallizes it at group level [30]. Step 3 externalizes the individual knowledge of the expert. Externalization is most often seen as the conversion from tacit to explicit knowledge, often through the use of such techniques as analogies, models, or metaphors [30]. Thereby knowledge that cannot be formalized and communicated to others easily, is converted into knowledge that is transmittable and articulable, such as words or numbers.[12] eduWEAVER gives employees the possibility to design, develop, and distribute the external representations of their knowledge (or in other words learning materials). Step 4 is about the actual usage of the learning materials and courses created in step 3. The materials and structures created in eduWEAVER are utilizable in different ways (cf. Section 4.3).

[12] Of course not all forms of knowledge are easily being externalized, as for instance personal knowledge, skills, and relationships. We refer here especially to the dimension of tacit knowledge, as e.g. in [32].

6 Conclusions

This paper presented the integration of two information systems to a comprehensive IT-based management approach for KM on the basis of the Modules of KM. After a short description of the Modules of KM by Probst et. al [14] we provided a brief introduction into the ICRB framework and the eduWEAVER approach. The ICRB framework proposes a procedure for generating comprehensive, comparable, and expressive intellectual capital reports. On basis of these reports knowledge gaps are identified and appropriate training is designed, developed, and deployed with the eduWEAVER approach. Furthermore, we proposed a conceptual model that supports the transformation of human into structural capital. The next steps include practical examples in different domains to test our proposed approach.

References

1. Specht, D.: Wissensbasierte Systeme im Produktionsbetrieb. Carl Hanser Verlag, Munich (1989)
2. Deking, I.: Management des Intellectual Capital. Deutscher Universitats-Verlag, Wiesbaden (2003)
3. Andriessen, D.: Making Sense of Intellectual Capital. Elsevier Butterworth Heinemann, Oxford (2004)
4. Nemetz, M.: Intellectual Capital Management and Reporting. PhD thesis, University of Vienny (2008)
5. Barney, J.: Integrating organizational behavior and strategy formulation research: A resource based analysis. Advances in Strategic Management 8, 39–62 (1992)
6. Barney, J.: Resource-based theories of competitive advantage. Journal of Management 27, 643–650 (2001)
7. Nelson, R.R., Winter, S.G.: An Evolutionary Theory of Economic Change. The Belknap Press of Harvard University Press, Cambridge (1982)
8. Grant, R.M.: The knowledge-based view of the firm. Long Range Planning 30(3), 450–454 (1997)
9. Bontis, N.: Managing organizational knowledge by diagnosing intellectual capital. In: Morey, D., Maybury, M., Thuraisingham, B. (eds.) Knowledge Management: Classic and Contemporary Works. MIT Press, Cambridge (2000)
10. Karagiannis, D., Telesko, R.: Wissensmanagement - Konzepte der Kuenstlichen Intelligenz und des Softcomputing. Lehrbuecher Wirtschaftsinformatik. Oldenbourg, Munich (2001)
11. Sveiby, K.E.: A knowledge-based theory of the firm to guide in strategy formulation. Journal of Intellectual Capital 2(4), 344–358 (2001)
12. Reimer, U., Karagiannis, D. (eds.): PAKM 2006. LNCS (LNAI), vol. 4333. Springer, Heidelberg (2006)
13. Sullivan, P.H.: Value Driven Intellectual Capital: How to Convert Intangible Corporate Assets into Market Value. John Wiley and Sons, Inc., New York (2000)
14. Probst, G., Raub, S., Romhardt, K.: Wissen managen. Gabler Verlag, Wiesbaden (2003)

15. Nemetz, M.: A meta-model for intellectual capital reporting. In: Reimer, U., Karagiannis, D. (eds.) PAKM 2006. LNCS (LNAI), vol. 4333, pp. 213–223. Springer, Heidelberg (2006)
16. Nemetz, M., Karagiannis, D.: Intellectual capital and it: Requirements and applications. In: Andriessen, D. (ed.) European Congress on Intellectual Capital, Haarlem, Netherlands (2007)
17. Waldner, F., Nemetz, M., Karagiannis, D.: The eduweaver contents approach: From provision to instruction. In: Uskov, V. (ed.) Web Based Education (WBE), Innsbruck, Austria (2008)
18. Waldner, F., Nemetz, M., Steinberger, C.: eduweaver: Integrated design, development and deployment of elearning scenarios. In: International Conference on Internet and Web Applications and Services, Athens, Greece. IEEE Computer Society Press, Los Alamitos (2008)
19. Karagiannis, D., Nemetz, M., Fochler, S.: Web-bases course design on the basis of eduweaver and advancements. In: Auer, M. (ed.) International Conference on Interactive Computer Aided Learning (ICL), Villach, Austria. Kassel University Press (2007)
20. Saint-Onge, H.: How knowledge managements adds critical value to distribution channel management. Journal of Systemic Knowledge Management (1998)
21. Wiig, K.M.: Integrating intellectual capital and knowledge management. Long Range Planning 30(3), 399–405 (1997)
22. Daum, J.H.: Intangible Assets oder die Kunst, Mehrwert zu schaffen. Galileo Business (2002)
23. Yu, D., Zhang, W., Chen, X.: New generation of e-learning technologies. In: First International Multi-Symposiums on Computer and Computational Sciences IMSCCS, vol. 2, pp. 455–459. IEEE Computer Society Press, Los Alamitos (2006)
24. Figl, K., Derntl, M.: A comparison of visual instructional design languages for blended learning. In: Proceedings of World Conference on Educational Multimedia, Hypermedia and Telecommunications 2006, Chesapeake, VA. E. Pearson, P. Bohman, London (2006)
25. Steinberger, C., Bajnai, J., Ortner, W.: Another brick in the courseware or how to create reusable learning objects. In: Kommers, P., Richards, G. (eds.) World Conference on Educational Multimedia, Hypermedia and Telecommunications (EDMEDIA) 2005, Montreal, Canada, pp. 1064–1071 (2005)
26. Dodds, P., Thropp, S.E. (eds.): SCORM, 3rd edn. Advanced Distributed Learning (2004), `http://www.adlnet.gov/`
27. Koper, R., Olivier, B., Anderson, T. (eds.): IMS Learning Design Information Model. IMS Global Learning Consortium (2003), `http://www.imsglobal.org/learningdesign/`
28. Stewart, T.: Intellectual Capital. Nicholas Brealey Publishing, London (1997)
29. Edvinsson, L.: Some perspectives on intangibles and intellectual capital. Journal of Intellectual Capital 1(1), 12–13 (2000)
30. Nonaka, I., Takeuchi, H.: The Knowledge-Creating Company. Oxford University Press, New York (1995)
31. Baecker, D.: The form of the firm. Organization: The Critical Journal on Organization, Theory and Society 13(1), 130–142 (2006)
32. Polanyi, M.: The tacit dimension. In: Prusak, L. (ed.) Knowledge in Organizations, pp. 135–146. Butterworth-Heinemann (reprint), Newton (1997)

Developing a Reference Method for Knowledge Auditing

Theodoros Levantakis, Remko Helms, and Marco Spruit

Institute of Information and Computing Science, Utrecht University,
Padualaan 14, 3584 CH Utrecht, The Netherlands
{tlevanta,helms,spruit}@.cs.uu.nl

Abstract. Knowledge management was regarded as the discipline of the 21st century but it has yet to bear all of its fruits. The expectations are high but a number of barriers limit the effects of knowledge management initiatives. A serious problem is the lack of measures to prepare the ground for the successful implementation of knowledge management initiatives. A promising solution to this problem is the area of auditing knowledge. But the theory and the techniques behind it are not yet mature. The following paper is presenting the development of a complete knowledge audit method that could serve as a standardized method for auditing knowledge. The proposed method was constructed by comparing/integrating 13 knowledge and information audit methods. The analysis of the existing methods was based on the notion of method engineering and the new method was applied and validated in a Dutch telecommunications company.

Keywords: Knowledge management, knowledge audit, method engineering, social network analysis, knowledge strategy process.

1 Introduction

Knowledge management and related strategy concepts are considered necessary components for organizations to thrive and maintain their competitive advantage [1], [2]. Consequently, both public and private sectors started implementing knowledge management initiatives. One survey shows that up to 72% of firms within Europe have some kind of a KM initiative underway [3]; this figure is even up to 80% according a global survey [4]. Still, knowledge management is not fully exploited and many KM initiatives fail [5]. An estimated 80 percent of KM projects, for example, have little or no impact on the organizations [6]. This resulted in a discussion on how to conduct KM initiatives and consequently several methods have been proposed [7]. However, different companies need different knowledge management solutions, and hence ad hoc or standard solutions are not a good option [8]. Therefore, investigating the unique environment and knowledge culture of an organization is a necessary step in a method for conducting KM initiatives. This step is also referred to as a knowledge audit. According to Hylton [7], many of the mistakes that both earlier and more recent adopters of knowledge management originate from, is the serious oversight of not including the knowledge audit in their overall KM strategies and method for conducting their KM initiatives. Furthermore, we found that even though

T. Yamaguchi (Ed.): PAKM 2008, LNAI 5345, pp. 147–159, 2008.

a knowledge audit can help to prepare the ground for implementing successful KM initiatives, the theory behind it is still rather abstract. This led to the question that dictated our research:

> *How should a reference method for auditing knowledge be formulated in order to prepare the ground for knowledge management initiatives?*

This paper presents the results of developing a reference method for knowledge auditing and has been validated in a case study in a Dutch telecommunications company. It is based on existing knowledge audits, but it is more than just a combination of existing methods. Firstly, because the goal is to be more complete and detailed than existing methods. Hence, it is assumed to be more valuable for practitioners in the KM domain. Secondly, also new concepts are introduced that are not part of existing methods and are derived from knowledge strategy formulation and social network analysis.

In the next section, an introduction and literature review concerning the development of knowledge audit methods is presented. Next, the third section will present the research method(s) used to develop the new knowledge audit method. The fourth section presents the steps of assembling the method and finally conclusions and suggestions for future research are provided in the fifth section.

2 Knowledge Audit

Section 2.1 provides an insight on the theory behind information and knowledge audit methods. Section 2.2 explains how an inventory of information audit methods was made and which criteria were used to include a method in this research.

2.1 Introduction to Information and Knowledge Auditing

Information audit as a term was first mentioned by Reynolds back in 1980 (as cited by [9]). Initially, an information audit was considered a process used in the user requirement phase of a management information system project where the ranking and the organizing of the information resources of the system was conducted. Later, in the 1990s information started to gain popularity as an integral strategic resource. As a result, "information audits extended the concept of auditing by evaluating an organization's accounting and financial procedures to that of the organizations overall information system" [10]. The term was used to suggest a method for the identification, evaluation, and management of information resources in order to fully exploit the strategic importance of information [9]. Typical objectives of an information audit include identifying an organisation's information resources (1) and identifying an organisation's information requirements (2). Later, Buchanan and Gibb [9] added more objectives so that the audit would become a truly comprehensive and integrated strategic approach. The most important objectives included: integrating IT investments with strategic business initiatives (1) and identifying information flows and processes (2).

Debenham and Clark [11] can be considered as proponents of the knowledge auditing theory and were among the first to introduce the term 'knowledge audit' in

1994. Knowledge auditing can be regarded as a successor of information auditing. Information audits mainly focused on codified or explicit knowledge, whereas, knowledge auditing also "tries to identify, evaluate and manage tacit knowledge along with explicit knowledge" [9]. Hence a knowledge audit can be defined as an assessment that "incorporates all the effective processes associated with the exploration (such as identify, evaluate, manage) of human knowledge (tacit and explicit) within a business unit or an organization" [32]. Such an audit is typically conducted before embarking on a KM initiative in order to make an inventory of current knowledge and knowledge processes in the organization and to assess if there is a gap between current situation and the desired situation that is required to achieve business goals [32].

In the next section the most commonly adopted methods for auditing knowledge will be presented. As we will see, the similarities between information audits and knowledge audits are, not surprisingly, great. Hence, information audit and knowledge audit methods were regarded of equal importance in this research.

2.2 Inventory of Existing Methods

The literature study was conducted using mainly the following sources: Citeseer, Google Scholar, Scirus, Picarta (national library database), and our university library. The main keywords for conducting the search on existing auditing methods were "knowledge audit", "information audit", "auditing knowledge" and "knowledge inventory". Not all methods that were found were included in our research. The following criteria have been used to select methods:

1. The methods should be available in open literature and published in textbook so that sufficient information is available
2. They should be methods that prescribe specific steps to follow
3. The methods should specifically engage on auditing information or knowledge

In total, thirteen methods have been found that satisfied these criteria: [9, 11–22]. In the remainder of this paper is explained how a new reference method for knowledge auditing has been derived from these methods.

3 Research Method

3.1 Design Research Approach

A design research approach has been applied to develop the reference method for auditing knowledge [23], [24]. Although an entirely new method was not created, the research can be considered as design research as the new method is a combination of existing methods to which some new elements were added. The design research cycle suggested by [24] and by [25] dictated the research. In a nutshell the phases of our design research cycle include:

- Awareness of the problem – Failing KM initiatives/abstract knowledge auditing theory
- Suggestion – Comparison of existing methods to determine differences/ overlap

- Development – Creation of reference method based on the comparison
- Evaluation – Case study for validation of reference method
- Conclusion – Conclusion and further research

Method development in phase 2 and 3 is based on techniques from the Method Engineering domain [27-28]. The process that is applied in this research for method development is discussed in the next section.

3.2 Method Development

Method engineering forms the theoretical background on how to develop (engineer) the new knowledge audit method. A method is defined as an "*approach to perform a systems development project, based on a specific way of thinking, consisting of directions and rules, structured in a systematic way in development activities with corresponding development products*" [27]. Although, it originally refers to IS development method it can also be applied to other methods. A meaningful part of a method is also called a method fragment [28]. Hence, a method can be decomposed in method fragments. Method engineering is defined by Brinkkemper as *"the engineering discipline to design, construct and adapt methods, techniques and tools for the development of information systems"* [27]. This can refer to the development of a completely new method or to the assembly of method fragments, originating from different existing methods, into a new method. Assembly of fragments entails more than just putting together the best fragments of each method. Fragments are assembled in an intelligent manner where the fragments are integrated and aligned with each other [33].

Table 1. Method development process

Step 1: Define steps/outline of new reference method for knowledge auditing
Step 2: Identify best methods for donating method fragments to the new reference method
Step 3: Select method fragments to be included in new reference method
Step 4: Assemble and describe new reference method for knowledge auditing

In this research, we developed a new reference method for knowledge auditing using the four steps described in table 1 and is based on Hong et al.'s comparison method [26]. The first step in the development method is to define an outline of the new reference method for knowledge auditing that outlines the main steps of the new reference method. Hong et al. also refer to this as the super method and defines it as the smallest common denominator of all activities covering all of the existing knowledge audit methods [26]. Hence, the super method is by definition more complete than each of the individual methods. In step 2, we choose a number of methods, out of the existing thirteen methods, as possible donators of method fragments to the new reference method. The method fragments that are actually donated are selected in step 3. This is done by first comparing the selected method against the outline of the reference method, i.e. super method (also based on Hong et al.'s approach). Based on this comparison those method fragments are selected that best match one or more activities in the outline of the reference method. In case more than one fragment is found the most descriptive

or detailed one is selected. Finally, in the fourth step the selected method fragments are assembled and the complete method is described.

In order to be able to compare methods we first described the existing methods in a standardized format using a meta-modeling technique called Process-Deliverable Diagram (PDD) [29]. A PDD is the integration of two separate diagrams, the meta-process diagram (process view) and the meta-deliverable diagram (deliverable view). The meta-process diagram is based on a UML activity diagram and the deliverable view is based on a UML class diagram [29]. An example of a PDD is shown in figure 1. Finally, we also completely described the new reference method for knowledge auditing using the PDD technique.

4 Development of the New Knowledge Audit Method

This section describes the development of the new reference method for knowledge auditing according the four steps as described in the previous section.

4.1 Defining the Outline for the New Reference Method (Step 1)

To find this list of activities that constitute the outline of the new reference method, all existing knowledge audit methods have been analyzed in detail by creating meta-process diagrams using the PDD technique from [29][1]. In creating the PDD's, we tried to stay as close as possible to the original terminology as was used by the author(s). Afterwards, the PDD's were compared and it was found that there was much difference in the labeling of activities and deliverables and also in the level of detail that was used to describe a method. In identifying an outline for the new reference method, it was decided to go through the identification in a top down manner, i.e. from high level to a lower, more detailed level. The result of the high level comparison is the distinction of eight main activities for the new reference method:

- Prepare audit: in this stage the initial meetings with the management take place, the scope and the objectives of the audit are defined and the environment of the targeted area is investigated
- Promote audit's benefits: the benefits of the audit are discussed with top management and employees are invited to take part in the process
- Investigate targeted area: the identification of the business processes and the stakeholders involved takes part in this stage
- Collect data: collecting data through interviews and/or surveys is facilitated here.
- Analyze data: the data are processed and the end products are graphs depicting knowledge flows, the knowledge inventory and others.
- Evaluate data: the identification of problems and bottlenecks starts along with suggestions to overcome these difficulties
- Conclude audit: the main product is the audit report and the action plan
- Re-auditing: the suggestion that the audit should be continuous is observed in most of the methods and it was included as the last activity of the supermethod

[1] Due to space limitations it is not possible to show all PDD in this paper. A report containing all PDD's can be downloaded from: http://people.cs.uu.nl/remko/

In the next step, each of these activities was decomposed into further sub-activities in accordance with the steps of the analyzed methods and the end result was a complete list of activities and sub-activities of the supermethod (see columns 1 and 2 in table 2) that covered the ground of all the methods.

4.2 Identification of Best Methods for Donating Method Fragments (Step 2)

To determine which method would be the best 'donator' of a method fragment to the reference method, we first compared each existing method with the reference method to find differences and similarities. This is done by creating a comparison table as shown in Table 2. All notations and the way of structuring the comparison table have been adopted from [26].

Table 2. Comparison of reference method with existing audit methods

Activity	Sub-activity	Orna	Henc zel	Burn ett	Perez-Soltero	Che ung
1. Preliminary	1.1 Organization brief			=	=	=
	1.2 Observe targeted area	=	>	><	=	=
	1.3 Set objectives of audit	>	=			
	1.4 Set scope of audit		=			=
2. Promote audit	2.1 Ensure management support	=	=			
	2.2 Foster collaboration with employees	=	=	=		
3. In depth investigation	3.1 Identify business processes		>		=	
	3.2 Identify key actors				=	>
	3.3 Identify knowledge requirements	=				
	3.4 Identify knowledge processes			=	><	
4. Collect data	4.1 Meet key actors				=	
	4.2 Conduct survey/interviews	=	><	=	=	=
5. Analyze data	5.1 Build knowledge inventory	>	><	><	=	=
	5.2 Build knowledge map	>	=	><	=	=
	5.3 Execute Social network analysis					>
	5.4 Execute Gap Analysis	><				
6. Evaluate data	6.1 Identify bottlenecks and gaps		=	><	>	>
	6.2 Prioritize problems		=			
	6.3 Suggest solutions	=	=	=	=	=
	6.4 Rank solutions		=	=		
	6.5 Develop action plan	=	=	=		>
7. Conclude audit	7.1 Write audit report	=	=	><	=	=
	7.2 Present results	=	=			
	7.3 Get approval of action plan	=	=	=		>
8. Re-audit		=	=		=	=

The first two left columns of the comparison table display the activities of the reference method. The remaining columns show how well the existing knowledge audit methods match with the reference method. If a cell in the comparison table is blank, this means that the activity on that row is not present in the method of the corresponding column. In any other case, there are three notations describing the relationship between the reference method and the other methods that should be taken into account when looking at the table.

- An "=" symbol indicates that a similar activity to the one of the super method is available in the method of the corresponding column
- An "<" or an ">" symbol indicates that the activity of the method of the corresponding column does more or less than the activity of the super method, respectively
- An "><" symbol indicates that a part of the activity of the method of the corresponding column overlaps a part of the activity of the super method and the other parts of both activities do not overlap

Table 2 does not show the complete comparison table, due to space limitations (see footnote 1). However, it does show that there are distinct differences in the activities that each method supports. Furthermore, there is no single method that covers all the activities of the reference method; the best match is Henczel's method that covers 18 out of 25 activities of the reference method.

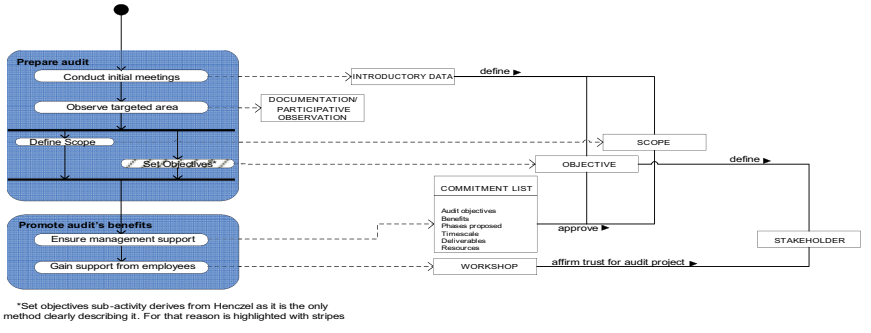


Fig. 1. Part of the knowledge audit method

4.3 Selection of Method Fragments and Assembly of the New Reference Method (Step 3 and 4)

Based on the comparison table (table 2), a corresponding method fragment is selected that best matches a particular activity or set of activities in the reference method. In those cases, where there were two or more method fragments to choose from, the criteria for selecting a particular method fragment was the amount of information provided for the deliverables of the method. Finally, we also took into account to what extent the method fragment was aligned with other selected method fragments. As already stated, there was a lack of terminology and information on the deliverables of activities in the existing methods. Thus, method fragments that provided a

substantial amount of information on the deliverable of an activity and/or providing templates and details on supporting techniques were selected against other similar method fragments with less information. A small part (containing the first two activities) of the method that was developed is shown in figure 1. The left part is the meta-process model which shows the processes, and the right part is the meta-deliverable model, which shows the deliverables of the respective processes. Due to space limitations it is not possible to describe the complete reference method for knowledge auditing here as it covers many pages. A link to the document containing the complete description of the reference method is provided in footnote 1.

5 Validation of the New Knowledge Audit Method

The iterative approach for applying the reference knowledge audit method made it possible to test and refine the method during the case study.

5.1 Case Study Organization

The case study organization is a Dutch Information and Communication Technology company, Getronics PinkRoccade. The case study was conducted in the Business Unit of Consulting (BUCO) which is specialized in offering solutions to a broad spectrum concerning IT-management. The target group within this business unit consisted of 58 consultants that are predominantly working on client's sites which are located within the Netherlands. The projects carried out are primarily on areas such as project management, and service and performance management. In the past, the business unit had attempted several times to implement a successful knowledge management system. The organization tried both extremes; either focusing on the people and the transfer of the so called tacit knowledge or focusing on the computer based technology and mainly the storage of and access to the so called explicit knowledge. Managing knowledge became even harder after the acquisition of PinkRoccade. This becomes more obvious within the business unit under investigation as most of the employees there come from PinkRoccade. Systems that were successful within PinkRoccade where shut down but there were no systems to properly replace them.

This history of failed implementations makes BUCO an interesting case to see if the knowledge audit can be used to prepare the ground for a new knowledge management initiative.

5.2 Evaluation of the Reference Knowledge Audit Method

The case study at BUCO was conducted to evaluate the newly developed knowledge audit method. It was conducted by the main author of this paper and took approximately two months to complete. Besides the author, the manager of the BUCO unit was involved who acted as the main sponsor of the project. Also the consultants of the BUCO unit were involved. First of all, the consultants had to provide input for the questionnaire and interviews that are part of the reference knowledge audit method. Secondly, the consultants provided feedback on the usefulness and completeness of the method if appropriate. Progress of the knowledge audit was regularly reported in department meetings once per week. During the execution of the

knowledge audit the researcher could evaluate the method himself. This resulted in two big changes to the method. The first problem that was encountered concerned the identification of knowledge that supports the business processes (3.1 - 3.3 in the reference method). The method fragments that were included were not descriptive enough to provide guidance to identify the required knowledge. To improve this part of the reference method, parts of the Knowledge Strategy Process by Van der Spek et al. were adopted [30]. It especially concerned the part that describes how knowledge areas can be identified within an organization. The second problem that was encountered concerned the construction and analysis of the knowledge map. A knowledge map provides an overview of who shares knowledge with who, i.e. shows the knowledge flows through the company (step 5.2, 5.3 and 6.1). The method fragment that was selected provided some instructions for executing a social network analysis. However, the analysis part in this method was rather straight forward. Therefore, this part was replaced by the Knowledge Network Analysis technique, which provides more elaborate guidelines of bottlenecks in knowledge networks [31].

The final version of the reference knowledge audit method was successfully applied in the business unit. Eight semi-structured interviews were conducted with identified key people. These interviews focused on:

- Verifying the importance of the identified knowledge areas,
- Providing with information about the systems installed in the business unit and what specific knowledge they supported,
- Identifying the knowledge requirements supporting these knowledge areas

Three crucial knowledge areas were identified and seventeen knowledge requirements supporting these three areas. Of the 65 people of the business unit, it was decided not to include 7 because their work position was not relevant to the audit (secretarial and technical job positions). Initially 44 out of 58 people filled out the questionnaire, which is a response rate of 76%. After an initial checking of the responses, four were excluded as insufficient bringing the final number of valid responses to 40. The response rate was reduced to 69% which is regarded as representative for the BUCO population where the audit was conducted [14].

The results of the survey helped to identify the expertise level of the respondents regarding the three identified knowledge areas and the knowledge requirements of each of these areas (identify existing knowledge). The existing knowledge of the respondents and the systems supporting the knowledge requirements were inputs for the construction of the knowledge inventory list. The knowledge inventory was constructed and from analyzing it, three problems were identified.

The knowledge inventory helped facilitate the execution of the gap analysis. The existing knowledge that is depicted in the knowledge inventory was compared against the desired level of expertise concerning the knowledge requirements and four main gaps were identified.

Social and knowledge network analysis was executed with the use of NetMiner. Separate network graphs were constructed for each of the identified knowledge areas and by integrating some relevant data like the expertise level or the frequency of knowledge transfer more graphs were produced that helped a more thorough analysis. For each of the graphs produced, there was a quantitative and a qualitative analysis.

Eighteen bottlenecks were identified which also related with the problems identified by the knowledge inventory and the gap analysis, thus, adding to their validity.

The next step was to present the identified problems and evaluate them within the context of the organization [14]. In a meeting with the top management all 25 problems (barriers, gaps and bottlenecks) were evaluated and ranked according to their significance. In the same meeting, the possible solutions suggested by the authors were discussed and ranked according to their suitability, effectiveness and cost. 8 of the 25 problems were regarded as highly significant. All possible recommendations for these 8 problems were measured regarding their suitability in terms of their effectiveness, cost, and compatibility with the existing or future knowledge management systems. 5 of the recommendations scored highly and will be taken into consideration for integrating them in the next knowledge management initiative that is on its way.

The final version of the reference knowledge audit method was evaluated by 5 persons. Main goal of the evaluation was to check for face and content validity. First of all, the clarity of the activities and deliverables of the reference method was evaluated. A consultant within the business unit of Getronics PinkRoccade, with expertise in the knowledge management area, approved the final version of the method by stating that the steps were clear and the expected results should satisfy the goals of the project. The method was also explained thoroughly to two consultants of BUCO. The feedback was positive as they could comprehend all of the steps depicted in the process deliverable diagrams of the knowledge audit method and they verified the importance of the deliverables.

The three consultants that were mentioned above also evaluated the reference method in terms of its applicability, usability and effectiveness. They were asked to score these criteria in a scale of 0-5. The method scored 4 out of 5 in applicability and effectiveness and a 3.5 in usability. Their most valuable feedback was that there was no need of expert knowledge in the area of auditing to carry out the method. They stated that the only prerequisites were some experience in the field of knowledge management and a capability to execute network analysis. Because the consultants did not have any auditing experience we also asked two practitioners of auditing knowledge (outside Getronics) to review the new knowledge audit method in terms of its potential usefulness and effectiveness. Their judgment was based on the method development report that was send to them by email. The remarks were very positive regarding the applicability and the completeness of the method. One practitioner had some doubts regarding its consistency but after further explanation regarding the relations between the gap analysis and the network analysis he reacted positively. Both of them were intrigued by the application of the Knowledge Strategy Process and agreed that it was a valuable contribution to the knowledge audit method enhancing its applicability and adjustability. The two practitioners were asked to evaluate the method in terms of its applicability, completeness, consistency, and clarity. In all these terms the method scored 3.5 and above by both practitioners. One last question was if the method can provide the expected results of a knowledge audit method. Both of the practitioners stated that the method is very detailed with explicit explanation of the steps and techniques to follow and if executed properly it should deliver the necessary results. One suggestion for improvement though, was to emphasize more on the tacit deliverables of the method.

6 Conclusions and Further Research

A "super" method was developed according to the Hong's comparison method [26]. The method was applied in the consulting business unit with success. All the objectives stated in the start of the project were met. Suggestions provided in the knowledge audit report can be implemented in the new knowledge management initiative to ensure its success. Instead of just implementing a promising new knowledge management system without preparation, the knowledge audit helps the organization to identify problems that should be taken into consideration. For example, the identification of knowledge gaps should engage the top management to assure the support the cultivation of this missing knowledge either by hiring an external expert or integrating some system that supports it. The knowledge audit method should be considered as a necessary step before the implementation of any new knowledge management initiative, in order to define exactly what needs to be implemented.

One crucial limitation of this research was the limit space which resulted in analyzing in detail five of the thirteen methods. Of course, the five selected methods covered the whole spectrum of activities and were the most elaborated ones, but in the future it is possible to analyze all of the methods with the use of meta-modeling. What is more, it is possible to include in the analysis, methods that focus on only one activity of the supermethod or even a sub-activity resulting in a much more extensive background that could resulted in a refined knowledge audit method.

Secondly, the knowledge audit method should be tested in different kind of companies to assess its adaptability, consistency and applicability in all possible circumstances. In case that some difficulties emerge it should be updated to improve its applicability if it is to be considered a standardized method.

Thirdly, it is possible to decompose further the activities and sub-activities of the knowledge audit method in order to have a lower-level view. Finally, as knowledge management is evolving the new knowledge audit method should be updated to keep its consistency and applicability. New techniques and methods are developed very often and if proven sound and relative they should be considered for integration to the method. It should be noted that a knowledge audit does not guarantee the success of a knowledge management initiative but its purpose is to improve its chances of success.

Acknowledgments

We would like to thank Ankie Eeken (former Getronics PinkRoccade) for her support and valuable contribution for initiating this research. Special thanks to Anne-Marie Soppe (Getronics PinkRoccade) for her forbearance and support throughout the duration of the project. Finally, we would like to thank Sjaak Brinkkemper and Inge van de Weerd from the Department of Information and Computing Sciences of Utrecht University for their input and feedback concerning the use of Method Engineering in this research.

References

1. Davenport, T.H., Prusak, L.: Working Knowledge: How Organizations Manage what They Know. Harvard Business School Press, Boston (1998)
2. Martensson, M.: A critical review on knowledge management as a management tool. Journal of Knowledge Management 3, 204–216 (2000)
3. Mertins, K., Heisig, P., Vorbeck, J.: Knowledge Management – Concepts and Best Practices. Springer, Heidelberg (2003)
4. Allee, V.: Knowledge Networks and Communities of Practice. OD Practiotioner 32, 4–13 (2000)
5. Chua, A., Lam, W.: Why KM projects fail: a multi-case analysis. Journal of Knowledge Management 9, 6–17 (2005)
6. Lucier, C., Torsiliera, J.: Why knowledge programs fail: A CEO's Guide to Managing Learning. In: Cortada, J.W., Woods, J.A. (eds.) The Knowledge Management Yearbook 1999-2000, pp. 262–279. Butterworth-Heinemann (1999)
7. Hylton, A.: A Knowledge Audit must be People-Centered and People Focused. Hylton Associates. ITtoolbox Knowledge Management Knowledge Base (Retrieved November 1, 2002), http://knowledgemanagement.ittoolbox.com/browse.asp?c=KMPeerPublishing&r=%2Fpub%2FAH
8. Edwards, J.S., Shaw, D., Collier, P.: Knowledge management systems: finding a way with technology. Journal of Knowledge Management 9, 113–125 (2005)
9. Buchanan, S., Gibb, F.: The Information Audit: an Integrated Strategic Approach. International Journal of Information Management 18, 29–47 (1998)
10. Ellis, D., Barker, R., Potter, S., Pridgeon, C.: Information Audits, Communication Audits and Information Mapping: A Review and Survey. International Journal of Information Management 13, 134–151 (1993)
11. Debenham, J., Clark, J.: The Knowledge Audit. Robotics and Computer Integrated Manufacturing Journal 11, 201–211 (1996)
12. Orna, E.: Practical Information Policies. Gower Publishing, Ltd., Aldershot (1999)
13. Liebowitz, J., Rubenstein-Montano, B., McCaw, D., Buchwalter, J., Browning, C., Newman, B., Rebeck, K.: The Knowledge Audit. Knowledge and Process Management 7, 3–10 (2000)
14. Henczel, S.: The Information Audit: A Practical Guide. Saur, Munich (2001)
15. Lauer, T.W., Tanniru, M.: Knowledge Management Audit – A Methodology and Case Study. Australian Journal of Information Systems (Special Issue on Knowledge Management), 23–41 (2001)
16. Hylton, A.: A KM Initiative is Unlikely to Succeed without a Knowledge Audit. Knowledge Board (Retrived November 1, 2007), http://www.knowledgeboard.com/library/the_need_for_knowledge_audits.pdf
17. Abell, A.: Conducting an information audit. TFPL (Retrived November 12, 2007), http://www.tfpl.com/assets/applets/Conducting_an_info_audit.pdf
18. Burnett, S., Illingworth, L., Webster, L.: Knowledge Auditing and Mapping: A Pragmatic Approach. Knowledge and Process Management 11, 25–37 (2004)
19. Iazzolino, G., Pietrantonio, R.: An innovative knowledge audit methodology: some first results from an ongoing research in Southern Italy. In: Accettato alla KMAP International Conference on Knowledge Management, University of New Zeland (2005)
20. Perez-Soltero, A., Barcelo-Valenzuela, M., Sanchez-Schmitz, G., Martin-Rubio, F., Palma-Mendez, J.T.: Knowledge Audit Methodology with emphasis on core processes. In: European and Mediterranean Conference on Information Systems 2006, Costa Blanca Alicante (2006)

21. Cheung, C.F., Li, M.L., Shek, W.Y., Lee, W.B., Tsang, T.S.: A systematic approach for knowledge auditing: a case study in transportation sector. Journal of Knowledge Management 11, 140–158 (2007)
22. Biloslavo, R., Trnavcevic, A.: Knowledge Management Audit in a Higher Educational System: A Case Study. Knowledge and Process Management 14, 275–286 (2007)
23. Hevner, R.A., March, T.S., Park, J., Ram, S.: Design Science in Information Systems Research. MIS Quarterly 28, 75–105 (2004)
24. Vaishnavi, V., Kuechler, W.: Design Research in Information Systems. Association for information systems (2007) (Retrived January 15, 2008), http://www.isworld.org/Researchdesign/drisISworld.htm
25. Takeda, H., Veerkamp, P., Tomiyama, T., Yoshikawa, H.: Modeling Design Processes. Al Magazine 11, 37–48 (1990)
26. Hong, S., van den Goor, G., Brinkkemper, S.: A Formal Approach to the Comparison of Object-Oriented Analysis and Design Methodologies. In: Proceedings of the Twenty Sixth Annual Hawaii International Conference on Systems Sciences, Hawaii, pp. 689–698 (1993)
27. Brinkkemper, S.: Method engineering: engineering of information systems development methods and tolls. Information and Software Technology 38, 275–280 (1996)
28. Saeki, M.: Embedding metrics into information systems development methods: An application of method engineering technique. In: CAiSE 2003: The 15th Conference on Advanced Information Systems Engineering, pp. 374–389 (2003)
29. van de Weerd, I., Brinkkemper, S.: Meta-modeling for situational analysis and design methods. In: Syed, M.R., Syed, S.N. (eds.) Handbook of Research on Modern Systems Analysis and Design Technologies and Applications, pp. 28–38. Idea Group Publishing, Hershey (2008)
30. van der Spek, R., Hofer-Alfeis, J., Kingma, J.: The Knowledge Strategy Process. Springer, Heidelberg (2002)
31. Helms, R.W.: Redesigning Communities of Practice using Knowledge Network Analysis. In: Kazi, A.S., Wohlfart, L., Wolf, P. (eds.) Hands-On Knowledge Co-Creation and Sharing: Practical Methods and Techniques, Knowledgeboard (2007)
32. Jashapara, A.: Knowledge Management, An Integrated Approach. Pearson Education Limited, Harlow (2004)
33. Brinkkemper, S., Saeki, M., Harmsen, F.: Meta-Modelling Based Assembly Techniques for Situational Method Engineering. Information Systems 24(3), 209–228 (1999)

An Empirical Study on the Correlation between Knowledge Management Level and Efficiency in Ceramic Tile Industry

Gholamreza Khoshsima and Mehdi Ebrahiminejad

Department of Management, Faculty of Management,
Vali-e-Asr University of Rafsanjan, Rafsanjan, Kerman
{Khoshsima,Ebrahiminejad}@vru.ac.ir

Abstract. This paper investigated the correlation between the level of knowledge management (KM) and efficiency of ceramic tile companies in Iran. KM was measured with four components: infrastructure, process, people, and strategy and then efficiency was measured with data envelopment analysis (DEA). First, factor analysis was used in order to determine KM factors. Then, the partial correlation between KM components and efficiency was computed. Finally, multiple linear regressions were computed with the four independent variables (factors) and efficiency. Results showed positive relationship between the KM level and the efficiency in ceramic tile industry.

Keywords: KM, Efficiency, DEA, Correlation and Regression, Factor Analysis.

1 Introduction

The Iranian ceramic tile industry has been growing in both quality and quantity. It has had a very significant growth in quantity. Between 2000-2005, the production level has grown by 309%, and is predicted to reach 300 million square meters by 2006. But almost for sure, the progress has been far more promising in the quality front. The marked increase in exports and the comments encountered during international exhibitions all point to this fact. A similar growth is expected in quality of products in the next decade, and it will not be far from reach when Iran will become a major international producer and exporter of ceramic tiles. Iran has ranked 6th in the world, in ceramic tile production.

The Iranian ceramic tile and flooring sector is one of the most dynamic and innovative industries in Iran, being characterized by dynamic production growth and constant improvements in the quality and finish of its products. The continuous incorporation of new product designs and its high technological level have helped Iranian companies become a world leader in the design and production technology of ceramic tiles and flooring. The acquiring of a leading position is due above all, to the high degree of productivity and competitiveness. The ability to keep the quality and services at the global cutting edge of technology and the sector's potential export capacity have contributed to Iran's position as a global leader in the twenty-first century. Iran represents approximately 3.2% of total global production and 6.1% of

T. Yamaguchi (Ed.): PAKM 2008, LNAI 5345, pp. 160–172, 2008.

Asian production. The production totaled 220,000,000 m^2 in 2005, equivalent to an increase of nearly 78% over the previous year. This sector comprises approximately 71 companies in Iran that directly employ 60,000 people. One of the principal characteristics of the Iranian ceramic tiling sector is its high level of geographic concentration in Yazd province. Approximately 70% of the companies are located in Yazd, accounting for 45% of total Iranian production.

Table 1. Growth of ceramic tile in Iran (%)

Year	Growth of Production	Growth of Share of world	Growth of Number of Company	Growth of Number of Employee	Growth of Export
2000	16.39	1.33	10.52	9.67	-26.56
2001	9.85	1.41	9.52	29.41	14.89
2002	21.79	1.64	8.69	22.72	11.11
2003	15.78	1.78	12	48.15	143.33
2004	11.82	1.87	14.28	25	43.83
2005	78.86	3.23	10.93	2	19.04

In an economy where uncertainty is the only certain fact, one source of lasting competitive advantage is knowledge and its manipulation[24]. Most of the differences between the market and book values of companies can be explained through the level of knowledge that is not recognized in the companies' balance sheets. KM has been recognized as a highly effective means of improving competencies of a company, and it is expected to improve the Iranian ceramic tile industry's competencies. With increased levels of competition in the marketplace and high costs associated with human resources, increase in employee transience, and shortage of qualified knowledgeable workers, organizations have pursued the notion of making more effective use of the knowledge and expertise. This notion of managing knowledge as a corporate source has been looked to as one of the few foundational means that delivers sustainable core competencies in the future[27][28][25][31]. An operational objective of KM is to ensure that the right knowledge is available to the right processors, in the right representations and at the right time, for performing their knowledge activities (and to accomplish this for the right cost) [20]. Ways of measuring the success of knowledge management components is still being explored by organizations, researchers and management consultants. Most of the solutions offered are geared with profit-making commercial firms[12][42][43], embedding knowledge management within the overall business performance model of public sector organizations[14], the impact of knowledge management process perceived knowledge management effectiveness[39][3]. Both outputs and inputs were considered in this study, and performance was assessed based on single-output and multiple-input items, employing econometric analysis. Data envelopment analysis (DEA) was used to obtain a measure of relative manufacturing performance. Performance evaluation generated classes of firms based on performance. KM components and performance evaluation are reviewed in the next section. KM components for ceramic tile industry in the new competitive environment are also discussed and classified using factor analysis questions. A performance metrics is then identified, and the relationships between KM components and superior manufacturing

performance are hypothesized. Empirical analysis through DEA is used to obtain relative manufacturing performance evaluations. Statistical analysis is subsequently undertaken to test the hypothesized relationships between KM components and superior manufacturing. A consistency test of the empirical results based on different displays of outputs is presented next. Finally, the conclusion and a discussion of limitations and future research directions are presented.

2 Research Methodology

This research focuses on correlation between the four components of knowledge management (infrastructure, people, strategy, and processes) and organizational efficiency in Iranian ceramic tile industry. This study has also taken into account four statistical methods, namely factor analysis, reliability analysis, bivariate and partial correlation, and multiple linear regressions, and one operation research method for measuring efficiency, the data envelopment analysis (DEA). Factor analysis was conducted to explore factors underlying the measurement of knowledge management level. It also evaluated the construct validation of the instrument[26]. Reliability analysis, or Cronbach's alpha, was used to investigate the level of internal consistency of the scale items. A regression analysis was conducted to investigate the impact of knowledge management factors on overall efficiency. Partial correlation analysis was conducted to illustrate the correlation between an independent variable (when other independent variables are controlled) and a dependent variable. Bivariate correlation analysis was used in order to investigate the correlation between KM level and efficiency. Furthermore, the efficiency was measured using DEA. The definition of research variables: Sveiby (1997) defined knowledge management as "leveraging the intellectual assets of the company to meet defined business objectives"[10]. Some of authors defined knowledge as "Justified true belief" and as the set of justified beliefs that enhance an entity's capability for effective action[22][23][1][3][39]. Davenport and Prusak (1998) have defined organizational knowledge as ranging from "complex, accumulated expertise that resides in individuals and is partly or largely inexpressible" to "much more structured and explicit content". The types of organizational knowledge are reflected in several classification schemes. For example, Venzin et al (1998) identify a number of categories of knowledge-including tacit, embodied, encoded, embrained, embedded, event, and procedural. Kogut and Zander (1992) distinguish between "information" and "know-how" as two types of knowledge, viewing them as "what something means" and "knowing how to do something". Singley and Anderson (1989) also identify the parallel distinction between declarative knowledge (facts) and procedural knowledge (how to ride a bicycle). Another classification of knowledge views it as tacit or explicit as presented by Polanyi (1966). Explicit knowledge can be expressed in numbers and words and shared formally and systematically in the form of data, specifications, manuals, and the like. In contrast, tacit knowledge which includes insights, intuitions, and hunches is difficult to express and formalize, and therefore difficult to share [3]. Lee and Choi attempt to find relationships among knowledge management factors such as enablers, processes, and organizational performance. An integrative research model is built on a process-oriented perspective and then tested empirically[31]. Some specific work has been done in particular domain of evaluation

related to knowledge management. Bohn (1994) proposes a framework for levels of technological knowledge in his article. Moore (1999) developed a set of metrics for measuring and forecasting knowledge work. His set of measures was based on software companies, evaluating knowledge work with respect to the software characteristics. Hendriks et al (1999) have developed a framework in which companies can measure their current situation with respect to intellectual capacity and related management structure, in other words, measure how good their knowledge management is. Chandler (1999) proposed a six-step framework to align macro knowledge management (where how the business will achieve the knowledge management targets is determined at this level) to micro knowledge management (what to target in knowledge management activities according to the company's mission statement and other strategies). Del-Rey-Chamorro et al (2003) proposed a framework using BSC in which the way to develop performance indicators for knowledge management solutions is presented.

Many researchers have emphasized four factors for managing knowledge separately: Strategy, processes, infrastructure and people (Khoshsima et al, 2004). In this paper, knowledge management was clustered into four elements of the KM measurement model and then operationalized by constructing indices of extensiveness of (1) strategy, (2) people, (3) process and (4) infrastructure.

Strategy: Concepts of fit, integration, vision/strategy planning, and customer satisfaction have been discussed and investigated in different approaches. Indices of strategy are derived from a series of questions that relate to four items[18]: First, integration between business and knowledge management domains; is the link between business strategy and knowledge management strategy. Second, dealing with the corresponding internal domains, namely, the link between organizational infrastructure and process and knowledge management infrastructure and process. Third, the mission describes why an organization wants to be involved in certain activities, the vision makes explicit what it wants to be in the future, and the strategy explains how this should be accomplished. Without knowing why, what and how, it will be very difficult to link knowledge to the business objectives; in such a case, the first step in particular (knowledge identification) would have no basis and could not be completed[9]. The topics of this structural field describe aspects of the corporate vision and goal-setting with regard to knowledge management. The behavior of the top management and the budget policy are analyzed (Ehms and Langen, 2002). Fourth, determining customer satisfaction is fundamental to delivery of services. Successfully being able to judge customers satisfaction levels and to apply that knowledge potentially gives a hospitality manager an advantage over competitors via such benefits as product differentiation, increased customer retention, and positive word-of-mouth communication[45].

Process: Knowledge management processes (knowledge management activities) can be thought of a structured coordination of managing knowledge effectively[13]. Typically, these processes include activities such as identification, acquisition, development, sharing and distribution, utilization, and retention[37]. In this paper, we have applied Nonaka's processes concepts. Nonaka proposed four knowledge management processes: internalization, externalization, socialization, and combination. Socialization is the process of sharing tacit knowledge of individuals. Sharing experiences is a key to understand others' ways of thinking and feeling. In a

certain sense, tacit knowledge can only be shared if the self becomes part of a larger self. Externalization requires the articulation of tacit knowledge and its translation into forms that can be understood by others. Individuals transcend the inner- and outer-boundaries of the self in dialogue. Dialogue, 'listening and contributing to the benefit of all participants', strongly supports externalization. In practice, externalization is supported by the use of metaphors and analogies. Combination involves the conversion of explicit knowledge into more complex sets of explicit knowledge. To diffuse fragmentary knowledge, editing and systemizing such knowledge are the keys to this conversion mode. Here, new knowledge generated in the externalization stage transcends the group. Internalization means the conversion of newly created explicit into tacit knowledge of individuals. Learning by doing, training and exercises are important to embody explicit knowledge. Thus on the-job training, simulations or experiments are used to induce internalization of new knowledge[21]. People: Includes employees [20] and managers [2] who are divided into top level managers and middle level managers. An example of knowledge activities in organization is those carried out and supported by a human-processor (e.g. an individual knowledge worker, a group). An organization's knowledge workers use their knowledge handling skills, plus the knowledge at their disposal, in categorizing knowledge activities. Such activities can be examined at various levels of analysis and characterized in various ways. Managerial influences emanate from organizational participants responsible for administering the management of knowledge. The framework partitions these influences into four main factors: exhibiting leadership in the management of knowledge, coordinating the management of knowledge, controlling the management of knowledge, and measuring the management of knowledge. Human participants' personal beliefs and experiences can affect their approaches to sharing[20]. Infrastructure: Includes technology, structure and culture[13]. A particular instance of knowledge activities in an organization can be carried out by a computer-based processor (e.g. an intelligent agent)[20]. Technology advances can affect the modes and channels of sharing. It can create means to break knowledge-sharing barriers such as geographically dispersed locations[20]. Degree of support for collative work, for communication, for searching and accessing, for simulation and prediction, and for systematic storing[31]. Through the linkage of information and communication systems in an organization, previously fragmented flows of information and knowledge can be integrated. These linkages can also eliminate barriers to communication that naturally occur between different parts of the organization. Organizational structure is important in leveraging technological architecture. Although intended to rationalize individual functions or units within an organization, structural elements have often had the unintended consequence of habiting collaboration and sharing of knowledge across internal organizational boundaries [13]. This key area describes matters relating to the organizational structure and the assignment of knowledge management roles. Emphasis is given to aspects of procedural organization within the context of a process-based organization. The aim is to discover how knowledge management activities can be added to these specific business processes (Ehms and Langen, 2002). The most significant factor that

affects knowledge management is organizational culture. Shaping culture is central to the firm's ability to manage its knowledge more effectively[13].

Factor Analysis: We followed Nunnally's (1978) guidelines in developing items and pre-testing them for clarity and appropriateness. For studying knowledge management, we constructed a survey instrument of 36 questions. In general, multi-item scales were designed to capture the common construct. To assess the reliability of the measures, the most general method of using reliability coefficient such as Cronbach alpha coefficient was used[8]. For each set of items measuring a specific scale, Cronbach alpha was computed. A value of more than 0.70 was considered appropriate for the analysis [26]. It means that all the factors and the overall scale in this study are acceptably reliable. In order to reach the objective of the study, a factor analysis has been performed on the variables that, in accordance with the specialized literature, could have an impact on the level of knowledge management. The principle components analysis with orthogonal varimax rotation was used for factor extraction. The criteria used in this analysis are outlined as follows:

- Eigen values or latent roots of all components should not be less than 1.0;
- Communalities of all items should be more than 0.5;
- The loadings should be 0.50 or greater to be considered practically significant; and
- Cronbach's alpha values of each factor extracted and overall measure should be greater than 0.7[17].

Table 2 shows the loading and Cronbach alpha based on the revised sets of items. The items in the questionnaire covered by the four factors are shown in Table 1. We attempted to name the factors based on the principle of being concise without losing clarity of meaning; the names and content of the four factors are as follows: strategy consists of six items, infrastructure consists of twelve items, people contains 6 items, and process contains twelve items. Some marketing research experts recommend 0.30 (e.g. [16]). Other researchers[40] recommend retaining loadings above 0.50 for principal components analysis results and above 0.40 for exploratory factor analysis results. High item retention level standards increase the chances of committing type II errors of falsely rejecting nonzero population loadings[30]. This research is based on above 0.5. A Kaiser–Meyer–Olkin (KMO) measure of sampling adequacy assesses the interrelatedness of item correlations. A KMO score 0.60 is mediocre, 0.70 is middling, and 0.80 is meritorious [16]. Principal components analysis was performed with varimax rotation. The statistical program used in this study was SPSS factor analysis. Four factors were found and confirmed by means of a scree chart. Bartlett's Test of Sphericity was significant (1561.424; $p < .001$), suggesting that the correlation matrix was not an identity matrix. The Kaiser-Meyer-Olkin measure of sample adequacy reached a value of 0.606. Kaiser (1974) characterizes measures in the 0.60s as mediocre; therefore, we can comfortably proceed with the factor analysis. For our factor analysis (first questionnaire) 43 usable responses were received from 23 companies. Data gathered were analyzed by the SPSS 11.5 application.

Table 2. Construct validity and reliability analysis (Rotated Component Matrix)

Cronbach Alpha for Factor 1=0.9374; Factor 2=0.9356; Factor 3=0.9169; Factor 4=0.8839											
Question *i*	1	2	Question *i*	1	2	3	4	Question *i*	1	2	4
Qestion1	0.792		Qestion13			0.807		Qestion25	0.692		
Qestion2	0.823		Qestion14			0.872		Qestion26		0.791	
Qestion3	0.799		Qestion15			0.832		Qestion27	0.807		
Qestion4	0.719		Qestion16			0.802		Qestion28	0.8		
Qestion5	0.711		Qestion17			0.859		Qestion29	0.796		
Qestion6	0.805		Qestion18	0.742				Qestion30		0.764	
Qestion7		0.715	Qestion19		0.689			Qestion31		0.839	
Qestion8		0.828	Qestion20				0.794	Qestion32		0.778	
Qestion9		0.582	Qestion21	0.742				Qestion33			0.703
Qestion10		0.847	Qestion22		0.741			Qestion34			0.885
Qestion11		0.754	Qestion23				0.732	Qestion35			0.684
Qestion12		0.735	Qestion24			0.804		Qestion36			0.811

Extraction Method: Principal Component Analysis. Rotation Method: Varimax with Kaiser Normalization.

Performance measurement: DEA is a linear programming technique for the construction of a non-parametric, piecewise linear convex hull to the observed set of input and output data. DEA is used to compare a set of decision making units (DMU's) providing a similar set of services or products and using a similar set of resources to evaluate their relative efficiencies. DEA objectively identifies the most efficient best practice DMU's and the inefficient less productive DMU's. An inefficient DMU is one that produces its volume, mix, and quality of outputs by more resources based on comparison with the best practice DMU's. Charnes et al. (1978) introduced DEA techniques in a paper that used mathematical programming to pursue Farrell's approach to technical efficiency measurement. Their approach embodied an important improvement: instead of a single-output and single-input measure. They developed a method to deal with the case of multiple outputs and multiple inputs. This goal was achieved by computing a maximal performance measure for each production unit relative to all other units in the sample. For each inefficient production unit, a measure of relative inefficiency can be calculated by comparing its observed behavior (through the observed input and output vectors) either with the behavior of a reference unit located on the technological frontier or with a convex combination of different efficient units. Consequently, the potential improvement of the inefficient unit's productive performance is not necessarily obtained by reference to an actual efficient unit, but can be calculated by taking as a standard of comparison a virtual unit, made up of a weighted mix of the output and input profiles of several efficient production units. The weightings are determined as part of the solution of the mathematical optimization program[38]. Specifically, DEA determines the followings[40]:

- The best practice - most productive group of DMU's;
- The inefficient - less productive DMU's compared to the best practice DMU's;
- The amount of excess resources used by each of the inefficient DMU's;
- The amount of excess capacity or ability to increase outputs present in inefficient DMU's without utilizing added resources;
- The best practice DMU's that most clearly indicates that excess resources are being used by the inefficient DMU.

This information clearly and objectively indicates that units should be able to improve productivity and the amount of resource savings and/or output augmentation that the inefficient DMU's can potentially realize to meet the level of efficiency of the best practice units. DEA uniquely obtains these insights by comparing simultaneously the actual set of outputs achieved with the actual set of inputs used by a set of DMU's providing a similar set of outputs, (but naturally with varying volume and mix of inputs and outputs). It directly incorporates multiple inputs and outputs, which means that the results will be explicitly sensitive to the complexity and mix of outputs. It can incorporate outputs that have no clear price or market value, like training and new product development activities. Each output and input can be measured in its natural units without any need to convert them to common currency units. DEA employs mathematical programming to obtain ex post facto evaluation of the relative efficiency of management accomplishment[6]. The input-output *BCC* model evaluates the efficiency of DMU (0= 1, 2… n) by solving the following (envelopment form) linear program[7] and where θ_B is a scalar. The dual multiplier form of this linear program is expresses as BCC_0 model and where v and u are vectors and z and u_0 are scalars and the latter, being "free in sign," may be positive or negative (or zero). The equivalent *BCC* fractional program is obtained from the dual program. In fact, DEA seeks optimization contingent of each individual companies' performance in relation to the performance of all other units, those with the greatest productivity have a score of 1, suggesting 100% efficiency when compared to those in the competitive set. These optimal units, lie on a multidimensional frontier; the efficiency frontier envelops the inefficient units within and quantifies the inefficiency by a relative score of less than 100% and a relational measure on each output and input. Input and output variable are considered based on [34][35] which is characterized by the production of one desirable output, ceramic products (y) measured in million squares meters, and three variable inputs are also used: tons of clay, kaolin, feldspar and limestone (x_1), labor (x_2), measured as the number of workers, and capital (x_3), proxied by energy consumption in kilowatts-hour. Table 1 presents some descriptive statistics. As for application and model specification, we used DEA-Solver as this software appeared well suited to handle the computation-incentive linear programming models. Regarding model specification, we employed an input-oriented BCC model given the assumptions discussion earlier.

Table 3. Calculated efficiency with BCC input oriented(C=Company)

	Efficiency		Efficiency		Efficiency		Efficiency		Efficiency		Efficiency
C1	0.916	C7	0.671	C13	1	C19	0.951	C25	0.725	C31	0.998
C2	0.949	C8	1	C14	0.738	C20	0.849	C26	0.609		
C3	1	C9	1	C15	1	C21	0.741	C27	0.649		
C4	0.881	C10	0.728	C16	0.655	C22	0.979	C28	1		
C5	0.686	C11	0.731	C17	1	C23	0.799	C29	0.641		
C6	0.651	C12	1	C18	0.818	C24	0.797	C30	0.554		

Hypothesizes: Our hypotheses are largely derived from theoretical statements made in the literature on knowledge management.

H 1: Infrastructure has positive correlation with the level of efficiencies of the ceramic tile companies.
H 2: Processes have positive correlation with the level of efficiencies of the ceramic tile companies.
H 3: Strategy has positive correlation with the level of efficiencies of the ceramic tile companies.
H 4: People have positive correlation with the level of efficiencies of the ceramic tile companies.

3 Data Collection and Analysis

The data used to examine the research model and hypotheses, were collected through surveys. A sample of 31 firms in the ceramic tile industry was selected from Yazd province, which produces more than 78% of the total ceramic tiles in Iran. A package containing a cover letter and two unsealed envelopes was sent to each CEO of selected organizations. One envelope (KM questionnaire) was sent to each of the departments of marketing, information technology, and production. The DEA questionnaire was sent to the department of finance. Each envelope contained the questionnaire and a postage paid business return envelope. The cover letter to the CEO explained the purpose and importance of the study, and requested the CEO to forward the enclosed envelopes to the appropriate executives. The questionnaire design used a Likert seven-point scale to measure each company's attitude and comment on each question. The measurement scales ranged from "completely agree" to "completely disagree".

Correlation analysis: This study adopted partial correlation analysis to determine the correlation of each knowledge management component and its correlation with efficiency. As the KM variables are independent, they probably have impacts on each other; therefore, partial correlation was measured. The results show a strong positive correlation among the factors, as shown in Table 4.

Table 4. Partial and bivariate correlation between variables

Controlled Variables	Independent Variable		Efficiency	Controlled Variables	Independent Variable		Efficiency
People, Strategy and infrastructure	Process	Correlation Sig(2-tailed) df	0.6499 0.000 26	People, strategy and process	Infrastructure	Correlation Sig(2-tailed) df	0.4979 0.007 26
Strategy, process and infrastructure	People	Correlation Sig(2-tailed) df	0.8743 0.000 26	People, process and infrastructure	Strategy	Correlation Sig(2-tailed) df	0.8445 0.000 26

Multiple-regression analysis is a simple and extended application used mainly for understanding the linear relationship between a group of predicted variables and a valid variable. A multiple-regression analysis examined our hypotheses. The results of the multiple-regression analysis used in this article are shown in Table 5. Our model would not be meaningful if the correlation between KM components and efficiency was not significant. Therefore, the knowledge components were considered as an aggregated variable, and its partial correlation with efficiencies were computed. Then each hypothesis was tested to find the component which was important in the efficiency of companies. We examined the linearity, constant variance, and normality to meet the assumptions of regression analysis. Since the scatter plots of individual variables did not indicate any nonlinear relationships, therefore the linearity is guaranteed. Plotting the studentized residuals against the predicted value shows that no variable violates the constant variance. The result of the normal probability plot and Kolmogorov-Smirnov tests indicates no violation of normality (statistic=0.01). The overall regression model for finding the relationship between the KM components and efficiency is significant

Table 5. Summary of regression results

Model	Unstandardized Coefficients		Standardized Coefficients	T	Sig. 0.05 level	Collinearity Statistics	Adjusted R Square	F	Sig.
	B	Std. Error	Beta			VIF			
Constant	-0.366	0.108		-3.396	0.002		0.832	38.258	0.000
Process	0.0075	0.017	0.338	4.360	0.000	1.074			
People	0.153	0.017	0.706	9.185	0.000	1.058			
Infrastructure	0.056	0.019	0.228	2.928	0.007	1.088			
Strategy	0.127	0.016	0.622	8.042	0.000	1.072			

(F=38.258, p<0.000). Adjusted R square (0.832) suggests that 83.2% of the variance is explained by the four variables. The result of the collinearity test (Table 5) shows no multi-collinearity problem. The P-value of the F-test was less than 0.05, which was significant, making the four factors valid in predicting the relationship between KM components and company efficiencies. As a result, the four factors are valid and vital to the company efficiency in the ceramic tile industry. In addition, regression coefficients were used to predict the effect of independent variables on dependent variables using the T-test. The results showed that factor 2 (people) and factor 4 (strategy) had higher significant effects than the other factors on enhancing the company efficiency. Factor like infrastructure (including technology, culture, and structure) doesn't have higher impact on efficiency. This is mainly due to the fact that, almost all of the companies in Maybod use the same technology and the companies are operating within the same cultural milieu.

4 Conclusion

Knowledge management includes four main components, namely KM infrastructure, KM strategy, KM people, and KM processes. This paper discussed the effect of knowledge management components on efficiency in the Iranian ceramic tile industry.

As indicated by the results of the multiple-regression analysis, shown in Table 5, the B-value, Beta-value, T-value and all other values achieved the significant positive level. The Beta values for the model were 0.338, 0.706, 0.228 and 0.622, respectively. The model is

$$\hat{y}_i = 0.338 \times \mathrm{Pr}ocess + 0.706 \times People + 0.228 \times Infrastructure + 0.622 \times Strategy$$

all showing positive significant relationships. The adjusted R square is 0.832, indicating a very high explainability of the variables.

The results reveal a tremendous positive effect of knowledge management components on efficiency. The hypothetical assumption "*the higher the level of knowledge management components you have, the more efficiency you possess*" has been verified by means of the illustrative statistical analyses. Therefore, when enterprises decide to increase their efficiency, they first need to improve their knowledge management level. In order to have a more efficient performance and survive in the competitive environment, enterprises must emphasize the KM process, KM people, KM infrastructure, and KM strategy.

Acknowledgements. We thank Mr. Hossein Eslami Yazdi.

References

[1] Alavi, M., Leidner, D.: Review: Knowledge management and knowledge management systems: conceptual foundation and research issues. MIS Quarterly 25(1), 107–136 (2001)
[2] Anderen, A.: APQC. The knowledge management assessment tool: External benchmarking version (1996)
[3] Becerra-fernandez, I., Sabherwal, R.: Organizational knowledge management: A contingency perspective. Journal of Management Information Systems 18(1), 23–55 (2001)
[4] Bohn, R.E.: Measuring and managing technological knowledge. Sloan Management Review 36(1), 61–73 (1994)
[5] Chandler, M.,C.: Aligning micro to macro knowledge management, MSc thesis, Cran.eld University (1999)
[6] Charnes, A., Cooper, W.W., Rhodes, E.: Measuring the efficiency of decision making units. European Journal of Operational Research 2(6), 429–444 (1978)
[7] Cooper, W.W., Seiford, L.M., Tone, K.: Introduction to Data Envelopment Analysis and its uses with DEA-Solver software and references. Springer Science Business Media, Inc. (2006)
[8] Cronbach, L.J.: Coefficient alpha and the internal structure of tests. Psychometrika 16, 297–334 (1951)
[9] CWA 14924-1, European Guide to good Practice in Knowledge Management -Part 1: Knowledge Management Framework (March 2004)
[10] Del-Rey-Chamorro, F.F., Roy, R., Van Wegen, B., Steele, A.: A framework to create key performance indicators for knowledge management solutions. Journal of Knowledge Management 7(2), 46–62 (2003)
[11] Dovenport, T.H., Prusak, L.: Working knowledge: how organization manage what they know. Harvard Business School Press, Boston (1998)
[12] Edvinsonn, L., Malone, M.S.: Intellectual Capital, HaroerCollins, New York, NY (1997)
[13] Gold, A.H., Malhotra, A., Segars, A.H.: Knowledge management: An organizational capabilities perspective. Journal of Management Information Systems 18(1), 185–214 (2001)
[14] Gooijer, J.: Designing a knowledge management performance framework. Journal of Knowledge Management 4(4), 303–310 (2000)
[15] Kaiser, H.F.: An index of factorial simplicity. Psychometrica 39, 31–36 (1974)
[16] Hair Jr., J.F., Anderson, R.E., Tatham, R.L., Black, W.C.: Multivariate Data Analysis with Readings, 4th edn. Prentice Hall, Englewood Cliffs (1995)
[17] Hair Jr., J.H., Anderson, R.E., Tatham, R.L., Black, W.C.: Multivariate Data Analysis. Prentice-Hall, New Jersey (1998)
[18] Henderson, J.C., Venkatraman, N.: Strategic alignment: Leveraging information technology for transforming organizations. IBM Systems Journal 32(1), 472–484 (1999)
[19] Hendriks, B., Swaak, J., Lansink, A., Van AmIsfort, T., Heeren, E., Kalff, P.: The knowledge management measure of Telematica Instituut of The Netherlands. The Journal of the Cap Gemini Applied Knowledge Management Natural Work Team, 17–23 (May 1999)
[20] Holsapple, C.W., Joshi, K.D.: An investigation of factors that influence the management of knowledge in organizations. Journal of Strategic information systems 9, 235–261 (2000)
[21] Nonaka, I., Konno, N.: The concept of Ba: Building a foundation for knowledge creation. California Management Review 40(3), 40–54 (Spring, 1998)
[22] Nonaka, I.: A dynamic theory of organizational knowledge creation. Organization Science 5(1), 14–37 (1994)

[23] Nonaka, I., Toyama, R., Konno, N.: SECI, Ba and leadership: A unified model od dynamic knowledge creation. Long Range Planning 33, 5–34 (2000)
[24] Nonaka, I.: The knowledge-creating company. Harvard Business Review, 96–104 (November-December 1991)
[25] Jans, B.D., Prasarnphanich, P.: Understanding the antecedents of effective knowledge management: The importance of a knowledge-centered culture. Decision Sciences 34(2), 351–384 (2003)
[26] Kerlinger, F.N.: Foundations of Behavioral Research, Holt. Rinehart & Winston, New York (1973)
[27] Khoshsima, G.: A model for measuring organizational agility in Iran television manufacturing industry: a fuzzy logic framework. In: Proceedings of IEMC 2003, Albany, New York (2003)
[28] Khoshsima, G., Lucas, C., Mohaghar, A.: Assessing Knowledge Management with Fuzzy Logic. Lecture Note In Artificial Intelligence, 425–432 (2004)
[29] Kogut, B., Zander, U.: Knowledge of the firm, combinative capability, and the replication of technology. Organization Science 3(3), 383–397 (1992)
[30] Lawrence, F.R., Hancock, G.R.: Conditions affecting integrity of a factor solution under varying degrees of overextraction. Educational and Psychyological Measurement 59(4), 549–579 (1999)
[31] Lee, H., Choi, B.: Knowledge management enablers, processes, and organizational performance: An integrative view and empirical examination. JMIS 20(1), 179–228 (2003)
[32] Moore, C.R.: Performance measures for knowledge management. In: Liebowitz, J. (ed.) The Knowledge Management Handbook. CRC Press, Boca Raton (1999)
[33] Nunnally, J.C.: Psychometric Theory. MacGraw-Hill, New York (1978)
[34] Picazo-Tadeo, A.J., García-Reche, A.: Why is environmental performance different between firms? p. 23 (2006)
[35] Picazo-Tadeo, A.J., Reig-Martínez, E., Hernaández-Sancho, F.: Directional distance functions and environmental regulation. Resource and Energy Economics 27, 131–142 (2005)
[36] Polanyi, M.: The tacit dimension. Anchor Day, New York (1966)
[37] Protest, G., Raub, S., Romhardt, K.: Managing Knowledge: Building blocks for success. John Wiley & Sons Ltd., Chichester (2000)
[38] Reig-Martínez, E., Picazo-Tadeo, A.J.: Analysing farming systems with Data Envelopment Analysis: citrus farming in Spain. Agricultural Systems 82, 17–30 (2004)
[39] Sabherwal, R., Becerra-Fernandez, I.: An empirical study of the effect of knowledge management processes at individual, group, and organizational levels. Decision Sciences 34(2), 225–260 (2003)
[40] Sharma, S.: Applied Multivariate Techniques. John Wiley & Sons, Inc., Chichester (1996)
[41] Smith, H.A., McKeen, J.D.: Developing and aligning a km strategy, Working Paper, WP 03-03 (2003)
[42] Sveiby, K.: Skandia AFS business navigator (1997) (accessed 15 January 2001), http://www.sveiby.com.au/IntangAss/SkandiaAFS.html
[43] Sveiby, K.E.: The new organizational wealth: managing and measuring knowledge – based assets. Berret-Koehler, San Francisco (1997)
[44] Venzin, M., Krogh, G., Roos, J.: Future research into knowledge management. In: Krogh, G., Roos, J., Kleine, D. (eds.) Knowing in Firms. Sage Publications, London (1998)
[45] Atila, Y., Mike, R.: Customer-satisfaction measurement. Cornell Hotel and Restaurant Administration Quarterly 39(6), 60–70 (1998)
[46] Ehms, K., Langen, M.: Holistic Development of Knowledge Management with KMMM (February 2008), http://www.kmmm.org

Appendix 1: Factors and items of knowledge management

To what extent are the following items present and observable in your company:

Factors	Items	Studies
Process	Face-to-face meetings	[3] [39]
	Learning by observation	
	On the job training	
	Learning by doing	
	Modeling based on analogies and metaphors	
	Capture and transfer of experts' knowledge	
	Use of apprentices and mentor to transfer knowledge	
	Brainstorming retreats or camps	
	Employee rotation across areas	
	Groupware and other team collaboration tools	
	Repositories of information, best practices and lessons learned	
	Cooperative projects across directions	
People	People can understand not only their own tasks but also others' tasks	[31]
	Can make suggestion about others' tasks	
	Can not only communicate well with their department members but also with other department members	
	Are specialists in their own part	
	Can perform their own task effectively without regard to environmental change	
	Can collaborate with other persons inside the organization	[13]
	Can collaborate with other persons outside the organization	
	In multiple locations to learn as a group from a single source or at a single point in time	
	In multiple locations to learn as a group from multiple sources or at multiple points in time	
	Senior management clearly supports the role of knowledge in our firm's success	
	Managers frequently examine knowledge for errors/mistakes	
Infra structure	Structure promotes collective rather than individualistic behavior	[13]
	Structures facilitate the discovery of new knowledge	
	Structures facilitate the creation of new knowledge	
	High levels of participation are expected in capturing and transferring knowledge	
	Use of database, web and network	
	Problem solving system is based on technology like case-based reasoning	
Strategy	Overall organizational vision is clearly stated	[13]
	Knowledge is regarded as an input to business strategy development	[41]
	Knowledge is regarded as a support for business strategy	
	KM strategy and business strategy are aligned	[18]
	KM infrastructure and process are linked	
	Customer satisfaction level is measured for applying knowledge	[45]

Web-Based Knowledge Database Construction Method for Supporting Design

Kiyotaka Takahashi[1], Aki Sugiyama[1], Yoshiki Shimomura[1], Takeshi Tateyama[1], Ryosuke Chiba[1], Masaharu Yoshioka[2], and Hideaki Takeda[3]

[1] Faculty of System Design, Graduate School of System Design, Tokyo Metropolitan University, 6-6 Asahigaoka, Hino-shi, Tokyo 191-0065, Japan
takahashi-kiyotaka@sd.tmu.ac.jp
[2] Graduate School of Information Science and Technology, Hokkaido University Kita 14 Nishi 9, Kita-ku, Sapporo, Hokkaido, 060-0814, Japan
yoshioka@ist.hokudai.ac.jp
[3] National Institute of Informatics, 2-1-2 Hitotsubashi, Chiyoda-ku, Tokyo 101-8430, Japan
takeda@nii.ac.jp

Abstract. In recent years, comprehensive problem solving by artefact designers has been required as demands related to artefacts become greater and more complicated. In relation to this background, we have proposed Universal Abduction Studio (UAS), a computer environment that synthetically supports a creative design. However, in general, it is difficult for a designer to acquire multiple domain knowledge because this knowledge is accumulated manually. This paper proposes a Web-based knowledge database construction method for supporting design using UAS.

Keywords: Creative Design, Knowledge Extraction, Analogy, Ontology.

1 Introduction

In recent years, the number of problems related to artefact design has increased for many reasons: the problem of manufacturers' responsibility, worsening environmental problems, the magnification and complication of artefacts and the diversification of consumers' values. Because of these issues, designers' roles and responsibilities have increased. In order to overcome these problems, the authors have developed a computer environment called *Universal Abduction Studio (UAS)* [13] to support creative design. The authors account for new knowledge generation processes in creative design; UAS realizes step-by-step knowledge extension by integrating different domains' knowledge groups.

In the above-mentioned knowledge integration, referring to a large amount of knowledge is effective. Therefore, designers need to store a huge amount of knowledge groups in multiple domains in the UAS design knowledge database. However, in the current UAS system, designers need to extend the knowledge stored in UAS by hand. Therefore, it is hard for them to acquire knowledge groups from multiple different domains. Therefore, the current prototype of UAS has a problem in

T. Yamaguchi (Ed.): PAKM 2008, LNAI 5345, pp. 173–184, 2008.

that designers only store a limited amount of a domain's knowledge in the design knowledge database. In addition, though we previously proposed a method to convert design knowledge data described in natural language into data with a format that could be processed in UAS [17], its processing accuracy is not so high. Therefore, designers currently need to correct design knowledge data by hand before conversion.

In this paper, for automatic and efficient knowledge construction, we propose a Web-based database construction method. For this purpose, we first propose a method of acquiring knowledge on the Web; this identifies design knowledge automatically, and extracts sentences that contain useful design knowledge information. Second, we propose a method to delete information that is useless for designers from the design knowledge information database. Third, we propose a method to change the acquired design knowledge information into available knowledge expression data for UAS.

We introduce the concept of UAS and the existing knowledge database construction method for UAS in section 2. In section 3, we propose the Web-based design knowledge construction method. In section 4, we verify the effectiveness of the proposed method. In section 5, we discuss the results of this verification. In the final section, we conclude the paper.

2 Universal Abduction Studio

2.1 Concept of UAS

UAS is a design support system based on abduction. Abduction, as proposed by C.S. Peirce, is a logical process by which an axiom can be found from a theorem [9]. Fig. 1 shows the system architecture of UAS. A workspace, a knowledge base group and a knowledge integration module group compose UAS. Designers put the design problem for operating and integrating with domain knowledge on the workspace. The knowledge integration module consists of multiple abductive reasoning mechanisms,

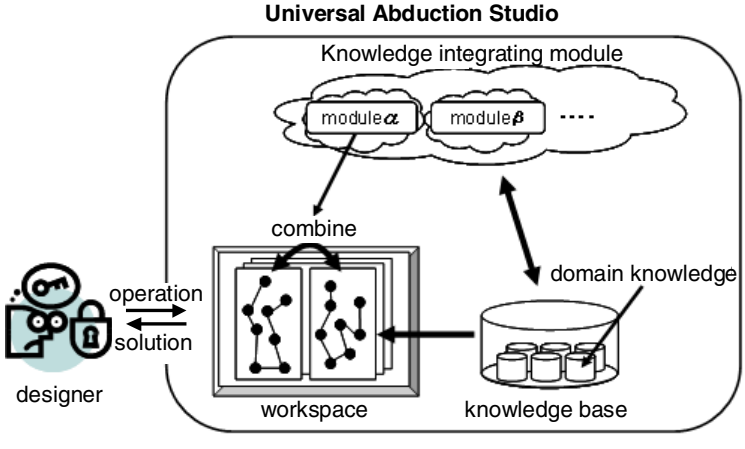


Fig. 1. System architecture of UAS [13]

and designers choose one or some of them depending on each design problem. The knowledge base consists of multiple domain knowledge bases that can potentially be integrated with the first domain‘s knowledge. The abductive reasoning system then performs knowledge integration. This fundamental concept requires a unified knowledge description among various domain knowledge bases.

2.2 Existing Knowledge Database Extension Method for UAS

In this section, we describe the existing knowledge database construction method for UAS.

1. Designers obtain design knowledge from paper texts that include the design knowledge and input design knowledge sentences to a text file line by line on a computer.
2. The text files are converted into a format that is processible in UAS.

In UAS, rule-type knowledge expressed in an If-Then form is used as a knowledge representation form (hereafter, we abbreviate this as If-Then-type design knowledge). We proposed an automatic document tagging method [17]. In this method, a computer automatically converts the design knowledge described by natural language into graph geometry as UAS knowledge.

The automatic document tagging system generates the design knowledge for UAS using the following procedures. The system:

1. Analyses a modification structure for the knowledge described by natural language using *Cabocha* [3], which is a parsing machine.
2. Defines a uniting relation between a predicate and an arbitrary case using the concept of *surface case* [14] with *GoiTaikei --- A Japanese Lexicon* [1].
3. Distinguishes a delimitation of the If-Part and Then-Part automatically, and divides them.
4. Outputs the result of procedure 3 (If-Then-type design knowledge) as an XML file.

The automatic document tagging system delimits as If-Parts and Then-Parts at the part including 10 patterns of words and Parts-of-Speech extracted from design documents [15] exist.

3 Proposed Method

3.1 Procedure of the Proposed Method

The information on the Web about multiple domain knowledge is updated hourly. Moreover, we can freely process that Web information using a computer, so we can extract information automatically. Therefore, the authors use Web information as the source in order to acquire multiple domain knowledge groups automatically and efficiently. Moreover, it is preferable that information stored in the design knowledge database is “New information that is highly reliable and useful for design objects”. In this paper, the authors use press releases on the Web as the source for design knowledge, because press releases include design knowledge information such as the demand that a new product targets, the mechanism for achieving the function, and the

function to satisfy the demand. Fig. 2 shows the construction method for the UAS design knowledge database. The details of the procedure are described as follows.

1. Acquire Web information including design knowledge using *Webstemmer* [7], a Web crawler for news sites, and store it in the Web information database.
2. Acquire design knowledge information from the Web information database using the method proposed in this paper.
3. Store the acquired design knowledge information in the design knowledge information database for UAS.
4. Convert the acquired design knowledge information into a text file with one sentence on each line using the proposed conversion system.
5. Convert the design knowledge information given as one sentence in each line into design knowledge for UAS using the automatic document tagging system, and store it in the design knowledge database for UAS.

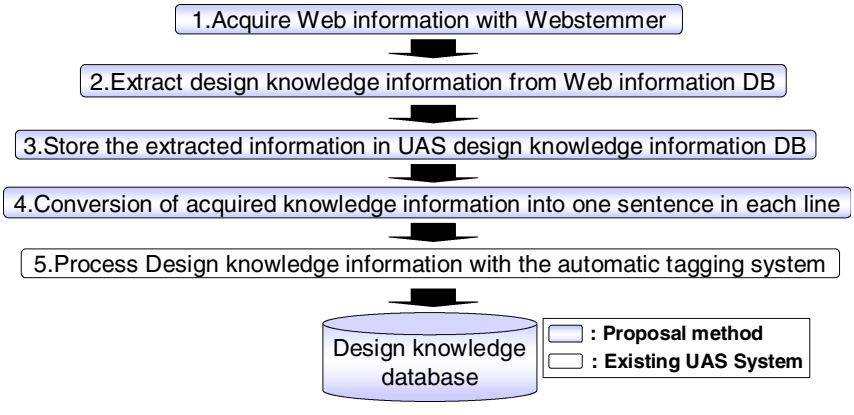


Fig. 2. Construction method for the UAS design knowledge database

3.2 Extracting Method of Design Knowledge Information from Web Information DB

In this section, the authors define the term design tag, introduced in this paper, explain how to create a design tag, and how to apply the design tag automatically to Web information. First, the authors define *design tag* as a tag for the efficient acquisition of useful design knowledge information from Web information. In this research, "The broader concept of the feature word of an article," is used as design tags, such as a physical phenomenon, an artificial material in GoiTaikei. The authors define a *feature word* is "A word existing with high frequency". In this research, a feature word is chosen by applying the *TF (Term Frequency) method* [6] to nouns, verbs, adjectives and adjectival verbs. We assume that a sentence that includes some design tags is a sentence that has useful design knowledge information. Therefore, we extract paragraphs that include design knowledge information sentences with design tags added. In addition,

Web information includes information that is irrelevant to design such as sales information, etc. Therefore, we also use design tags as labels to distinguish relevant information. Hence, we create a design tag set by analysing press releases.

Next, we explain the *method for creating the design tag set*. First, we acquire multiple pieces of information from the Web and extract candidates for the design tag set from Internet. In this paper, we use press releases published in *Nikkei Net* [8] as an example of press releases about products. For example, we acquire articles on the four different domains of beverages, digital cameras, electric power equipments and Web services from Nikkei Net, one by one, to avoid bias in the knowledge domains. The design tag set is created using the following procedure. Fig. 3 shows an example of creating a design tag.

1. Execute a *morphological analysis* of sentences in the press releases and a *frequency analysis* for each morpheme.
2. Extract the concepts of the extracted feature words in procedure 1 based on GoiTaikei.
3. Abstract the extracted broader concept in procedure 2 to make it "a concept which does not depend on knowledge domains," and define the highest concepts as design tags.

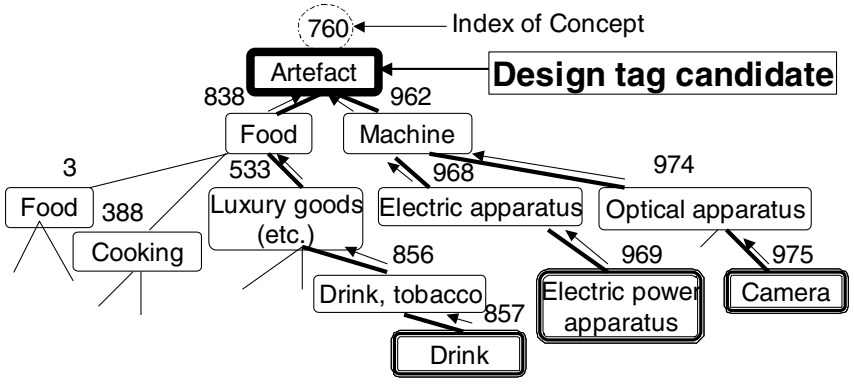


Fig. 3. Concrete example of the method for creating a design tag set

Procedure 3 makes it possible to acquire useful information for the design regardless of the domain of the press release. In addition, we applied procedure 3 to the concepts of all the listed feature words in procedure 1 and created multiple design tags.

For the frequency analysis for feature word extraction, we use a morpheme frequency analysis program produced by the authors. We adopted the name of a concept in GoiTaikei as the name for a design tag and we added an initial D, meaning "Design," to their heads. If a term had multiple broader concepts, we allocated the appropriate broader concept to the term by considering the contextual meaning of the term subjectively. Table 1 shows the design tag set created in this research.

Table 1. Design tag set created in this research

Japanese concept	Design tag
Organization	<D-ORGANIZATION>
Artefact	<D-ARTIFACT>
Physical phenomenon	<D-PHENOMENA>
Property	<D-PROPERTY>
States	<D-STATES>
Right	<D-RIGHT>
Aspect (etc.)	<D-ASPECT>
Possible	<D-CAN>

Next, we explain the *automatic tagging method for design tags*. First, designers execute a morphological analysis by using *Chasen* [16] for each article acquired by Webstemmer. Second, they match each morpheme in the article and each term that should have a design tag added among the terms in GoiTaikei and add design tags to the morphemes matched to them.

In this thesis, objects that should be tagged are all terms that belong to the subordinate concept of each tag. Because prototypes and inflections exist in Japanese verbs, we perform the matching by using prototype information for vocabularies outputted as morphological analysis results by Chasen. Finally, all the paragraphs, including the sentence tagged by the above-mentioned procedure, are extracted. Each task in the automatic tagging method for design tags is completed automatically by a computer (processing with the grep command of Linux or an existing substitute software program).

3.3 Filtering Method

In the information included in press releases related to new products' release information, most of the sales information and corporate information is less likely to give designers new insights for design. Therefore, if we process the press releases that include this information using the proposed method, a large amount of useless information will be included in the design knowledge database.

As a result, computational complexity and processing time for the inference of UAS will increase pointlessly. Consequently, we propose the following *filtering method* to reduce information that is useless for design in the design knowledge information database. We assume that the following six *named entity tags* defined by *IREX* [12], <ORGANIZATION>, <PERSON>, <LOCATION>, <DATE>, <TIME>, and <MONEY> are added to sales information that seems not to be useful for the design (Refer to Table 2). First of all, designers create articles that include the named entity tag IREX by morphological analysis with Chasen. Next, they reduce the paragraphs including those tags. These tags are called "*Reduce tag*". This procedure is implemented using Cabocha because Chasen and the named entity analysis program are built into Cabocha.

Table 2. The named entity tag set [12]

Object	Start position tag	Finish position tag
Organization-name, Government organization-name	<ORGANIZATION>	</ORGANIZATION>
Person-name	<PERSON>	</PERSON>
Location-name	<LOCATION>	</LOCATION>
Artefact-name	<ARTIFACT>	</ARTIFACT>
Date-phrase	<DATE>	</DATE>
Time-phrase	<TIME>	</TIME>
Money-phrase	<MONEY>	</MONEY>
Percent-phrase	<PERCENT>	</PERCENT>

3.4 Automatic Conversion Method for Knowledge Representation

The following two kinds of problems exist in the automatic document tagging system.

- Erroneous handling of sentences that include two or more design knowledge domains
- Low processing accuracy. In particular, this system cannot correctly process a lot of If-Then type design knowledge data

The former is a problem such that sentences related to two or more design knowledge are not processed correctly. This is because we have developed the automatic document tagging system based on the assumption that designers usually describe only one piece of design knowledge in each line of the input text file. Therefore, when the system processes design knowledge information acquired through the above-mentioned proposed method, the possibility of generating a large amount of incorrect design knowledge is high. As a result, an increase of useless information in the design knowledge database for UAS and computational complexity in the inference of UAS will arise. Therefore, we need to convert design knowledge information into a text file in which one piece of design knowledge is described in each line. The latter is a problem given the low accuracy of the current automatic document tagging system. According to our experiments, when the automatic document tagging system processed 82 kinds of design knowledge information sentences, it could convert less than 10% of the design knowledge information into design knowledge for UAS correctly. Therefore, it is necessary to improve the accuracy of the automatic document tagging system.

Therefore, we propose following two methods in this research as *automatic conversion methods for knowledge representation*.

- A converting method to include one design knowledge sentence in each line
- An improved method of the automatic document tagging system

The Converting Method to Include One Design Knowledge Sentence in Each Line. The automatic tagging method for design tags acquires paragraph including design knowledge information. However, we think it is often case that the causal relationship between the If-Part and Then-Part is described in one sentence. Therefore, we think that

it is appropriate to divide design knowledge information into one sentence for each line. Hence, in this research, we propose a system that can place a period "." to punctuate the sentence and divide design knowledge information into multiple sentences automatically with the regular expression. This system is implemented in Ruby.

The Improved Method of the Automatic Document Tagging System. It is necessary to clarify the method of distinguishing between the If-Part and Then-Part in order to improve the accuracy of delimiting the If-Part and Then-Part. We describe related work on this measure as follows. Inui et al. analyzed newspapers, and categorised words that showed a causal relation as *causal markers*. In addition, they proposed a method for acquiring cause and effect knowledge including "*tame* (one of the *conjunctive particles* indicating causality)"[2]. On the other hand, Sakaji [10] researched a method of acquiring knowledge about causal relations using causal markers for all cause and effect knowledge. Based on the above-mentioned related works, we propose effective causal markers as the measure for distinguishing between the If-Part and Then-Part based on the result of analysing design knowledge information in press releases. To be specific, we acquired design knowledge from 50 press releases, converted them into the If-Then forms by hand. Herewith, we enumerated the description seen in the If-Then delimitation. As a result, we identified the following two heuristics for the If-Then delimitation.

- Seen in [2][10], "conjunctive particle" indicates causality often exist.
- "*infinitive form of the verb*" indicates causality with comma often exist.

Based on the above heuristics, we use the causal markers in Table 3. The method of distinguishing a concrete delimitation between the If-Part and Then-Part is described as follows. This system is implemented in Ruby.

1. Match the causal markers and each morpheme using the morphological analysis machine Mecab [4], working in the automatic document tagging system. The output data describe the morpheme and Part-of-Speech information.
2. If the causal markers correspond to an arbitrary morpheme, the system divides the sentence into the If-Part and Then-Part before and behind the morpheme.

Table 3. Causal markers in this paper

	Term "The romanized Japanese word":(the anglicized Japanese word)	Kind of term
	"yori, ":(by,) "de, ":(by,) "niyoru":(by) "niyotte":(by) "noga":(by) "ba,":(if…,) "baai":(when) "node":(because) "houga":(with) "kotoha":(that is)	Word
Item 1	"toki":(when) "tame":(because)	Reading
Item 2	"soubi si,":(is installed, so) "jitugen si,":(is achieved, so), etc.	Infinitive form of verb

4 Verification

4.1 Verification of Automatic Tagging Method of Design Tags and Filtering Method

In our experiment, we verify the proposed method by applying it to 20 press release articles acquired by Webstemmer. First, we mark the paragraphs including useful

sentences for design in each acquired press release. In this experiment, we use press releases for a time clock, a wool coat, a game controller, etc. We assume these marked paragraphs to be the correct answers. We apply the proposed method to the articles and examine whether the method can output the correct answers.

The method is evaluated according to *F-measure* [11], which gives a value in assessing both *accuracy* and *recall ratio*. The accuracy evaluates how much unnecessary information a system can remove. The recall ratio evaluates how much necessary information a system can acquire without omitting necessary information. F-measure is a value that shows the effectiveness of an information extraction system. The value ranges from 0 to 1. We aim to maximize the F-measure in this paper. A computer executes the processes (tagging and filtering) automatically.

Next, we give the experimental conditions and results, as well as our discussion. In this research, we assume that several kinds of design tags are put to use in finding information for design. Because of this assumption, if the number of kinds of design tags in a sentence is greater than a threshold value, we regard the paragraph which includes such a sentence as useful information for design. We define the threshold value as the *design tag acquisition threshold*. We classify the *design tag acquisition threshold* into four stages, as shown in experiment 1-8 in Table 4. Moreover, we verify the automatic tagging method for design tags with and without the filtering method.

Table 4. Results of the verification of the automatic tagging method for design tags and the filtering system

	Design tag's acquisition threshold	Filtering	Average accuracy	Average recall ratio	F-measure
Experiment 1	One kind of design tag	Non-executed	9.19%	97.14%	0.17
Experiment 2		Executed	9.25%	60.00%	0.16
Experiment 3	Two kinds of design tag	Non-executed	14.95%	91.43%	0.26
Experiment 4		Executed	19.00%	54.29%	0.28
Experiment 5	Three kinds of design tag	Non-executed	22.83%	82.86%	0.36
Experiment 6		Executed	33.33%	48.57%	0.40
Experiment 7	Four kinds of design tag	Non-executed	27.06%	65.71%	0.38
Experiment 8		Executed	39.39%	37.14%	0.38
Further experiment Experiment 9	Three kinds of design tag	Executed to only the sentences put three kinds of design tag	27.72%	80.00%	0.41

When we compare the F-measure of each experiment without the filtering method, we confirm that the F-measure improves proportionally as the number of kinds of design tag increases. Therefore, the automatic tagging method for design tags can result in the efficient acquisition of useful information for designs by setting the design tag's threshold to be high. Therefore, we confirm that the method tends to tag multiple design tags as representing useful information for design. The experimental results show the effectiveness of the information extraction method based on the linguistic features of design knowledge information. As a result of applying filtering system to automatically tagged sentences, the accuracy and the F-measure improve the design tag's threshold in most cases. We can confirm that applying the filtering

system can reduce the noise in the Web information database, and the effectiveness of the proposed method has been demonstrated.

In addition, we carry out experiment 9 in order to examine about adding the condition to the filtering system based on the result of automatic tagging system of design tags. In experiment 9, design tag's acquisition is three kinds of design tag, and we add condition that designers apply the filtering system to only the sentences put three kinds of design tag are applied. As a result, in comparison with experiment 6, the accuracy improves without decreasing the recall ratio. Therefore, we consider that we can construct a more effective system by combining the result of the automatic tagging method for design tags and the filtering system.

4.2 Verification of the Automatic Conversion Method for Knowledge Representation

First, we verified the effectiveness of the converting method for one design knowledge sentence in each line. When we applied this proposed method to 100 press releases, it was automatically able to convert the design knowledge from all the press releases to one sentence in each line using punctuation. As mentioned above, the effectiveness of the converting method for placing one design knowledge sentence on each line is shown.

Next, we verify the effectiveness of the improved method of the automatic document tagging system. As a validation methodology, we obtain the correct data for the delimitation of If-Parts and Then-Parts for 32 kinds of design knowledge information that describes customer demands, the functions to satisfy the demand and the mechanism to achieve the function by hand. As a result, we verified whether the improved automatic document tagging method could correctly process If-Then-type knowledge information. We analysed 82 kinds of design knowledge information extracted from 17 kinds of press releases. This system extracted 32 kinds of If-Then-type knowledge information.

The part of the press release regarding a motherboard is shown below as an example of an experimental result.

<If-Part> By adopting low generation of heat and low RDS (on) MOSFET
<Then-Part> Longevity of the capacitor has improved by six times or more. [5]

The verification result for the improved automatic document tagging system is shown in Table 5. As shown, 65.60 % of the If-Then-type knowledge information can be converted to If-Then-type design knowledge. The improved automatic document tagging method can correctly convert more than half of 32 kinds of If-Then-type design knowledge information.

Table 5. Result of the verification of the improved method of the automatic document tagging systems

	Quantity	Proposal method	
		Accuracy before it applies	Accuracy after it applies
If-Then type knowledge	32	6.30%	65.60%

As mentioned above, the effectiveness of the improved method of the automatic document tagging system is shown.

Finally, when we apply the design knowledge acquired by the proposal method to inference mechanism of UAS, it operates correctly.

5 Discussion

We confirmed the feasibility of performing a series of operations for constructing the design knowledge database for UAS automatically and efficiently by the proposed methods. The following are the main results:

- *Automatic tagging method of design tags:* This method could extract design knowledge information automatically on the basis of linguistic features.
- *Filtering method:* This method could reduce the amount of redundant information such as sales information (sale dates, offer prices, etc.) using the named entity analysis automatically. In addition, it could improve the effectiveness of the automatic tagging method of design tags.
- *Automatic conversion method for knowledge representation*: This method could automatically convert more than half of the extracted If-Then-type information into If-Then-type design knowledge for UAS.

We found that the proposed methods have the following drawbacks:

- Because of errors commonly occurring in the named entity analysis, the filtering method sometimes erroneously reduces the amount of design knowledge information by including reduce tags.
- When If -Then-type information includes active sentences and itemized sentences, the improved automatic document tagging method can not process it. This is because such information does not include causal markers.

We will solve the above problems in the future.

6 Conclusion

In this thesis, we proposed a Web-based knowledge database construction method for supporting design. The verification results demonstrated the possibility of constructing a design knowledge database for UAS automatically and efficiently. We will improve the database construction method on the basis of the results obtained in this study. In addition, we will develop the system including all proposed method in order to automate the manual processing part, and verify about it intimately.

References

1. Ikehara, S., Miyazaki, M., Yokoo, A., Shirai, S., Nakaiwa, H., Ogura, K., Ooyama, Y., Hayashi, Y.: Nihongo Goi Taikei—A Japanese Lexicon. Iwanami Shoten 5 (1997) (in Japanese)
2. Inui, T., Okumura, M.: Investigating the Characteristics of Causal Relations in Japanese Text. In: The 43rd Annual Meeting of the Association for Computational Linguistics, Workshop on Frontiers in Corpus Annotation II: Pie in the Sky (2005)

3. Kudo, T., Matsumoto, Y.: Japanese Dependency Analysis Using Cascaded Chunking. In: Proc. 6th Conference on Natural Language Learning (CoNLL 2002), pp. 63–69 (2002)
4. Kudo, T.: Mecab: Yet Another Part-of-speech and Morphological Analyser, http://mecab.sourceforge.net/
5. Links International, Inc.: GA-MA790X-DS4 GIGABYTE Motherboard—Product Information (This press release is available in NIKKEI NET [8] dated (November 21, 2007) (in Japanese), http://www.gigabyte.com.tw/Products/Motherboard/Products_Overview.aspx?ProductID=2695
6. Luhn, H.P.: The Automatic Creation of Literature Abstracts. Journal of Research and Development 2(2), 159–165 (1958)
7. Niiyama, Y.: Webstemmer, http://www.unixuser.org/~euske/python/Webstemmer/index-j.htm
8. Nikkei Inc.: NIKKEI NET, http://release.nikkei.co.jp
9. Peirce, C.: Collected Papers of Charles Sanders Peirce, vol. 5. Harvard University Press, Cambridge (1935)
10. Sakaji, Y., Takeuchi, K., Sekine, S., Masuyama, S.: Causative Relation Extraction Using Syntactic Pattern. In: 14th Annual Meeting of the Japanese Association for Natural Language Processing, Tokyo, Japan (2008)
11. Sarwar, B., Karypis, G., Konstan, J., Riedl, J.: Analysis of Recommendation Algorithms for E-Commerce. In: Proc. of the 2nd ACM Conference on Electronic Commerce (EC 2000), pp. 285–295 (2000)
12. Sekine, S., Eriguchi, Y.: Japanese Named Entity Extraction Evaluation—Analysis of Results. In: Proc. of The 18th International Conference on Computational Linguistics, Saarbrucken, Germany, July 31–August 4, pp. 1106–1110 (2000)
13. Takeda, H., Sakai, H., Nomaguchi, Y., Yoshioka, M., Shimomura, Y., Tomiyama, T.: Universal Abduction Studio—Proposal of a Design Support Environment for Creative Thinking in Design. In: The Fourteenth International Conference on Engineering Design, ICED 2003 (2003)
14. The National Institute for Japanese Language: Relation between Surface Case and Deep Case in Japanese, Sanseido (1997) (in Japanese)
15. Watanabe, H.: Zoku Mechanical Design Acquaintanceship Note, The Nikkan Kogyo Shimbun (1988) (in Japanese)
16. Yamashita, T., Matsumoto, Y.: Language Independent Morphological Analysis. In: Proceedings of 6th Applied Natural Language Processing Conference, pp. 232–238 (2000)
17. Yoshioka, M., Satoh, T., Morimoto, K., Takeda, H., Shimomura, Y.: Proposal of Automatic Document Tagging Method for Hypothetical Knowledge Generation. In: Creative Design Support System, 20th Annual Conference of the Japanese Society for Artificial Intelligence, 3B3-02, pp. 1–4 (2006)

Classifying Digital Resources in a Practical and Coherent Way with Easy-to-Get Features*

Chong Chen[1,2], Hongfei Yan[1], and Xiaoming Li[1]

[1] Computer Networks and Distributed Systems Laboratory, School of EECS, Peking University, 100871, Beijing, China
[2] Department of Information Management, School of Management, Beijing Normal University, 100875, Beijing, China
{cc,yhf}@net.pku.edu.cn, lxm@pku.edu.cn

Abstract. With a rich variety of forms and types, digital resources are complex data objects. They grows fast in volume on the Web, but hard to be classified efficiently. The paper presents a practical classification solution using features from file names and extensions of digital resources. The features are easy to get and common to all resource. But they are generally low frequency and sparse, which implies that statistical approach may not work well. Our solution combines Naive Bayes (NB) classifier with Simple Good-Turing (SGT) probability estimation, which shows great promise for this condition with a total accuracy of 80%. In our opinion, the results are due to 1) the features fit the NB's conditional independence hypothesis well; 2) the abound one-time-occurrence features lead to reasonable probability estimation on unobserved features, which also means general feature selection strategy is not needed in this case. A 7.4TB digital resource collection, CDAL, is used to train and evaluate the model.

Keywords: Digital resource, classification, feature, probability estimation.

1 Introduction

In this paper, a digital resource refers to the data object composed of one or more files with diverse data formats, usually having its own directory organization structures, and representing a certain thing, entity or topic. Digital resources are popular and widely shared on the Web. [1] found that the "resource-seeking" goals account for a large part of web searches. CNNIC also announced that digital resources have enjoyed the greatest requirement by Chinese Web users from 2007[2]. Lots of the resources are public and valuable mental treasures. In many cases, such as building e-learning repositories, digital libraries or other special collections, there is a need to efficiently reorganize the resources collected from the Web in uniform classification architecture for better access. Traditionally, it is done manually, because the digital resource is 1) hard to quantitatively define the composition, structure and size; 2)

*The research is supported by PRC MOST Grant 2006BAH02A10, 863 Grant 2006AA01Z196 and 863 Grant 2007AA01Z154.

T. Yamaguchi (Ed.): PAKM 2008, LNAI 5345, pp. 185–196, 2008.

composed of members of various data types, such as text, video, etc., other than particular one; 3) disorder as a kind of Web-based data usually with chaotic naming and diverse organization. Obviously, manpower-based classification cannot meet the need of huge resource volume. People need solutions to automatically process the hodgepodge data objects in a coherent way no matter what members, structures, data types or sizes they are.

The paper first studies on the resources attributes and finds out suitable features, then applies the classification scheme which can make good use of these features. If we can deal with kinds of resources with an acceptable accuracy, say 80%, it is no doubt labor-saving and time-efficient in speed-up the initial stage of building large digital resource repositories. That is, people only need to refine the results based on the primary automatic category result. It is the realistic motivation for our work.

In what follows, we list related work in section 2. Section 3 shows several primary concepts and definitions. Section 4 discusses on how to select features. Section 5 states the classification approach and the evaluation measures. Section 6 presents the experiment and demonstrates the suitable classification solution; we also analyze the effects on precision in this part. Section 7 is the conclusion and possible future works.

2 Related Works

The Internet Archive [3] has reserved a large amount of digital resources. Their huge amount of resources is either uploaded by the public through the Internet or collected by their staff. In both cases, the process of labeling suitable categories to the resources is labor-cost. The same problem came up when we built China Digital Asset Library (CDAL) [4] by collecting digital resources from public ftp sites in China.

In [5], we proposed a tree-merge model to illustrate how the time saved by judging a batch of resources to be the same category at a time. This approach is based on the assumption that resources of same category are likely to be put in same parent directory. By grasping the organization knowledge embedded in the original directory structures, we could get an acceptable accuracy with reasonable initial cost. It works well when the original organization quality is good, and the meaning of the host directory's name matches some category in target taxonomy. But things are not always like that — original Web resource collections are full of personality. Thus the way of using initial organization knowledge of the resources is not suit to automatic classification methods. We consider the attributes inside of the resources, such as their member names, member file extensions, or even file sizes.

[6] flattened the tree-structure product catalog and modeled a category vector with the product items' features of this category. The features aggregate upward from low level is similar to our treatment on digital resource. In our case, each digital resource is taken as a tree, and is represented by the features extracted from its members. That is, we represent each digital resource without considering its inner hierarchy.

[7] studied the characters of the file size and file name length of files shared in public ftp sites. And they also classified the files by their extensions, e.g. the "Sound" includes files of ".mp3", ".wav"; the "Video" includes ".rm", ".avi", etc.. [7] only concerned files, which are much simpler in treatment and granularity than digital resources. Things are different when considering the self composition and the topic

integrity. In this paper, we auto-classify the resources by their content or topics instead of merely by file extension. We also study the correlation between file extensions and the resource categories and find that file extensions do not necessarily give clues to category from mutual information viewpoint.

[8] proved that feature selection could notably improve the text category performance and compared several feature selection methods. While in section 6.3, we analyze why the feature selection measures are not necessary in our case.

3 Primary Definition

1. Digital resource representation

Since normally stored in the file system and arranged in the directory tree, a digital resource can be considered as a sub-tree of a big label tree. Its members, files or sub-directories, are supposed to be nodes in this sub-tree. A classification hierarchy can also be denoted as a tree composed by the category nodes, and resources belong to the leaf nodes according to their content topics. The classification task can be defined as mapping a resource from its source tree to certain leaf node of the category tree.

Each resource has at least one member file to represent the content. Not only the member files or sub-directories have names in source directory tree, the resource its own also has a name no matter it really or virtually exists in the directory tree. In the former case, the resource is named as the folder's name. And in the latter case, if the members are mixed with items not belonging to the resource, they need to be picked out for integrity. And the resource name could either be given by people or just take the name of its biggest-size member file.

A resource r can be represented by its feature vector, which only focus on the features' appearance and ignores their sequence or hierarchy. R is the digital resources set, S denotes the features taken from the R's name set; E, the file extensions set; and Z, the byte of R's files member. Then r could be represented as $\forall r \in R$, $r : <S, E, Z>$.

2. Resource frequency (*RF*)

RF is the number of resources where a feature occurs. It borrows the definition of document frequency (*DF*) in information retrieval, and reflects the popularity of a feature among different resources.

3. Category set (*C*)

C includes 12 elements as shown in Table 2 in section 6.1, e.g. "music", "songs", and "drama and opera", etc.. Each category labels a group of digital resources.

4. Segmentation (*seg1* & *seg2*)

Two approaches are used to segment a name string to different snippets. A general assumption on writing custom is people are likely to indicate the change of semantic meaning with explicit or implicit delimiters. So we can segment a name string according to the delimiters, and the result is denoted as *seg1*. The delimiters include punctuations, special symbols, the changes between uppercase and lowercase, etc. These snippets can be further segmented by [9] according to lexicons to split Asian languages without delimiters, and the result is denoted as *seg2*.

5. Data set
CDAL is a 7.4 TB digital resource set used in this paper. It had archived 16.3 thousand resources by sampling public ftp sites of the China from 2003 to 2006. The resources keep the original names and internal structures.

The members of these resources include total 617 thousand files of various data types and sub-directories. In this paper, R_n, F_n and SD_n denote the name set of resource, the file and the sub-directories of resources.

Each resource has a category label in CDAL. The category tree includes 8 branches, such as sound, text, image, etc.. Each branch is sub-divided according to different taxonomy, e.g., image includes "TV program", "film", "picture & photo", and so on.

4 Features for Digital Resource

In this section, we consider the resource attributes and find out how to get suitable features. Each digital resource has external and the internal attributes. The external ones depend on the owner's viewpoint of how a digital resource should be organized. The internal ones reflect the construct of a resource.

The primary external resource attribute is the directories' names along the path from the root to a digital resource in the source tree. The directory names reflect the owner's ontology from how he arranges the content. It is a kind of hard-to-integrate knowledge, since the digital resources are originally scattered in different source trees organized by different users, these trees are heterogeneous from semantic meaning. So the idea of using external attributes is not practical in machine-based resource category, although it works well in human-intelligence-based process [5].

The internal attributes are information inside of a digital resource, such as names of the resource and its members, the files size, file extensions. We will discuss these attributes in detail.

4.1 Names

The functions of an object's name are expression and identification. If merely for identification, names need not to be long, e.g. strings less than 10 characters are enough to distinct the files or sub-directories in a folder. But we found about 2/3 of total 620 thousand names, i.e. elements in R_n, F_n, and SD_n set, range from 10 to more than 100 characters. The names' length distributions and the compositions, studied on [10], show that extra information is added to the names and make them condensate description of the resources. The information includes personal opinions, metadata, and so on. So the names can be treated as descriptive text on the resources, and the segmented snippets as category features for their diversity and relative independence. What to be noted is these descriptions is generally added by resource owners or sharers for better understanding to the resources when they are spread in the Internet. The long and chaotic naming is rather common in Web digital resources.

Snippet Independence. The delimiters are largely used in name strings. In CDAL data set, there are 63 different delimiters in resource name set R_n, 107 in file name set

F_n, and 51 in sub-directory name set SD_n. On average, there are 2 delimiters in a resource name, 1 in a file name and 3 in a sub-directory name. It means that if we want to tokenize the name strings according to delimiters, the total 620 thousand names would provide a snippet set for text statistic analysis. What's more, according to the assumption that people usually separate the name strings when switching semantic meanings, we can roughly consider these snippets are independent.

We also find that there are common naming patterns and frequent snippets in same category, and different naming custom shown by Internet users in different category. For example, the pure-digit file names are more likely to occur in "TV" or "books" resources since there are usually serial content such as of TV plays or book chapters.

As a result, name snippets segmented by *seg1* or *seg2* are considered as category-indicating factors to digital resources classification. They are denoted as feature *S*.

4.2 File Extensions

File extension is roughly taken as clue to resource category, such as the ".pdf" would likely to be in a literal-content resource rather than in a film or a song. If a resource's category can be predicted through its member files' extensions, the classification problem would be greatly simplified. While in this section, we prove their predicting ability should not be over-estimated. It is a related but not the definitive factor.

We use mutual information (MI) to measure the dependence between the file extensions E and categories C. Given $e_v \in E$, $C_i \in C$, the $MI(e_v, C_i)$ means how much knowing feature e_v reduces the uncertainty about it occurs in category C_i. If e_v and C_i are independent, $MI(e_v, C_i)$ is 0, which means knowing e_v does not give information about C_i. In formula (1), $P(e_v, C_i)$ is the probability of e_v and C_i co-occurred. A is the number of times e_v and C_i co-occur; B, the number of times that e_v occurs without C_i; G, number of times C_i occurs without e_v; and N, the total number of digital resources.

$$MI(e_v, C_i) = \log \frac{P(e_v, C_i)}{P(e_v)P(C_i)} \approx \log \frac{A \times N}{(A+B) \times (A+G)} \tag{1}$$

We use resources of six categories in CDAL for demonstration. In each category, the top 10 frequent (*RF*) occurrence file extensions are listed on the left of each column in Table 1, denoted as FT10. If the mutual information of these features giving to their category C_i also keep high rank and keep same sequence with their *RF* list, the file extensions are considered to be important category-indicating features.

In each category C_i, we measure the correlation between the rank list of *RF* and $MI(e_v, C_i)$ with Kendall Tau, marked by "Tau" in Table 1. MIR denotes the rank of $MI(e_v, C_i)$ in given C_i,. If the elements in one list keep same sequence in another list, the Kendall Tau is 1, otherwise 0, and be -1 if the two ranks are in reverse sequence.

Table 1 shows the file extensions of high *RF* in given category are not as crucial as what have been considered in category prediction. The reason is because the digital resource is usually composed by multiple file members to describe an entity or object together, and a frequent observed file extension in the resources of C_i is probably not the typical files for content. E.g., the ".txt" widely exists in "Film" resources, it is by no means a film file but the description or script of the film.

Table 1. The rank of *RF* and MI of file extensions in given categories

Books (1332)		Music (45)		Songs (78)		Film (160)		TV (90)		Game (490)	
FT10	MIR	FT10	MIR	FT10	MIR	FT10	MIR	FT10	MIR	FT10	MIR
pdf	904	mp3	14	mp3	19	avi	29	rm	31	wav	327
txt	1225	mid	13	txt	48	rm	43	rmvb	28	mp3	486
gif	1135	wma	5	wma	24	txt	49	jpg	54	txt	437
htm	1135	txt	19	jpg	41	jpg	51	avi	45	bmp	395
jpg	1097	mpc	3	mpg	23	rmvb	37	mpg	32	mdl	324
html	1303	jpg	18	avi	43	idx	26	asf	26	tga	292
zip	1112	m4a	2	rm	50	sub	27	txt	47	vos	315
pdg	89	mpga	8	asf	31	srt	25	rar	72	map	334
ext	1094	wav	27	wmv	40	rar	59	wmv	21	exe	404
chm	1001	ape	4	dat	32	nfo	46	dat	44	ini	381
Tau=0.378		Tau=0.111		Tau=-0.156		Tau=0.022		Tau=-0.022		Tau=1	

4.3 File Size

We found the size distributions of some typical file extensions vary in different categories. E.g. the 'rm' files in resources of category C_1 ("Music") are most possibly be several megabytes; while in C_2 ("TV") they may range from a few dozens to more than 100 megabytes and in C_7 ("Film") they're likely to be a few hundreds megabytes. Suppose there is a potential size distribution of given file extension in categories, i.e. the size is probably correlated with both category and file extension, denoted as $P(Z|C, E)$, where C is digital resource categories; Z, the file sizes; and E, the file extensions.

The names, file extensions and file sizes are common to digital resources of different categories and data formats. The features of these attributes are easy to get and can be processed in mature methods. If the final category result reaches a practically usable accuracy, the way of using such simple and easy-to-get features can be regarded as both effectiveness and cost-efficiency, comparing with complexity content-based analysis and complete manpower work.

5 Digital Resource Classification

5.1 Naive Bayes Classification

We use Naïve Bayes for digital resources category for two reasons. 1) It performs better than many other supervised learning methods in the case of strong (naive) independence assumption; 2) For features like name snippets and file extensions, it is reasonable to assume their independence.

The predicted category of a resource r is the one with the maximum posteriori probability, i.e. $argmax P(C_i|r)$ [11]. $P(C_i|r)$ can be calculated by formula (2) under the feature independence assumption. $P(C_i)$ is the prior probability of the categories.

$$P(C_i \mid r) = \frac{P(C_i)P(r \mid C_i)}{P(r)} \propto P(C_i)P(r \mid C_i) = P(C_i)P(f_1 \mid C_i)P(f_2 \mid C_i)...P(f_j \mid C_i) \quad (2)$$

If we consider the file size Z as category-indicating features besides S and E, r can be represented as $<s_1,\ldots,s_k, e_1,\ldots,e_m, z_{11},\ldots,z_{mp}>$, where $s_u \in S$, $u \in [1,k]$, $k=|S|$; $e_v \in E$, $v \in [1,m]$, $m=|E|$, and z_{vh} is an integer variable to which the log of a file's size whose extension is e_v maps. We use formula (3) for NB considering the factor Z.

$$P(C_i \mid r) \propto P(C_i)P(s_1,...,s_k \mid C_i)P(e_1,...,e_m, z_{11},...z_{mp} \mid C_i) = P(C_i)P(s_1 \mid C_i)...P(s_n \mid C_i)\bullet$$
$$P(z_{11},...z_{mp} \mid e_1,...,e_m, C_i)P(e_1,...,e_m \mid C_i) = \underline{P(s_1 \mid C_i)...P(s_n \mid C_i)P(e_1 \mid C_i)...P(e_m \mid C_i)P(C_i)}\bullet \quad (3)$$
$$\underline{P(z_{11},...,z_{1k} \mid e_1, C_i)...P(z_{j1},...,z_{jh} \mid e_j, C_i)...P(z_{m1},...,z_{mp} \mid e_m, C_i)}$$

5.2 Evaluation

We use AC as total accuracy metrics, shown in formula (4). For the performance in each category, the standard definition of precision is used, see formula (5).

$$AC = \frac{num\ of\ \ rsc\ which\ category\ is\ correctly\ predicted}{total\ \ testing\ \ rsc\ number} \quad (4)$$

$$prec_i = \frac{num\ of\ \ rsc\ predicted\ to\ be\ C_i\ \ and\ \ correct}{num\ of\ \ rsc\ predicated\ to\ be\ C_i} \quad (5)$$

5.3 Probability Estimation Methods

In statistic learning, a basic need is to estimate the probability of unobserved data since the data in test set is likely not occurred in train set. To avoid a zero-probability in the Naive Bayes product in formula (2-3) and get a reasonable probability estimation, we use SGT [12][13] methods in features of name snippets and file extensions; and for file size, add-delta (also called "Lidstone's Law") instead[14].

The basic idea of SGT is to discount the high frequency data to estimate the "real" probability of low frequency data [12]. In many text category tasks, the terms whose DF is 1 are considered as non-informative to categories. But SGT can make good use of them in estimating the probability of unobserved data. In our case, the features whose RF is 1 are dominate, as shown in 6.2. So we choose SGT.

The frequency distribution is different in file size, where the SGT is not suitable. We use add-delta, which idea is to add a small positive number λ, $\lambda \in (0,1]$ to the observed frequency of the samples. I.e. the original zero probability is assigned a small value. In our experiment, $\lambda = 0.5$.

6 Experiment and Analysis

6.1 Training Set and Testing Set

The digital resources in CDAL are split to train set and test set by stratified random sampling (roughly 4:1) according to the category distribution. There are 12,496 resources in train set and 3137 in the test. Table 2 shows the serial numbers, the category names and their prior probability distribution from C_1 to C_{12}. The resource number in each category is uneven, ranging from less than 100 to nearly 3000, which is one of the reasons why the precision is different in different categories.

Table 2. Prior probability of categories

No.	Category	P	No.	Category	P
1	music	0.0383	7	films	0.1909
2	TV programs	0.0905	8	articles	0.0329
3	drama & opera	0.0046	9	learning materials	0.0195
4	songs	0.1902	10	books	0.2612
5	software	0.1031	11	picture & photo	0.0298
6	voice listening	0.0073	12	games	0.0316

6.2 Distribution of Name Snippet and File Extension Features

Fig. 1 shows the frequency distribution of *seg1* name snippet features and that of the file extensions in the training set. The distribution of the seg2-snippets is similar. In Fig.1, the low frequency snippets are dominant. Only 33,364 unique snippets which *RF* is greater than 1, while the remaining snippets' *RF* is 1. The reason is that there are snippets try to identify the files, sub-directories, or resources in the name strings. While, there are also some common snippets or patterns in names of each category.

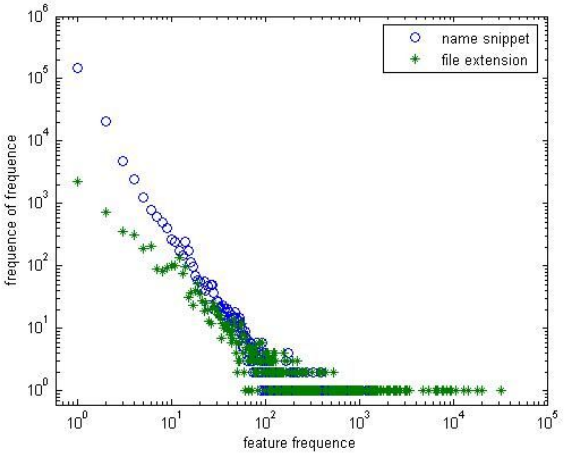


Fig. 1. Distribution of feature frequency. The *X* and *Y* axis respectively denote the frequency of unique features and the frequency of frequency. The circle is the frequency distribution of *seg1* name snippet features, and the star is that of the file extension features. There are 177,224 unique *seg1* snippets and 2,626 unique file extensions in the train set.

6.3 Performance of Naive Bayes Classifier

Precision in different feature groups. In this section, we want to find out which kind of features fit with the NB classifier and help to get the best performance.

We had wondered whether the *RF*=1 features could be neglected as the traditional text classification do, where these features are usually discarded by some feature selection scheme. We also wanted to know which kinds of features, i.e., name snippets, file extensions and file sizes, work more for the category. To study the

problems, 7 different feature groups are constructed to train the classifiers. We will compare their total accuracy *AC* for the answers. See the Table 3.

These feature groups in Table 3 have different composition. The ext_only means we only use file extensions as training features. The snip08_seg2 is composed by snippets whose sum of *RF* occupy 80% of the total snippets' *RF* in training set, and the snippets are got by *seg2* method. The snip10_ext_seg1 means using all name snippets got by *seg1* and all file extensions as training features. And the snip10_ext_seg2_fsize means adding the file size features to snip10_ext_seg2, etc.

1) If the total accuracy *AC* of ext_only is high, the digital resource classification can be simplified, i.e., the file extension features are good enough to distinguish the topic of such a kind of complex data object. 2) If the *AC* value of snip08 and snip10 are close, the features in snip10 but not in snip08 do not contribute much to classification, and thus feature selection need be taken into consideration to discard the useless features. 3) A higher *AC* between snip10_ext_seg1 and snip10_ext_seg2 indicates a better segment method suiting for resource category, since the difference of the two feature groups only lies in their segmentation. 4) The *AC* change between snip10_ext_seg2_size and snip10_ext_seg2 states the improving or decreasing effect of file size features do to the classification performance.

Table 3. Classifier performance of different feature groups with Naive Bayes

Feature group ID	Feature Group	Total accuracy (*AC*) %
1	ext_only	63.8
2	snip08_seg2	69.2
3	snip10_seg2	73.2
4	snip10_ext_seg1	76.1
5	snip08_ext_seg2	79.0
6	**snip10_ext_seg2**	**79.8**
7	snip10_ext_seg2_fsize	73.8

The analysis on the data in Table 3 is as the follows.

1. The lowest accuracy of ext_only shows that the file extensions are not as determinate as assumed, although they reflect partial category information. After all, files in resources are organized together by content not by extensions.
2. There are only name snippet features in snip08_seg2 and snip10_seg2, and the accuracy improvement from 69.2% to 73.2% totally comes from the *RF*=1 features. It is amazing because we have thought those one-time-occurrence snippets have little category information and would not lead to accuracy increasing. Same thing happens in snip08_ext_seg2 and snip10_ext_seg2. The main reason is that adding the snippets of *RF*=1 helps to enlarge the estimated probability space on unobserved data and decreases the over-estimated probabilities of observed data.
3. When combining all seg2-snippets with all file extensions of the training set, i.e. snip10_ext_seg2, the total accuracy is nearly 80% (79.8%), which is the best result in these feature groups. By comparison with the snip10_ext_seg1 (*AC* is 76.1%), the reason of *seg2* superior to *seg1* may be snippets tokenized by word can further split the long snippets to short ones, and then increase the *RF* of some snippets.

4. The file size does not improve the category performance. The *AC* value lowers down when adding file sizes factor to the known best group snip10_ext_seg2. The reason may be file size does not simply depend on category and file extension; or it is not a correlation variable to category at all.

We can draw the conclusion that 1) feature selection used in general text category is not necessary in our case. The once-occurred features are useful in probability estimation on zero-occurred data; 2) file extensions indicate the resource category to some extend, but it is better to combine with the name snippet features since resources are usually organized according to content or topic, which require files with kinds of extensions; 3) *seg2* is better than *seg1* in the case of segmenting name strings for resource category using Naive Bayes; 4) among the 7 feature groups using Naive Bayes and SGT, the best performance is acquired by snip10_ext_seg2.

Precision in different categories. Table 4 shows the precision of different categories in snip10_ext_seg2, the winner group of *AC*. It illustrates the classifier trained by the group is practicable in predicting the categories like "songs", "drama and opera", "software", "films", "books" and "pictures and photos". In these categories, the precision of auto-classification tools can reach or exceed 80%. In real application, it can help to reduce the human cost and speed up the resource reorganization process. As to the low precision categories, their sample numbers make it difficulty to find out category-selective features in statistic meaning.

Table 4. Precision of Naive Bayes in different categories on snip10_ext_seg2

C_i	$prec_i$ (%)	C_i	$prec_i$ (%)
1	68.0	**7**	**83.6**
2	65.1	8	41.3
3	**80.0**	9	46.0
4	**87.4**	**10**	**83.8**
5	**86.7**	**11**	**82.4**
6	27.6	12	73.3

Analysis on Effects Factors of Precision. Besides the training samples' number, we focus on the effect of unobserved features' probability. That is, how to allocate the whole unobserved probability space, estimated by SGT, to single unobserved feature. It is influential since the unobserved features are majority in real cases, the unreasonable probability will hurt the category precision.

The distribution of observed features in Fig. 1 inspires us to assume that the *RF* of unobserved features is probably 1 too. I.e., there may be a uniformity distribution in these features. As a result, we deduce the estimated probability of single *RF*=0 feature P_0 should close to that of the *RF*=1 feature P_1. The data in Table 5 prove that the smaller the difference of P_1 and P_0 is, the higher *AC* value is. As we've known, from snip08_seg2 to snip10_seg2, only the *RF*=1 features are added. I.e., the increase of the *AC* is aroused by these *RF*=1 features which not only adjust the whole probability space allocating from the observed features to unobserved ones but also close the difference of their single probability by reducing the number of unobserved features.

Table 5. The mean and standard deviation between estimated probability on *RF*=1 and *RF*=0 in snip08_seg2 and snip10_seg2 groups of every category. P_1 and P_0 denote the estimated probability of features observed once (*RF*=1) and zero time (*RF*=0).

Feature Group	$Mean(P_1 - P_0)$	$Std(P_1 - P_0)$	$AC(\%)$
snip08_seg2	2.50×10^{-5}	1.62×10^{-5}	69.2
snip10_seg2	1.17×10^{-5}	0.95×10^{-5}	73.2

What's to be mentioned is the mass *RF*=1 features lead to reasonable probability estimation on unobserved features. In the case where *RF*=1 is not dominant in amount, the SGT smoothing might be replaced by other method.

7 Conclusions and Future Work

In this paper, we pay attention to Web digital resources which are numerous and play an important role in knowledge sharing. We evaluate features from resources attributes and find out a practicable classification scheme to process different kind of digital resources in a coherent treatment which have few studies reported before.

We represent each resource with flatten features, i.e. the extensions, file sizes, and the names snippets. These features reflect the diverse composition, general length and various data formats of a resource. We built 7 feature groups to study the problem and drew the conclusions that 1) feature selection used in traditional text category is not necessary when using SGT in the case of dominating low-frequency features and plenty of unobserved features. The *RF*=1 features are especially useful in probability estimation on unobserved ones; 2) file extensions could only indicate the resource category to some extend since resources are usually organized according to content; 3) *seg2* is better than *seg1* in the case of segmenting name strings for resource category using Naive Bayes; 4) among the 7 feature groups using Naive Bayes and SGT, the best performance is acquired by snip10_ext_seg2 feature group, that is, using all name snippets of *seg2* and extensions.

The features we used are suit to NB's independence condition, which helps to lead to the final result. Now, we can predict the category of a given digital resource using the easy-to-get features, which are common to resources of all kinds of category and composition. The auto-classification tool combining Naive Bayes and Simple Good-Turing smoothing is practicable in integrating or reorganizing the mass resources from elsewhere in the Internet. The total accuracy reaches 80%, and the precision in half of the tested categories ranges from 80% to 87%, which can greatly release heavy human work in the initial stage of building a huge digital resource repository or digital libraries. In stead of start from scratch, people only need to refine the result of a rough classification result.

Future work on the digital resource classification includes expanding the related text about the digital resource to improve the category prediction. The related text can be getting from Web pages using the roughly judged category labels and the semantic snippets extracted from the name strings with getting rid of the noise information. We also want to evaluate more features from digital resource attributes.

References

1. Rose, D.E., Danny, L.: Understanding user goals in web search. In: Proceedings of the 13th international conference on World Wide Web(WWW 2004). ACM Press, New York (2004)
2. China Internet Network Information Center, Chinese Internet Development Report (2008), http://www.cnnic.cn/index/0E/00/11/index.htm
3. Internet Archive, http://www.archive.org
4. Chinese Digital Asset Library (restricted to public access for content copyright), http://cdal.net.pku.edu.cn
5. Chen, C., Yan, H.F., Li, X.M.: CDAL: A Scalable Scheme for Digital Resource Reorganization. In: Liu, W., Shi, Y., Li, Q. (eds.) ICWL 2004. LNCS, vol. 3143, pp. 193–200. Springer, Heidelberg (2004)
6. Wolin, B.: Automatic classification in product catalogs. In: Proceedings of the 25th Annual international ACM SIGIR Conference on Research and Development in information Retrieval (SIGIR 2002), Tampere, Finland, August 11 - 15, 2002, pp. 351–352. ACM Press, New York (2002)
7. Chen, H., Wang, J., Han, J.Q., Xie, X.: FTP Files Distribution Characteristics and Their Implications. Computer Engineering and Applications 1, 129–133 (2004)
8. Yang, Y., Pedersen, J.O.: A Comparative Study on Feature Selection in Text Categorization. In: Proceedings of the Fourteenth International Conference on Machine Learning, ICML 1997 (1997)
9. Chinese word parser, ICTCLAS, http://www.nlp.org.cn/
10. Chen, C., Yan, H.F.: Web Resource Naming Conventions and User Behavior Analysis. Journal of the China Society for Scientific and Technical Information (2008) (accepted)
11. Maron, M.E.: Automatic Indexing: An Experimental Inquiry. Journal of the ACM (JACM) 8(3), 404–417 (1961)
12. Gale, W.A., Sampson, G.: Good-Turing Frequency Estimation Without Tears. Journal of Quantitative Linguistics 2, 217–237 (1995)
13. Good Turing Frequency Estimation, http://www.grsampson.net/RGoodTur.htmlH
14. Chen, S.F., Goodman, J.: An Empirical Study of Smoothing Techniques for Languages Modeling. In: Proceeding of 34th Annual Meeting of ACL (1996)

Finding Functional Groups of Objective Rule Evaluation Indices Using PCA

Hidenao Abe[1], Shusaku Tsumoto[1], Miho Ohsaki[2], and Takahira Yamaguchi[3]

[1] Department of Medical Informatics, Shimane University, School of Medicine
89-1 Enya-cho, Izumo, Shimane 693-8501, Japan
abe@med.shimane-u.ac.jp, tsumoto@computer.org
[2] Faculty of Engineering, Doshisha University
1-3 Tataramiyakodani, Kyo-Tanabe, Kyoto 610-0321, Japan
mohsaki@mail.doshisha.ac.jp
[3] Faculty of Science and Technology, Keio University
3-14-1 Hiyoshi, Kohoku Yokohama, Kanagawa 223-8522, Japan
yamaguti@ae.keio.ac.jp

Abstract. To support data mining post-processing, which is one of the important procedures in a data mining process, at least 40 indices are proposed to acquire valuable knowledge. However, since their behaviors have never been elucidated, domain experts are required to spend their time to understanding the meanings of each index in a given data mining result. In this paper, we present an analysis of the behavior of objective rule evaluation indices on classification rule sets by principle component analysis (PCA). Therefore, we carried out a PCA to a dataset consisting of the 39 objective rule evaluation indices. In order to obtain the dataset, we calculated the average values of the bootstrap method on 32 classification rule sets learned by information gain ratio. Then, we identified the seven functional groups of the objective indices based on the PCA. Using this result, we discuss a rule evaluation interface for use by human experts.

1 Introduction

In recent years, enormous amounts of data have been stored on information systems in natural science, social science, and business domains. People have been able to acquire valuable knowledge due to the development of information technology. Besides, data mining techniques combine different types of technologies such as database technologies, statistical methods, and machine learning methods. Data mining has been well-known technique for utilizing data stored on database systems. In particular, if-then rules, which are generated by rule induction algorithms, are considered as one of the highly usable and readable outputs of data mining. However, for large datasets containing hundreds of attributes having noise, the process often yields many thousands of rules. It is difficult for human experts to acquire valuable knowledge from such a large rule set; valuable knowledge is rarely present in the rule set.

T. Yamaguchi (Ed.): PAKM 2008, LNAI 5345, pp. 197–206, 2008.

To support such a rule selection, many studies have been conducted using objective rule evaluation indices such as recall, precision, and other interestingness measurements [1,2,3] (hereafter, we refer to these indices as "objective indices"). Although their properties are identified by their definitions, their behavior on rule sets has been not investigated by any promising method.

We have developed a rule evaluation support method [4] to acquire valuable rules iteratively by modeling rule evaluation criteria specified by human experts based on objective indices as rule evaluation models. However, displaying many rules to human experts is problematic, because they are usually unfamiliar with the detailed meaning between each objective index and a given problem of each data mining project.

With regard to the above-mentioned issues, we carried out principle component analysis to identify the functional behavior groups of objective indices, described in Section 3. Then, using 39 objective indices and classification rule sets from the 32 UCI datasets, we identified the following functional behavior groups: measuring the correctness of a rule, measuring the gain of a promising rule, measuring the difference between theoretical distribution and actual distribution of the classification result of a rule, and so on. Based on the result in Section 4, we discuss a rule evaluation interface for use by human experts using the functional groups of the objective indices.

2 Related Work

Since it is difficult for a human expert to completely evaluate a large number of rules, many conventional studies on rule selection have been conducted in order to eliminate redundant rules from the obtained rules. However, these rule selection techniques are not so active because in each iterative evaluation process, only the calculated values of each rule are presented to human experts. As a more active rule selection method for use in such iterative processes, we have developed a rule evaluation support method to acquire valuable rules using an active human-system interaction. Our method combines multiple objective rule evaluation indices into a composite function called "rule evaluation model".

2.1 Rule Evaluation Index

Many conventional studies have investigated the selection of valuable rules from a large mined rule set by using objective rule evaluation indices. Some of these studies propose the use of indices to determine interesting rules from a large number of rules [1,2,3]. These interestingness measures are based on two different approaches [5]: an objective (data-driven) approach and a subjective approach.

By focusing on the selection of interesting rules using objective indices, researchers have developed more than forty objective indices based on the number of instances, probability, statistics values, information quantity, distance or attributes of rules, and complexity of rules. The behavior of each of these indices with respect to their functional characteristics has been investigated in a number

of studies [6,7,8]. However,any functional relationship among objective indices on a obtained classification rule set has not been analyzed completely.

2.2 A Rule Evaluation Support Method Based on Objective Rule Evaluation Indices

The rule evaluation support method involves two major phases: the training phase and the prediction phase. The system supports iterative rule evaluation processes using objective rule evaluation indices and learning algorithms.

In the training phase (Figure 1), first, a human expert evaluates all or a part of the obtained rules using his/her subjective criterion. Subsequently, by these evaluations and the objective rule evaluation values for the given rules,

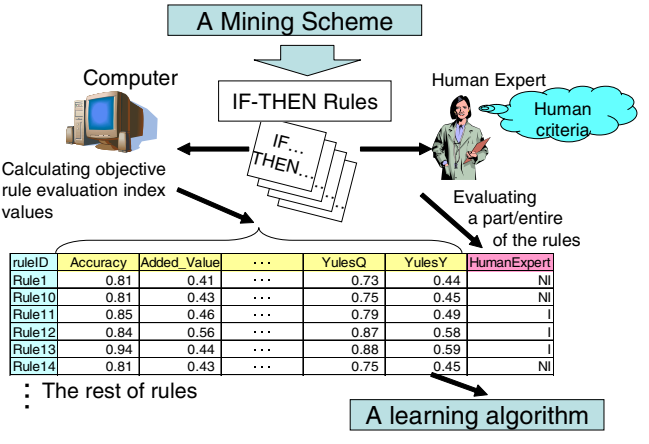


Fig. 1. Training phase of the rule evaluation support method

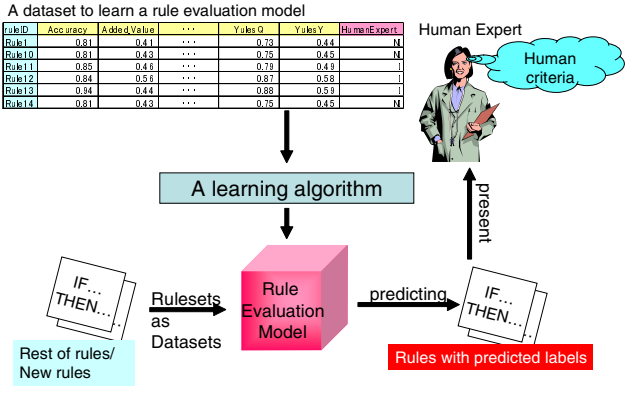


Fig. 2. Prediction phase of the rule evaluation support method

the system obtains a dataset to learn a rule evaluation model. After learning the rule evaluation model, the prediction phase is commenced.

In the prediction phase, the system predicts the evaluations of the remaining rules or new rules based on the values of the objective rule evaluation indices, as shown in Figure 2. The predicted results are presented to a human expert through a user interface. Subsequently, the next training phase is commenced.

3 PCA of a Dataset of Objective Rule Evaluation Indices

In this section, we describe the process of an analysis for identifying functional groups based on the behavior of objective indices. To analyze functional relationships between objective indices, we should determine the following materials: values of objective indices of each classification rule set learned from each dataset and correlation between valued objective indices. The process of the analysis is illustrated in Figure 3.

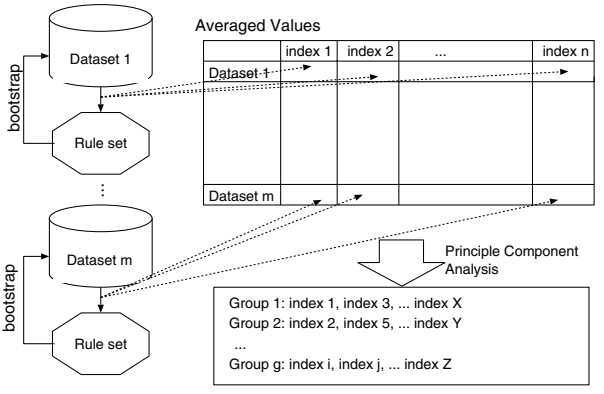


Fig. 3. An overview of the correlation analysis method

First, we obtain multiple rule sets from some datasets to determine the values of objective indices. When determining these values, we should ensure the statistical correctness of each value. Therefore, an adequately large number (> 100) of values obtained from bootstrap samples are averaged.

Using the dataset, we carried out PCA to identify the functional groups of the objective indices. The number of components of the PCA depends on parameter such as λ input by a user.

4 Analysis of the Objective Rule Evaluation Indices on UCI Datasets

In this section, we describe the analysis of 39 objective indices with 32 UCI datasets. Table 1 shows the 39 objective indices investigated and reformulated for determining classification rules by Ohsaki et al. [9].

Table 1. Objective rule evaluation indices for determining classification rules. **P:** Probability of antecedent and/or consequent of a rule. **S:** Statistical variable based on P. **I:** Information of the antecedent and/or consequent of a rule. **N:** Number of instances included in the antecedent and/or consequent of a rule. **D:** Distance of a rule from others, obtained by rule attributes.

Theory	Index Name (**Abbreviation**) [Reference Number of Literature]
P	Coverage (**Coverage**), Prevalence (**Prevalence**)
	Precision (**Precision**), Recall (**Recall**)
	Support (**Support**), Specificity (**Specificity**)
	Accuracy (**Accuracy**), Lift (**Lift**)
	Leverage (**Leverage**), Added Value (**AddedValue**) [2]
	Klösgen's Interestingness (**KI**) [10], Relative Risk (**RelativeRisk**) [11]
	Brin's Interest (**BI**) [12], Brin's Conviction (**BC**) [12]
	Certainty Factor (**CertaintyFactor**) [2], Jaccard Coefficient (**Jaccard**) [2]
	F-Measure (**FMeasure**) [13], Odds Ratio (**OddsRatio**) [2]
	Yule's Q (**YuleQ**) [2], Yule's Y (**YuleY**) [2]
	Kappa (**Kappa**) [2], Collective Strength (**CollectiveStrength**) [2]
	Gray and Orlowska's Interestingness weighting Dependency (**GOI**) [14]
	Gini Gain (**GiniGain**) [2], Credibility (**Credibility**) [15]
S	χ^2 Measure for One Quadrant (**ChiSquareone**) [16]
	χ^2 Measure for Four Quadrant (**Chisquarefour**) [16]
I	J-Measure (**JMeasure**) [17], K-Measure (**KMeasure**) [18]
	Mutual Information (**MutualInformation**) [2]
	Yao and Liu's Interestingness 1 based on one-way support (**YLI1**) [3]
	Yao and Liu's Interestingness 2 based on two-way support (**YLI2**) [3]
	Yao and Zhong's Interestingness (**YZI**) [3]
N	Cosine Similarity (**CosineSimilarity**) [2]
	Laplace Correction (**LaplaceCorrection**) [2]
	ϕ Coefficient (**PhiCoefficient**) [2]
	Piatetsky-Shapiro's Interestingness (**PSI**) [19]
D	Gago and Bento's Interestingness (**GBI**) [20]
	Peculiarity (**Peculiarity**) [21]

As for the datasets, we have taken the 32 datasets from UCI machine learning repository[22], which are distributed with Weka [23].

By using the datasets, we obtained rule sets using PART [24] implemented in Weka. PART constructs a rule set on the basis of information gain ratio. This implies that the obtained rule sets are biased with the correctness of each classification problem.

4.1 Dataset Consisting of the 39 Objective Indices and the Setup of PCA

For the 32 datasets, we obtained the rule sets using PART. This procedure is repeated 1000 times with bootstrap resampling for each dataset. As a representative value of each bootstrap iteration, the average for a rule set was calculated. Then, we averaged the average values obtained by the 1000 iterations.

Table 2. Summary of PCA of the dataset consisting of the 39 objective rule evaluation indices

Component	λ	Acc. Contribution Ratio (%)
1	20.43	52.39
2	3.74	61.98
3	3.39	70.68
4	2.58	77.30
5	2.57	83.90
6	2.04	89.12
7	1.23	92.26

Table 3. Contribution ratios of the PCA

	COMPONENT						
	1	2	3	4	5	6	7
Coverage	0.91	−0.31	0.06	−0.09	0.02	−0.08	0.07
Prevalence	0.21	−0.66	0.18	0.38	−0.03	−0.56	−0.07
Precision	0.96	−0.09	0.03	0.17	−0.14	0.06	−0.05
Recall	0.90	0.14	−0.02	−0.29	0.22	−0.10	−0.02
Support	0.91	−0.31	0.06	−0.09	0.02	−0.08	0.07
Specificity	−0.51	0.51	−0.17	0.43	−0.39	0.09	0.02
Accuracy	−0.01	0.65	−0.19	−0.37	−0.01	0.59	0.09
Lift	0.44	0.53	−0.02	0.21	0.43	0.16	−0.29
Leverage	0.97	0.11	−0.05	−0.10	−0.10	−0.08	0.04
AddedValue	0.97	0.00	−0.01	0.10	−0.09	0.12	−0.08
RelativeRisk	0.34	0.42	0.15	0.55	0.58	0.01	0.01
Jaccard	0.98	0.06	0.01	−0.12	0.09	−0.08	0.02
CertaintyFactor	0.97	−0.03	0.00	0.12	−0.12	0.10	−0.06
OddsRatio	0.24	0.14	0.20	0.38	0.72	−0.09	0.29
YulesQ	0.79	−0.37	−0.28	0.16	−0.05	0.26	0.08
YulesY	0.93	0.30	0.01	0.06	−0.10	−0.01	−0.07
Kappa	0.94	0.31	−0.06	−0.08	0.03	−0.07	−0.01
KI	0.98	0.06	0.00	0.04	−0.12	0.04	−0.04
BI	0.81	−0.53	0.09	0.08	−0.03	0.14	0.03
BC	−0.75	0.54	0.17	0.01	−0.10	−0.27	−0.01
GOI	0.96	−0.15	−0.01	0.08	−0.06	0.13	−0.07
CollectiveStrength	0.17	0.32	−0.08	−0.59	0.22	−0.33	0.19
Credibility	−0.06	0.31	0.16	0.58	−0.22	0.01	0.31
LaplaceCorrection	0.91	−0.05	0.11	0.27	−0.23	−0.02	−0.03
ChiSquareone	0.12	0.17	0.93	0.00	0.10	0.15	0.01
ChiSquarefour	0.14	0.01	0.96	−0.13	−0.02	0.13	0.00
GiniGain	0.92	0.20	0.09	−0.16	−0.06	−0.15	0.09
MutualInformation	−0.20	0.03	0.14	−0.11	0.14	−0.09	−0.74
JMeasure	0.85	−0.17	−0.04	−0.07	0.08	0.04	0.21
YLI1	0.82	0.48	0.01	0.18	0.00	0.07	−0.13
YLI2	0.95	0.22	−0.03	−0.13	−0.07	−0.11	0.04
YZI	0.90	0.26	0.08	−0.21	0.00	−0.18	0.10
KMeasure	0.75	−0.54	−0.17	−0.01	0.10	0.27	0.01
PhiCoefficient	0.55	0.72	−0.16	−0.04	−0.20	−0.24	0.06
PSI	0.09	−0.10	0.95	−0.19	−0.11	0.09	−0.09
CosineSimilarity	1.00	−0.01	0.01	−0.04	0.00	0.00	−0.03
GBI	−0.68	−0.33	0.08	−0.20	0.39	0.17	0.21
Peculiarity	−0.32	0.15	0.60	−0.03	−0.37	0.02	0.20
FMeasure	0.99	−0.01	0.00	−0.07	0.03	−0.02	−0.02

Using the average values for each dataset, we carried out the PCA. We set $\lambda_i > 1.0$ to identify significant i components on the dataset. We performed the PCA using SPSS 14.0J with varimax rotation.

4.2 Result of PCA of the Dataset

By the PCA, we identified seven components by performing 16 rotations. Figure 2 shows the summary of the PCA.

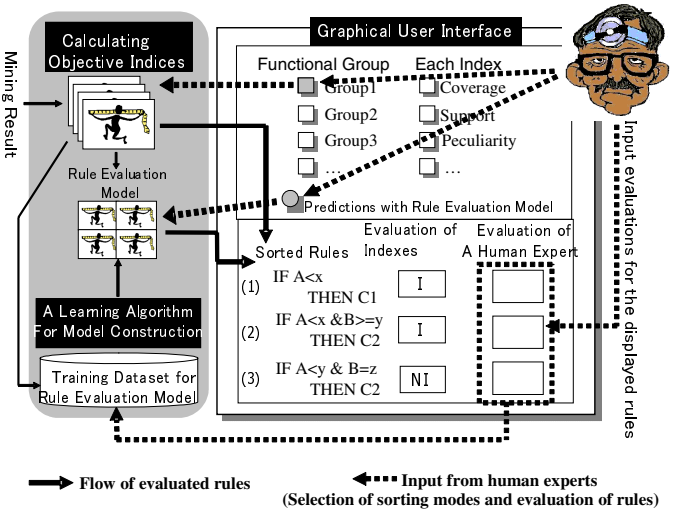


Fig. 4. Overview of the rule evaluation support system

Table 3 shows the contribution ratios of each objective on the seven components.

The first component consists of objective indices that measure the correctness of the classification result of each rule. Objective indices greatly contribute to the second component measure the gain of the classification result of each rule as compared to just predicting the mentioned class value. Objective indices that contribute to the third component are those measuring the difference between the theoretical and actual distribution of the classification result of a rule. The forth component mainly comprises objective indices such as OddsRatio, RelativeRisk and Credibility. The fifth component is similar to the forth one except that the fifth component yields a slightly different correctness value than that yielded by the fourth component. The sixth component measures the accuracy of each rule by considering both the correctness of a mentioned/unmentioned class value and the fewness of fatal classifications. The seventh component provides a negative evaluation based on Mutual Information of each rule.

5 Discussion

The results shows that the functional groups of the objective indices in this study are clearer than those of our previous study [25]. Most of these groups are similar to those of described in [25]. However, for example, the first component expresses the function of the biggest group of similar pairs, which consists of Coverage, Precision, Recall, Support, Leverage, AddedValue, Jaccard, CertaintyFactor, YulesQ, YulesY, Kappa, KI, BI, GOI, LaplaceCorrection, GiniGain, JMeasure, YLI1, YLI2, YZI, KMeasure, CosineSimilarity, and FMeasure, and also expresses the discrepant pairs consisted of GOI, BI, YulesQ, KMeasure and BC.

Considering the result of the PCA, we can extract more actual evaluation axes by using this method to implement a rule evaluation support system, as shown in Figure 4. By presenting multiple granularities of objective rule evaluation criteria, the rule evaluation support system can collect evaluations from human experts more exactly. The system yields some functional groups ("Functional Group" in Figure 4) corresponding to components obtained by PCA as described in Section 4. On the basis of the components shown in Table 2 and Table 3, system developers describe the meaning of each functional group. Subsequently, human experts receive the results of sorting done on the basis of the functional groups or each objective index.

In this framework, we consider the prediction results of rule evaluation models as being coarse grain information presented to human experts. Then, the functional groups represent finer grain information on the rules. Each objective index provides the finest grain information on the rules. By using the multiple granularity of information on objective rule evaluation, the system not only collects evaluations from human experts rapidly but also presents objective features of their interests based on the indices.

6 Conclusion

In this paper, we described a method to identify the functional groups of the objective rule evaluation indices.

We investigated the behavior of objective indices using the 32 UCI datasets and their rule sets as an actual example. With regarding to the result based on PCA, the seven groups are found as the functional groups in cross-sectional manner for theoretical backgrounds of the objective indices. These groups summarize the detailed function of each objective rule evaluation index. The result of this study are expected to aid human experts in browsing through many if-then rules in order to acquire valuable rules through a rule evaluation support system.

In the future, we intend to implement a system for supporting rule evaluation by human experts based on the functional groups of objective indices and rule evaluation models.

References

1. Hilderman, R.J., Hamilton, H.J.: Knowledge Discovery and Measure of Interest. Kluwer Academic Publishers, Dordrecht (2001)
2. Tan, P.N., Kumar, V., Srivastava, J.: Selecting the right interestingness measure for association patterns. In: Proceedings of International Conference on Knowledge Discovery and Data Mining KDD 2002, pp. 32–41 (2002)
3. Yao, Y.Y., Zhong, N.: An analysis of quantitative measures associated with rules. In: Zhong, N., Zhou, L. (eds.) PAKDD 1999. LNCS (LNAI), vol. 1574, pp. 479–488. Springer, Heidelberg (1999)

4. Abe, H., Tsumoto, S., Ohsaki, M., Yamaguchi, T.: A rule evaluation support method with learning models based on objective rule evaluation indexes. In: Proceeding of the IEEE International Conference on Data Mining ICDM 2005, pp. 549–552 (2005)
5. Freitas, A.A.: On rule interestingness measures. Knowledge-Based Systems 12(5-6), 309–315 (1999)
6. Vaillant, B., Lenca, P., Lallich, S.: A clustering of interestingness measures. In: Suzuki, E., Arikawa, S. (eds.) DS 2004. LNCS (LNAI), vol. 3245, pp. 290–297. Springer, Heidelberg (2004)
7. Huynh, X.H., Guillet, F., Briand, H.: A data analysis approach for evaluating the behavior of interestingness measures. In: Hoffmann, A., Motoda, H., Scheffer, T. (eds.) DS 2005. LNCS (LNAI), vol. 3735, pp. 330–337. Springer, Heidelberg (2005)
8. Blanchard, J., Guillet, F., Gras, R., Briand, H.: Using information-theoretic measures to assess association rule interestingness. In: Proceedings of the fifth IEEE International Conference on Data Mining ICDM 2005, pp. 66–73. IEEE Computer Society Press, Los Alamitos (2005)
9. Ohsaki, M., Abe, H., Yokoi, H., Tsumoto, S., Yamaguchi, T.: Evaluation of rule interestingness measures in medical knowledge discovery in databases. Artificial Intelligence in Medicine 41(3), 177–196 (2007)
10. Fayyad, U.M., Piatetsky-Shapiro, G., Smyth, P., Uthurusamy, R. (eds.): Explora: A Multipattern and Multistrategy Discovery Assistant. In: Advances in Knowledge Discovery and Data Mining, pp. 249–271. AAAI/MIT Press, California (1996)
11. Ali, K., Manganaris, S., Srikant, R.: Partial classification using association rules. In: Proceedings of the International Conference on Knowledge Discovery and Data Mining KDD 1997, pp. 115–118 (1997)
12. Brin, S., Motwani, R., Ullman, J., Tsur, S.: Dynamic itemset counting and implication rules for market basket data. In: Proceedings of the ACM SIGMOD International Conference on Management of Data, pp. 255–264 (1997)
13. Rijsbergen, C.: Information retrieval, ch. 7 (1979), `http://www.dcs.gla.ac.uk/Keith/Chapter.7/Ch.7.html`
14. Gray, B., Orlowska, M.E.: CCAIIA: Clustering categorical attributes into interesting association rules. In: Wu, X., Kotagiri, R., Korb, K.B. (eds.) PAKDD 1998. LNCS (LNAI), vol. 1394, pp. 132–143. Springer, Heidelberg (1998)
15. Hamilton, H.J., Shan, N., Ziarko, W.: Machine learning of credible classifications. In: Australian Conference on Artificial Intelligence AI 1997, pp. 330–339 (1997)
16. Goodman, L.A., Kruskal, W.H.: Measures of association for cross classification. Springer Series in Statistics, vol. 1. Springer, Heidelberg (1979)
17. Smyth, P., Goodman, R.M.: Rule induction using information theory. In: Piatetsky-Shapiro, G., Frawley, W.J. (eds.) Knowledge Discovery in Databases, pp. 159–176. AAAI/MIT Press (1991)
18. Ohsaki, M., Kitaguchi, S., Kume, S., Yokoi, H., Yamaguchi, T.: Evaluation of rule interestingness measures with a clinical dataset on hepatitis. In: Boulicaut, J.-F., Esposito, F., Giannotti, F., Pedreschi, D. (eds.) PKDD 2004. LNCS (LNAI), vol. 3202, pp. 362–373. Springer, Heidelberg (2004)
19. Piatetsky-Shapiro, G.: Discovery, analysis and presentation of strong rules. In: Piatetsky-Shapiro, G., Frawley, W.J. (eds.) Knowledge Discovery in Databases, pp. 229–248. AAAI/MIT Press (1991)
20. Gago, P., Bento, C.: A metric for selection of the most promising rules. In: Żytkow, J.M. (ed.) PKDD 1998. LNCS (LNAI), vol. 1510, pp. 19–27. Springer, Heidelberg (1998)

21. Zhong, N., Yao, Y.Y., Ohshima, M.: Peculiarity oriented multi-database mining. IEEE Transactions on Knowledge and Data Engineering 15(4), 952–960 (2003)
22. Hettich, S., Blake, C.L., Merz, C.J.: UCI repository of machine learning databases, University of California, Department of Information and Computer Science, Irvine, CA (1998), http://www.ics.uci.edu/~mlearn/MLRepository.html
23. Witten, I.H., Frank, E.: Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations. Morgan Kaufmann, San Francisco (2000)
24. Frank, E., Witten, I.H.: Generating accurate rule sets without global optimization. In: The Fifteenth International Conference on Machine Learning, pp. 144–151 (1998)
25. Abe, H., Tsumoto, S.: Analyzing behavior of objective rule evaluation indices based on pearson product-moment correlation coefficient. In: An, A., Matwin, S., Raś, Z.W., Ślęzak, D. (eds.) Foundations of Intelligent Systems. LNCS (LNAI), vol. 4994, pp. 84–89. Springer, Heidelberg (2008)

Organizational Knowledge Transfer of Intelligence Skill Using Ontologies and a Rule-Based System

Masao Okabe[1], Masahiko Yanagisawa[1], Hiroshi Yamazaki[1], Keido Kobayashi[2], Akiko Yoshioka[2], and Takahira Yamaguchi[2]

[1] Tokyo Electric Power Co., Inc., 1-1-3 Uchisaiwai-cho, Chiyoda-ku Tokyo 100-8560, Japan
{okabe.masao,yanagisawa.masahiko,yamazaki.hiroshi}@tepco.co.jp
[2] Department of Administration Engineering, Faculty of Science and Technology, Keio University, 3-14-1 Hiyoshi, Kohoku-ku, Yokohama-shi Kanagawa-ken 223-8522, Japan
{k_koba,y_aki,yamaguti}@ae.keio.ac.jp

Abstract. In recent automated and integrated manufacturing, so-called intelligence skill is becoming more and more important and its efficient transfer to the next generation is a key to keeping a factory productive and competitive. But, currently, it needs costly on-the-job training (OJT) and it is crucial to reduce its cost. In this paper, we propose a new approach without OJT, that is, combinational usage of ontologies and a rule-based system. It helps domain experts externalize their tacit intelligence skill and helps novices internalize it.

1 Introduction

Recently, in manufacturing, there has been increasing concern on transfer of skill to the next generation (see e.g. [1][2]). It is a key to keeping a factory productive and competitive. The problem is that most of the skill is not externalized and its transfer has to be mainly done by on-the-job training (OJT) and costs a lot of time and money.

To reduce its cost by OJT, there are several proposals using information technology, mainly, multimedia and virtual reality technology (see e.g. [3]). They focus on so-called "craft skill", which is skill such as to create a complex mold with high precisions. Chuma [4], however, points out that in manufacturing, there is another crucial skill called "intelligence skill". Intelligence skill is skill to do almost any kinds of jobs in manufacturing and to detect expected flaws. In a recent automated and integrated manufacturing plant, to operate it efficiently, intelligence skill is becoming more and more important, while craft skill is being replaced by computerized numerical control. Intelligence skill requires integrated knowledge regarding various aspects of a whole plant but it can be decomposed into pieces of knowledge. Hence, the key issue is how to combine pieces of knowledge logically, depending on various situations and, therefore, for intelligence skill, ontology and a rule-based system are more suggestive than multimedia and virtual reality technology.

SECI model by Nonaka et al. [5] is also suggestive for organizational knowledge transfer. It shows how organizational knowledge is created by syntheses of tacit and explicit knowledge, Socialization, Externalization, Combination and Internalization.

T. Yamaguchi (Ed.): PAKM 2008, LNAI 5345, pp. 207–218, 2008.

In Socialization, new organizational tacit knowledge is formed from personal tacit knowledge. For that, the spiral of Externalization, Combination and Internalization is necessary. In Externalization, new explicit knowledge is formed from tacit knowledge, in Combination, combined explicit knowledge is formed from different kinds of explicit knowledge, and in Internalization, new personal tacit knowledge is formed from combined explicit knowledge. A place where these syntheses are conducted is called “ba” [6]. Hijikata et al. [7] propose a computer-aided model that adopts SECI model and that helps two experts externalize and combine their tacit knowledge. In their proposal, the computer verifies the consistency of the knowledge externalized by two experts and provides them issues to be discussed to get consistently externalized and combined knowledge. Hence, the computer here can be said to provide a communication “ba” for externalization and combination.

In this paper, we focus on organizational knowledge transfer rather than organizational knowledge creation, but it can be viewed as the inverse of Socialization in SECI model and therefore it can be done by Externalization, Combination and Internalization. Hence, we propose ontologies and a rule-based system that help domain experts externalize and combine their intelligence skill and also help novices internalize it.

In the next section, we outline our proposal and also introduce the ontology repository called GEN we have developed. In section 3, we present a case study, which has been done, developing ontologies and a rule-based system using GEN. In section 4, we evaluate and discuss our proposal based on the case study. Finally, in section 5, we summarize our proposal and point out some future works.

2 Ontologies and a Rule-Based System for Organizational Transfer of Intelligence Skill, and GEN

We propose ontologies and a rule-based system for organizational transfer of intelligence skill. Ontology is defined as "a specification of a conceptualization" [8]. Usually, it consists of terms and semantic relations among them. But, in this paper, we take ontology in broader sense. It is not only about terms but also about pieces of knowledge. Therefore, what consists of pieces of knowledge and semantic relations among them is also an ontology. Well-structured explicit knowledge that is externalized from tacit intelligence skill and combined can be an ontology.

Fig. 1 illustrates the overall structure of our proposal with a simplified example. In the following sub-sections, we first explain ontologies we propose and how novices can internalize them as intelligence skill using a rule-based system and, then, explain how domain experts can externalize and combine their tacit intelligence skill as ontolgies using a rule-based system. Finally, we briefly introduce GEN.

2.1 For Internalization by Novices

Even if intelligence skill is decomposed into pieces of knowledge and they are externalized in a natural language, it is difficult for novices to understand them. There are two reasons. One is that natural language description of a piece of knowledge includes technical terms that novices cannot understand. The other is that to understand a piece of knowledge, novices need to know other related pieces of knowledge

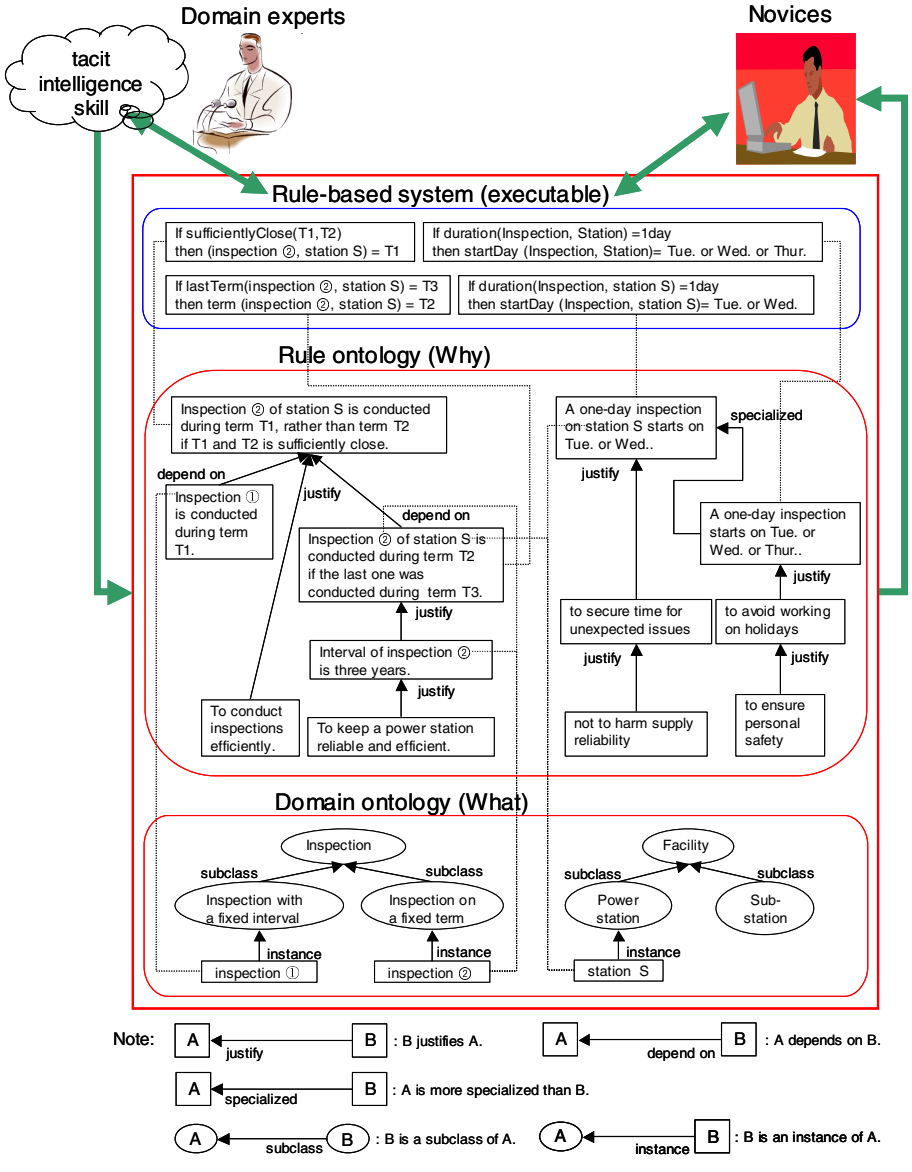


Fig. 1. Overall structure with a simplified example

such as its deep knowledge that justifies it [9] and to recognize where it stands among the whole knowledge. To solve the problem, we propose two kinds of ontology. One is domain ontology and the other is rule ontology. Rule ontology is about rules in a target domain, where as domain ontology is about technical terms in the domain. A rule, here, is a piece of knowledge that governs a job in the domain. In rule ontology, each rule may be written in a natural language but is treated as a primitive. A rule

ontology consists of rules as primitives and semantic relations among them. The semantic relations include;

- relation “justify”, which is a relation between a deep rule and its shallow rule
- relation “depend on”, which is a relation between a rule and a rule whose application is prerequisite for its application
- relation “specialized”, which is a relation between a specialized rule and its generalized rule
- relation “override”, which is a relation between a specialized rule with some overrides and its generalized rule that it overrides.

A domain ontology helps novices understand technical terms in rules, having relations to a rule ontology. Semantic relations in a rule ontology gives novices related rules and help novices understand where each rule stands among the whole rules.

But even with these helps, it is still difficult for novices to internalize them. To internalize them, novices need experience to perform a job properly using them. For that purpose, the rules directly used for the job are converted to a rule base and its rule-based system is also provided. Novices can try performing the job by themselves as many times as necessary and check the results, comparing with the outputs of the rule-based system. Then, novices can sufficiently internalize them.

2.2 For Externalization and Combination by Domain Experts

Domain experts are not familiar with externalizing and combining their intelligence skill to ontologies because usually they have got the skill by OJT and have never externalized them. Hence, some kind of initial ontologies are necessary to guide them to develop ontologies properly. Especially, it is not easy for them to recognize rules explicitly since they perform a job based on their tacit intelligence skill and not on a rule ontology. A rule-based system helps them refine the rule ontology, giving them an opportunity to check the outputs that it produces based on the rules externalized.

Ontologies, especially a rule ontology, need to be updated as the intelligence skill evolves. Domain experts are expected to maintain them but need motivation for that since they do not need them to perform the job. A rule-based system gives them the output of the job semi-automatically and motivates them.

2.3 GEN (General knowlEdge Navigator)

To support and examine our proposal, we have developed an ontology repository specialized for knowledge management called GEN (General knowlEdge Navigator) [10]. GEN is similar to Protégé [11] but is more end-user oriented and suitable for structured textual information like a rule ontology. GEN intends to provide a “ba” where domain experts externalize and combine their tacit intelligence skill and concurrently novices internalize it. Therefore, it supports multi-user concurrent use. It also has a reasoner so that domain experts can develop a rule-based system by themselves.

Fig. 2 shows how a semantic relation of a rule ontology is represented in GEN. When a slot, for example, “override”, is defined and its link is added in GEN, GEN automatically defines its inverse slot “overriddenBy” and also adds the inverse link.

Fig. 2. Example of a semantic relation in GEN

3 A Case Study from TEPCO

Having described our proposal, let us now turn to a case study and examine our proposal in details. In this case study, we have focused on some specific job on hydroelectric power stations at some remote control and maintenance office in Tokyo Electric Power Co., Inc. (TEPCO). It is so-called "inspection and maintenance work scheduling job". What we did first is similar to OJT. We, as novices, tried to do the job, referring to the previous performance, under the direct supervision of the domain experts of this job. Second, we externalized what we learned from the domain experts and reorganized and combined them as a rule ontology and a domain ontology in GEN. Third, we also developed a rule-based system specific to this job, that is, a scheduling subsystem on GEN. Finally, we examined whether the ontologies and the rule-based system worked as we expected.

3.1 Case Description

The system of hydroelectric power stations in TEPCO has a long history and is now highly automated and integrated. All the hydroelectric power stations are unmanned and a remote control and maintenance office is responsible to remote-operate and maintain all the hydroelectric power stations along a river system, which vary from very old small ones to state-of-the art large-scale pumped storage ones.

The "inspection and maintenance work scheduling job" is mainly to make out yearly inspection and maintenance work schedules of generators and other devices of all power stations controlled by a remote control and maintenance office. The schedule is made out so that it minimizes discharged water that is not used for power generation, under various constraints such as;

- statutory inspection interval of each device
- agreements with agricultural unions and other outside associations

- natural environment conditions
- operational conditions among interrelated facilities etc.

This is a typical job that needs intelligence skill since it requires a variety of knowledge on the whole power stations. Moreover, this is not a well-defined optimization problem. Some of the constraints are not mandatory but desirable and in most cases there is no feasible solution that strictly conforms to all the constraints. In case that there is no strictly feasible solution, sophisticated intelligence skill is required to determine what constraints should be loosened, depending on a situation. Most of the knowledge was not externalized and the skill for the job had to be transferred by costly OJT.

3.2 Rule Ontology

What we learned from the domain experts was decomposed into rules, which can be expressed by sentences with several tens of words. Most of the rules we got are shallow knowledge that can be used directly to make out the schedule, but some are deep knowledge that justifies the shallow knowledge. Since relation "justify" between a deep rule and a shallow rule is relative and hierarchal, we chose a set of the deepest rules, which are the objectives and basic constraints of the job. Fig. 3 shows some examples. There are also rules that do not belong to the set of the deepest rules but that no other rules are deeper than, such as "Scale grows during the warm season" as shown at Fig. 3. These rules are objective facts that justify the shallow rules.

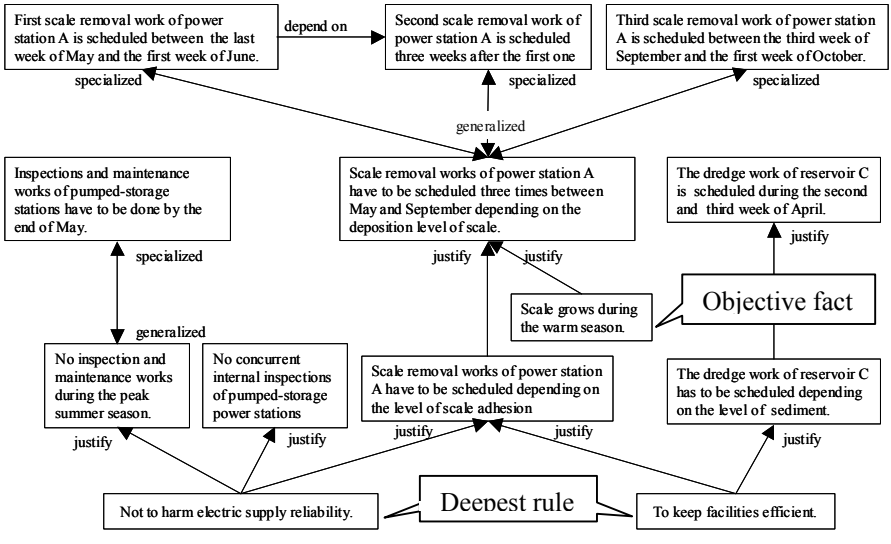


Fig. 3. Example of semantic relations of the rule ontology

We also developed a class hierarchy of the rule ontology, where each rule is treated as an individual of the rule ontology. There were two intensions. One is to make it easy to apply necessary rules to proper works. The other is to make it easy to maintain

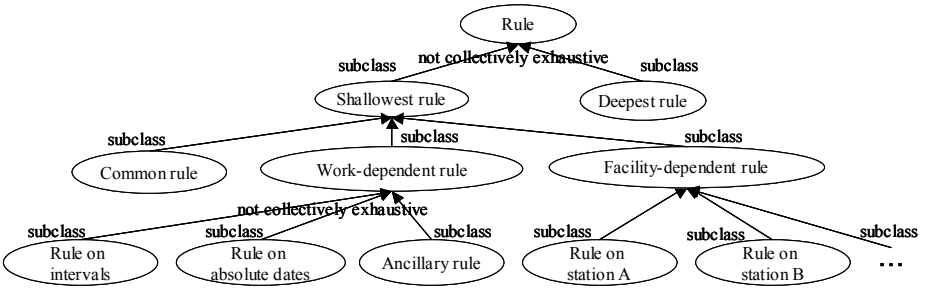


Fig. 4. Class hierarchy of the rule ontology

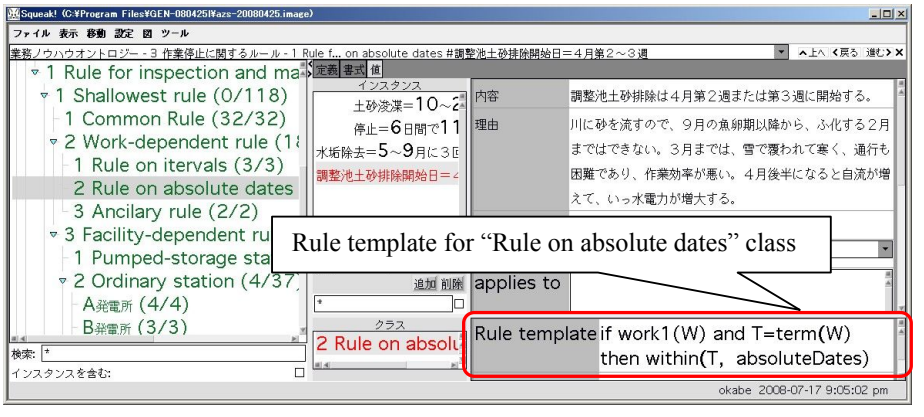


Fig. 5. Example of a rule template

Table 1. Number of classes and instances of the ontologies

	Classes	Instances	Note
Rule ontology	20	133	90 instances among 133 intances were converted to a rule base.
Domain ontology	55	292	

the rules. Fig. 4 shows the class hierarchy of the rule ontology we developed. It helped novices apply appropriate rules to appropriate works. In addition, we could add a class property called “rule template” to several classes as shown at Fig. 5. The rule template was a good guide for the domain experts to add a new rule.

Table 1 shows number of classes and instances of the rule ontology.

3.3 Domain Ontology

Domain ontologies were created from the technical terms in the rules. Table 1 shows number of classes and instances of the domain ontologies. Among them, a typical one is a facility ontology, which consists of such as power stations, sub-station and

transmission lines. TEPCO already has a kind of class hierarchy of facilities with a long history although TEPCO does not call it a class hierarchy. Basically we adopted this class hierarchy for the facility ontology because we thought that the facility ontology should be commonly applicable to the broader domain.

The facility ontology has semantic relations to the rule ontology. One is a class-individual relation between a class of the facility ontology and the rules that apply to the class. The other is an individual-individual relation between an individual of the facility ontology and the rules that apply to the individual. Fig. 6 shows how the facility ontology is represented and related to the rule ontology in GEN.

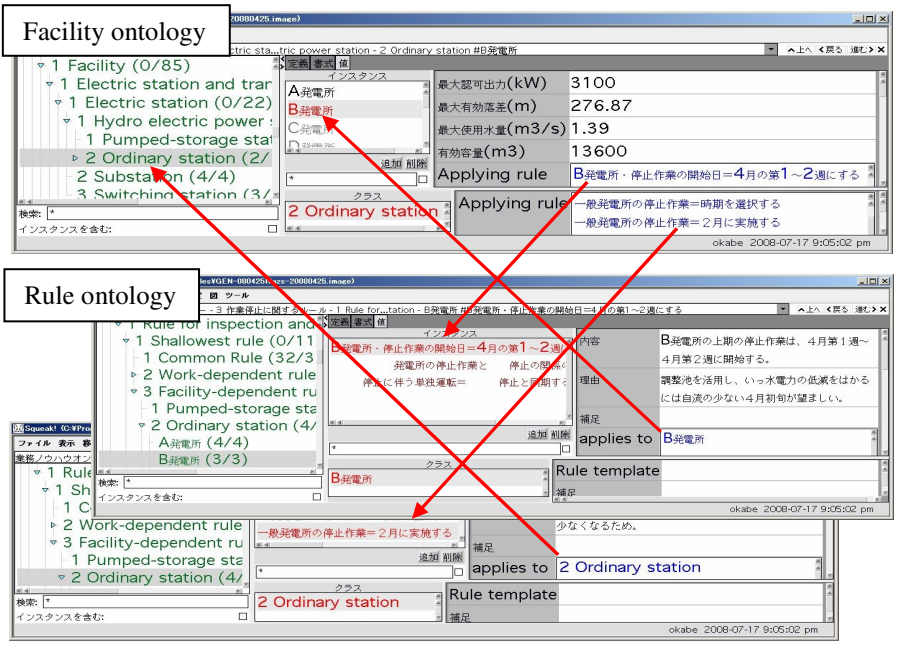


Fig. 6. Examples of the facility ontology and the rule ontology

3.4 Scheduling Subsystem

The shallowest rules (90 rules among 133 rules, see Table 1.) were converted to if-then style syntax so that they run on a reasoner on GEN. However, it was not easy to get a proper solution. As mentioned before, this is not a well-defined problem and most of the rules are not mandatory and may contradict. Here is a simplified example to illustrate how they contradict.

Example 1

Rule 1. if generator(G) and inspection(I, G) and T=term(I) then within(T, January)
Rule 2. if inspection(I, generator1) and T=term(I) then within(T, February)
where G, I, T are variables that implicitly mean a generator, an inspection, a term, respectively and generator1, January and February are individual names.

In this case, apparently, Rule2 is an exception of Rule 1 and Rule 1 can be refined as follows so that they do not contradict.

Rule 1. if generator(G) and not (G=genetor1) and inspection(I, G) and T=term(I)
then within (T, January)

But, in real cases, this kind of refinements makes rules complex and difficult especially for novices. The pair of the original Rule1 and Rule2 is easier since novices usually intuitively understand that Rule2 overrides Rule 1 since Rule 2 is more specialized than Rule1. Hence, instead of making this kind of refinement of Rule1, it is better that GEN specifies explicitly that Rule2 overrides Rule1 using the semantic relation "override" and that its reasoner understands this relation.

Even having done so, however, it was still not easy to get a suitable feasible solution. Here is another simplified example to illustrate the difficulty.

Example 2
Rule 3. day(startDate(inspection1))=Wednesday
Rule 4. startDate(inspection1)= March 1st
where inspection1 is an individual name for some specific inspection.

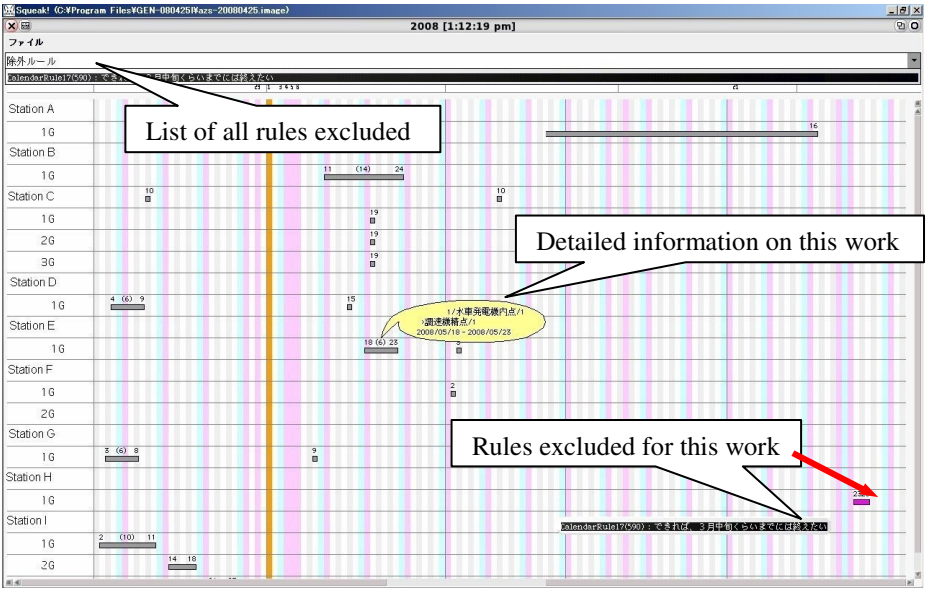


Fig. 7. Output of the scheduling subsystem

Then, a question is which of Rule 3 or Rule 4 should be excluded or loosened if March 1st is not Wednesday in the target year. This kind of situation occurs very often usually in much more complex manner. To treat these situations properly occupies an important part of the intelligence skill of this job. First, we tried to enumerate all such cases and give a specific solution for each case. But, there are a lot of cases and it is difficult to investigate all the cases. Moreover, even if all the cases are investigated at

one time with a great effort, a lot of new cases may appear when a rule is added or updated and it is almost impossible to maintain them.

Hence, we changed the strategy. Giving up enumerating all the infeasible cases, we decided to give each rule a priority number. If there is no feasible solution, then rules that have the least priority are simply ignored. Now, the way to externalize the tacit knowledge to treat infeasible cases became dramatically simple. It is now to give a priority number to all the rules by trial and error using the scheduling subsystem so that they can work well in any cases. Practically, we may assume that if they work well in many cases, they presumably work well in all other cases. By trial and error, we were successful in giving a suitable priority number to each rule. Fig. 7 shows a part of the output based on them.

4 Evaluation and Discussion

In the previous section, we have presented the case study of "inspection and maintenance work scheduling job" using GEN. It has given us some empirical evidences of the effectiveness of our proposal although one case is not enough to evaluate it fully. We discuss them from both of the points of externalization and combination by domain experts and of internalization by novices.

4.1 Externalization and Combination by Domain Experts

Apparently the domain experts had the intelligence skill to do the job well. But they had no idea what extent of details and carefulness is necessary in externalization for the novices to understand it. For example, they tended to externalize only the deepest and the shallowest rules because they did not recognize that the novices had difficulty in relating them without in-between rules.

The ontologies we developed had a major role in solving the problem. First, the description in a natural language of each rule in the rule ontology was a good guide to show what extent of details and carefulness is necessary. Also the semantic relations among the rule ontology and between the rule ontology and the domain ontologies showed what are necessary to be externalized.

Nevertheless, it was also true that the ontologies we developed, especially the rule ontology, were incomplete before the trial and error refinement using the outputs of the scheduling subsystem. The scheduling subsystem was highly evaluated by the domain experts and was successful to motivate them to maintain the ontologies. But, the reason was a little bit different from what was expected. They highly evaluated it because it could do easy but painful routines that may cause human errors, such as simple calendar checking. It still needs to be demonstrated whether domain experts keep motivated and can maintain whole ontologies for a long time.

All the results are based on the ontologies that we initially developed. In the case that domain experts have to develop ontologies and a rule-based system from scratch, it needs to be investigated what kind of support is necessary so that they can do that. It may, in particular, be tough for domain experts to develop a rule-based system that helps novices internalize the externalized intelligence skill.

4.2 Internalization by Novices

Unfortunately, in this case study, we were the only novices that could evaluate our proposal. We, as novices, could well internalize the intelligence skill of the job, using the ontologies and the scheduling subsystem. Referring to the facility ontology, we understood the exact meanings of the rules. The semantic relations among the rules such as "justify" helped us understand where each rule stands among the whole rules. The comparison of the outputs of the scheduling subsystem and what we did manually helped us internalize the intelligence skill so that we could do the job by ourselves. However, since we ourselves developed the ontologies, under the direct supervision of the domain experts, it might be necessary to be evaluated by novices who do not have direct supervision of the domain experts.

5 Summary and Future Work

In this paper, we have proposed a domain ontology, a rule ontology and a rule-based system that enables organizational transfer of intelligence skill without OJT. Intelligence skill is decomposed into pieces of knowledge and they are externalized as rules and technical terms that constitute them. They are combined as a rule ontology and a domain ontology respectively. A rule ontology consists of rules as primitives and semantic relations among them such as "justify", "depend on" etc. The shallowest rules are translated into a rule base so that they are executed on a rule-based system. The rule-based system motivates domain experts to externalize and combine their tacit intelligence skill to the ontologies and also help domain experts refine them and novices internalize them. The case study showed the effectiveness of our proposal and the possibility that it can replace costly OJT.

In the next step, we will do a more comprehensive and long-term case study using GEN, to examine whether

- domain experts can develop ontologies and a rule-based system by themselves,
- domain experts can maintain them for longer term,
- novices can internalize intelligence skill using them without OJT.

In addition, once ontologies and a rule-based system are developed in GEN, then GEN becomes a "ba" where the intelligence skill is organizationally created and accumulated. Hence, we plan to expand GEN to provide a "ba" where domain experts and even novices collaboratively externalize their tacit intelligence skill and refine them by communicating each other.

Acknowledgement

The authors wish to thank Miho Kato and Masahiro Ohira for their contribution.

References

1. Liu, C.-Y., Kume, Y.: Skill Transfer from Expert to Novice – Instruction Manuals Made by Means of Groupware. In: Proceedings, Part II of Human Interface and the Management of Information. Interacting in Information Environments, Symposium on Human Interface 2007, Beijing, China, pp. 423–429 (2007)

2. Peherstorfer, T., Schmiedinger, B.: Structured Knowledge Transfer in Small and Medium Sized Enterprises. In: Reimer, U., Karagiannis, D. (eds.) PAKM 2006. LNCS (LNAI), vol. 4333, pp. 234–242. Springer, Heidelberg (2006)
3. Watanuki, K., Kojima, K.: New Approach to Handing Down of Implicit Knowledge by Analytic Simulation. Journal of Advanced Mechanical design, Systems, and Manufacturing 1(3), 48–57 (2007)
4. Chuma, H.: Problem Finding & Solving Skill in Manufacturing Factories. The Japanese Journal of Labor Studies 510 (2002) (in Japanese)
5. Nonaka, I., Takeuchi, H.: The Knowledge-Creating Company: How Japanese Companies Create the Dynamics of Innovation. Oxford University Press, Oxford (1995)
6. Nonaka, I., Konnno, N.: The Concept of ba: Building a foundation for knowledge creation. California Management Review 40(3), 40–54 (1998)
7. Hijikata, Y., Takenaka, T., Kusumura, Y., Nishida, S.: Interactive knowledge externalization and combination for SECI model. In: Proceedings of the 4th international conference on Knowledge capture, Whistler BC Canada, pp. 151–158 (2007)
8. Gruber, T.R.: A Translation Approach to Portable Ontology Specifications. Knowledge Acquisition 5(2), 199–220 (1993)
9. Yamaguchi, T., Mizoguichi, R., Taoka, N., Kodaka, H., Nomura, Y., Kakusho, O.: Explanation Facilities for Novice Users Based on Deep Knowledge. The Transactions of the Institute of Electronics, Information and Communication Engineers D J70-D(11), 2083–2088 (1987) (in Japanese)
10. Iwama, T., Tachibana, H., Yamazaki, H., Okabe, M., Kurokawa, T., Kobayashi, K., Kato, M., Yoshioka, A., Yamaguchi, T.: A consideration of ontology that supports organized business knowledge accumulation and retrieval. In: 3rd Annual Conference of Information Systems Society of Japan, Niigata Japan (2007) (in Japanese)
11. Protégé, `http://protege.stanford.edu/`
12. Yoshioka, A., Ohira, M., Iijima, T., Yamaguchi, T., Yamazaki, H., Yanagisawa, M., Okabe, M.: Supporting Knowledge Transfer and Scheduling task with Ontologies. In: The 21st Annual Conference of the Japanese Society for Artificial Intelligence, Miyazaki Japan (2007) (in Japanese)
13. Kobayashi, K., Kato, M., Yoshioka, A., Yamaguchi, T.: Constructing Business Rules and Domain Ontologies to Support Externalization of Business Knowledge in Knowledge Transfer. In: The 22nd Annual Conference of the Japanese Society for Artificial Intelligence, Asahikawa Japan (2008) (in Japanese)

Ontology Based Object Categorization for Robots

Benjamin Johnston[1,*], Fangkai Yang[2], Rogan Mendoza[1], Xiaoping Chen[2], and Mary-Anne Williams[1]

[1] University of Technology, Sydney, Australia
[2] University of Science and Technology of China, Hefei, China
johnston@it.uts.edu.au

Abstract. Meaningfully managing the relationship between representations and the entities they represent remains a challenge in robotics known as grounding. Useful insights can be found by approaching robotic systems development specifically with the grounding and symbol grounding problem in mind. In this paper we use Semantic Web technologies to provide a powerful extension to existing proposals for grounding robotic systems and have consequently developed OBOC, the first robotic software system with an ontology-based vision subsystem. OBOC has been tested and evaluated in the robot soccer domain against an internationally recognized RoboCup system. The architecture can be extended from not just vision but to also enhance decision making and support tasks such as information retrieval, reasoning, planning, communication and collaboration. The interoperability and openness of the Semantic Web technologies that underlie this approach increase the ability for robots to introspect, communicate and be inspected—benefits that ultimately lead to more grounded systems with open-ended intelligent behaviour and scope for collaboration.

Keywords: Symbol grounding problem, ontologies, robot soccer.

1 Introduction

The symbol grounding problem (Coradeschi and Saffiotti 2003; Kent 1978; Williams *et al.* 2005; Ziemke 1997) is a prevalent topic in artificial intelligence and software engineering relevant to problems in knowledge representation, psychology and robotics. The problem is that of making the interpretation of a formal symbolic system intrinsic to the system itself, rather than depending on the interpretation in the minds of the human designers of the system (Harnad 1990).

The symbol grounding problem is related to the more recent and general concept of groundedness. Groundedness is the degree to which the entities of a system's representations correspond *meaningfully* to the entities that they represent (Williams *et al.* 2005). Grounding is a rich capability and therefore the groundedness of a system is a multidimensional and graded property: different groundings of representations can afford many different degrees of expressiveness, relevance, faithfulness, correctness, accuracy/precision, robustness, adaptability, timeliness, efficiency, self-awareness,

* Corresponding author.

T. Yamaguchi (Ed.): PAKM 2008, LNAI 5345, pp. 219–231, 2008.

awareness of others, functionality, transparency, testability and uncertainty management. Given these many potential dimensions, there is often no clear answer as to whether the entities in one system are definitively more meaningfully connected to reality than another, however it is possible to draw conclusions and make useful comparisons about these individual characteristics of a system's groundedness. For example, while human vision exceeds robotic systems with respect to robustness to visual noise and lighting conditions, robotic solutions may display greater faithfulness in the face of optical illusions.

One view of the development of robotic systems is, as a process, principally concerned with grounding: creating meaningful representations, maintaining those representations and acting upon them in an appropriate and timely manner. In theory, robots need not make use of any form of high-level representation—they could, for example, be purely reactive—however, many practical robots in use and under development today make some use of symbolic high-level representations. While these symbols may be 'second-class' entities that have been created on an ad-hoc basis and may be difficult to inspect, they are nevertheless abstract entities intended to refer to external entities or concepts. Groundedness with respect to these symbols is then the problem of how the symbols meaningfully correspond to real-life entities and concepts: the symbol grounding problem.

If the symbols in robotic systems are promoted to 'first-class' entities with independent existence and meaning, it then becomes possible to relate these symbols to concepts in a formal ontology. Doing so can enrich both the robotic system and the ontology. A robotic system enriched with a formal ontology is more grounded—the ontology can allow for greater adaptability, self-awareness, awareness of others, functionality, transparency and testability. Conversely, an ontology enriched with concepts of a robotic system has symbols whose meaning have been ground to reality via the robot: the robotic system fixes the meaning of the ontology, helps identify problematic constructions and can even be used as a reasoning mechanism in itself (a difficult problem may, in fact, be best solved by experimentation or exploration in the real world).

In this paper we introduce OBOC (Ontology Based Object Categorization), a combination of Semantic Web technologies and robotic systems that has been implemented on a Sony four-legged AIBO robot. While OBOC is narrowly focussed on the problem of recognition of objects and communication, related work by Vogt (2003)and by Gärdenfors and Williams (2003) has shown that categorisation is an essential process in understanding and constructing grounding capabilities. Categorisation is a crucial capability for robots because it allows them to deal with and efficiently store, reason with and communicate complex information captured by their internal and external senses. In OBOC, Semantic Web technologies enable categorization, communication and reasoning by providing standard protocols and languages for defining and sharing ontologies (using the Ontology Web Language, OWL). The result, a system combining the robot's physical and sensory capabilities with high-performance reasoning capabilities of the Racer inference engine, is a vast improvement over closed robotic systems that are unable to rapidly adapt to novel situations. And while our experiments with OBOC have, to date, focused on categorization and communication; there is no reason the framework would be unsuitable for extensions such as sharing actions and intentions.

2 Grounding, Symbol Grounding, Anchoring and Categorization

Grounding concerns managing the relationship between representations and the entities they represent in a *meaningful* fashion. This relationship is important because it affects the way an intelligent system can potentially behave, how it can interact with its environment, and what it is capable of achieving (Williams *et al.* 2005). Related to grounding are sub-problems of symbol grounding, anchoring and categorization. Conceptualizing these sub-problems as restrictions on the general problem of grounding creates a clearer understanding of the issues.

It is important to realize that groundedness is a multidimensional process or capability: it is not possible to merely claim that a system's representations *meaningfully* correspond to external entities, for the definition of *meaningful* is a complex and loaded term. By conceptualizing groundedness as a multidimensional and graded property, we can analyse different capabilities of systems and thereby compare different systems even though it may not be meaningful to claim that a system is objectively grounded or that one system is objectively more grounded than another entirely different system. Williams *et al.* (2005) offer a framework for analysing groundedness; in their framework, the groundedness of a system is analysed in terms of a set of gradings along multiple dimensions. While they make no claims that their analysis is exhaustive, they offer sixteen dimensions: expressiveness, relevance, faithfulness, correctness, accuracy/precision, robustness, adaptability, timeliness, efficiency, self-awareness, awareness of others, functionality, transparency, testability and uncertainty management.

The Symbol Grounding Problem (Harnad 1990) is, in fact, a restriction on the general grounding problem (or conversely, grounding is a generalization of symbol grounding). Where grounding is concerned with the relationship between a system's representations and reality, the symbol grounding problem is concerned only with those systems whose high-level representations form a *symbolic system*. A symbolic system is a set of arbitrary tokens that are manipulated on the basis of explicit rules (that are also defined in terms of arbitrary tokens), and subject to a semantic interpretation (Harnad 1990). Intuitively, a symbolic system may be seen as systems that are implemented using high-level programming languages (C, C++, Java, Python, Lisp, etc.) or that resemble the logic-based approaches of "Good Old Fashioned AI" (GOFAI).

In fact, symbol grounding relates to work that dates as far back as the ancient Greeks in their search for "true reality" and their study of metaphysics. More recently, prominent work by Kent (1978), Searle (1980) and Harnad (1990) have raised the challenge in the context of AI. Searle (1980) introduced the well known Chinese Room thought-experiment. Searle uses the Chinese Room to argue that computer programs are syntactic and lack the semantics that allow the computer to *understand*. Since no amount of syntax will ever produce semantics, Searle concluded that a purely symbolic system will never be able to understand what it is doing because of the lack of intentionality i.e., the inability to link internal representations to external objects or states (Ziemke 1997). Harnad (1990) later identified a core challenge of the argument, and posed the question of how 'the semantic interpretation of a formal symbol system be made intrinsic to the system, rather than just parasitic on the meanings in our heads?' In other words, how can arbitrary symbols be grounded and given

meaning? This is the challenge underlying the symbol grounding problem and the more general problem of grounding representations.

Harnad's (1990) own response to the symbol grounding problem was to propose combining connectionism (sub-symbolic) methods with symbolism (symbolic systems) into a hybrid model. Harnad uses connectionism as a way of representing icons and categories—internal analog transformations of sensory data—that 'pick out' or 'distinguish' concepts. He sees the relationship of icons and categories to sensory input as beyond the need for semantic interpretation—the icons and categories are merely causal responses to sensations. These can then be fixed to the elementary symbols of a symbolic system: when one has a sufficient set of elementary symbols grounded via an iconic connectionist network, then the rest of the concepts and symbols in a complex language can be generated by symbol composition alone. This, Harnad argues, is a solution to the symbol grounding problem.

If we, however, view symbol grounding as a multidimensional property then one recognizes that the problem isn't merely a matter of finding a 'solution'. Different systems have different degrees of groundedness—a more grounded system might have the capability to introspect upon the relationships between its icons and sensations, or it might be based on the idea of anchoring meanings to external objects rather than sensations as in work by Vogt (2003). Furthermore, different choices of elementary symbols can result in systems that are more expressive, robust, adaptable, transparent and testable, and therefore more grounded. For example, a system that can represent the categorization of a 'soccer ball' in terms of its shape, texture and purpose would be more grounded than another system that treats 'soccer ball' as an elementary symbol.

3 Symbols, Ontologies and the Semantic Web

Many 'intelligent' systems today incorporate some form of symbolic system—even if those symbols and rules are tightly embedded as the atoms and constructs of the mainstream programming languages used to implement the system. For example, a soccer playing robot might use a set of tokens to represent the position of a ball: it may have a set of rules hard-coded as C++ functions to predict the path of the ball, and move to a strategic position. In seeking to create more grounded systems, we recognise however, that such hard-coded rules restrict our ability to adapt such systems to novel situations. By exposing the symbols of a system as first order constructs in and of themselves (rather than entities in the software engineer's mind) it is possible to create more grounded systems.

Formal ontologies, particularly ontologies described using expressive formal languages, is one approach to constructing symbolic systems. Such ontologies are logical structures intended to present and independent model of reality—each symbol is defined for consistent semantics, robust reasoning and communication.

In recent years, Semantic Web efforts have resulted in the development of standard languages for expressing, reasoning with and communicating ontologies. The Web Ontology Language (OWL) is an established standard with widely available and efficient reasoners. By integrating (or implementing) the high-level symbolic systems of a robot with Semantic Web technologies, it becomes possible to make use of widely available tools, communication protocols, reasoners and best practices, thereby enhancing and improving the groundedness of the robotic system.

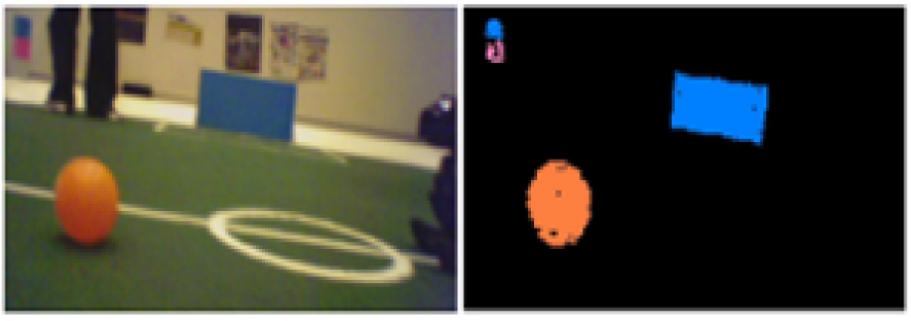


Fig. 1. Raw camera image and extracted objects in the raw image

The use of ontologies in robots is novel (Gärdenfors and Williams 2003) and few, if any, implementations and evaluations have been reported. Perhaps the most related research is Schlenoff (2002) who described the use of an ontology of obstacles to aid in path planning and obstacle avoidance. Other applications of ontologies have primarily been demonstrated in systems that are used for image and document classification (Breen, Khan and Ponnusamy 2002; Schober, Hermes and Herzog 2004; Song 2004), communication and object mapping in mobile robots (Limketkai, Liao and Fox 2005; Vogt 2003) and also object learning (Modayil and Kuipers 2004).

4 Object Recognition in Robot Soccer

RoboCup is an international robot soccer initiative designed to advance the field of Robotics. It involves an annual competition involving soccer, rescue and household robotic systems. The robot soccer domain is a complex and dynamic environment. In the 4-Legged League, Sony AIBO robots compete for possession of an orange ball in order to pass it to team-mates or shoot for goal. The playing field is 5.4m by 3.6m, made of green felt, and defined using white boundary lines and colour coded beacons. The four robots on each team are physically identical: principally interacting via their single camera, single chest-mounted range-finding sensor, joint actuators, joint-feedback sensors and wireless LAN (for robot-to-robot communication). The robots independently and autonomously process their motor-sensory data in order to construct a model of the current state of the field. A robot must be able to recognise the goals, the ball, other players, to reason about their position, and to select appropriate responses. This domain demands not merely on competent ball skills, but the fusion of sub-symbolic information into symbolic representations that guide strategy and that are shared with other robots to coordinate play.

We have targeted RoboCup as the domain to explore our ideas on object categorisation through ontologies and grounding because of the mixed demands for sophisticated object recognition, prediction, planning and communication in a dynamic environment. The UTS Unleashed! Robot Soccer System that competed in the Four-Legged RoboCup Competition in 2003 and 2004 (and has since been used for non-competitive research) was selected as the platform for development as it allowed us to focus entirely on the design and implementation of OBOC without having to

redevelop the extensive infrastructure required for fundamentals such as locomotion and vision.

The RoboCup environment has been largely crafted to assist with the AIBO's sensory capabilities. Standardized colours and lighting conditions allow for objects to be identified using straightforward colour segmentation techniques. For example, the ball is usually perceived as a large, round area of orange colour (although this can be complicated by occluding objects and shadows). The robots ground their colour concepts to sensory stimuli derived from their cameras. Colour concepts effectively form icons and of the physical objects; the colour segmentation capabilities of the robot effectively form the causal link and elementary concepts of Harnad's (1990) approach to symbol grounding. These elementary concepts can be used to build knowledge: combined into predicates that are used to describe the world and its behaviour within the robots knowledge base.

5 Applying OWL to Robot Soccer

The use of expressive ontologies in RoboCup allows for both feature-based and context-based categorisation. Furthermore, ontologies enable these concepts and categorisations to be shared with other robots. These are discussed in further detail below.

5.1 Feature Based Categorization

Using OWL properties, we can define abstract concepts (such as the ball) from the elementarily grounded concepts. For example, a concept corresponding to ***Ball*** can have the necessary property ***hasColor*** constrained to the concept ***Orange***. One can furthermore use OWL to define both sufficiency conditions: all objects with ***hasColor Orange*** are instances of ***Ball***.

In practice, three factors govern the robot's performance with feature-based categorization:

1. The quality of the elementary symbols for describing terms in the ontology and distinguishing objects in the real world (can we distinguish colours, shapes, movements?);
2. The quality of the ontology in defining objects correctly and in sufficient detail
3. The scope of the robot's application and the similarity of objects that need to be distinguished

For example, no teams in the RoboCup 4-legged league would currently be able to distinguish an orange soccer ball from a piece of ripe orange fruit: given the elementary grounding of these systems, there is no attribute that can be used to distinguish them. In practice, these recognition problems can have a significant impact on play—stray orange objects in the background can sometimes be identified as balls.

We cannot, however, expect to be able to solve these problems completely. Even as human beings, our highly sophisticated object recognition system is perceptually limited and suffers from a range of mistakes including optical illusions, hallucinations and other phenomena such as change blindness. Some uncertainty in an ontology (with respect to the elementary grounding) is inevitable—we must accept this fact and handle it gracefully.

5.2 Context Based Categorization

Often contextual information is able to distinguish similar objects. For example, a round orange object in a fruit bowl is likely to be an orange or mandarin, whereas a round orange object on a RoboCup field is more likely to be a ball. In both cases the percept of round orange thing is similar, yet the classification is very different. An OWL ontology can be used to represent such contextual information. For example, we can express contextual information about the goal box in the following way:

$$\textit{GoalBox} \sqsubseteq \exists_1 \textit{isBehindOf}.\textit{GoalKeeper} \sqcap \exists_1 \textit{isNear}.\textit{OwnBeacon}$$

Using this technique, a robot would be able to use the ontology to determine the relationships between recognised objects and be able to successfully categorise them based on context. This is similar to how some systems use ontologies for image classification (Breen, Khan and Ponnusamy 2002; Schober, Hermes and Herzog 2004) and also for the use of Relational Object Maps (Limketkai, Liao and Fox 2005).

5.3 Concept Learning and Sharing

OWL has been specifically designed for inter-operability among semantic web systems, and so is ideal for allowing robots to communicate. Ontologies allow robots to share knowledge about a single recognised object, even if they have different categorization capabilities. Ontologies not only allow two concepts to stand for the same object in respect to different groundings (Sowa 2000), but for the defined semantics and relationships to be compared, integrated and fused.

Consider the following example where Robot 1 fuses information from Robot 2:

Robot 1 (Initial):

$$\textit{Ball} \sqsubseteq \exists_1 \textit{hasShape}.\textit{Round} \sqcap \exists_1 \textit{isMovable}.\textit{True},$$
$$\textit{RoboCupBall} \sqsubseteq \textit{Ball}$$

Robot 2 (Initial):

$$\textit{PinkBall} \sqsubseteq \exists_1 \textit{hasShape}.\textit{Round} \sqcap \exists_1 \textit{isMovable}.\textit{True} \sqcap \exists_1 \textit{hasColour}.\textit{Pink}$$

Robot 1 Fused (After):

$$\textit{Ball} \sqsubseteq \exists_1 \textit{hasShape}.\textit{Round} \sqcap \exists_1 \textit{isMovable}.\textit{True},$$
$$(\textit{RoboCupBall} \sqcup \textit{PinkBall}) \sqsubseteq \textit{Ball},$$
$$\textit{PinkBall} \sqsubseteq \exists_1 \textit{hasColour}.\textit{Pink}$$

If Robot 2 recognises a round, pink object and categorises it as a ***PinkBall*** any robot that receives a message describing all the ontological features of ***PinkBall*** will be able to infer that this object is a type of Ball and respond accordingly. Obviously, a key assumption here is that overlapping symbols are jointly grounded or are defined in separate name-spaces—this is necessary, and implied in our use of ontologies that are intended to represent an 'objective' view of reality. If the meaning of ***Round*** is not known to be identical between Robot 1 and 2, then these should be given different symbols or name-spaces in the ontology: they should only be related or equated when more information is known. Note, also, that Robot 1 need not be able to perceive ***Pink***—if it is merely looking for any ***Ball*** to kick, it can still communicate with Robot

2 about the existence of the ***PinkBall*** and reason about it as a ball, even if it cannot independently classify an object as being specifically a ***PinkBall*** without the assistance of Robot 2.

6 Ontology, System Design and Implementation

The aim of our research is to implement and evaluate an ontology-based approach to categorization. The architecture, design and development of this system, OBOC, is briefly described below.

6.1 RoboCup Ontology

Using Protégé for development, a domain specific ontology was created for the 4-legged robots. Represented in this ontology are both ***ConcreteObjects*** such as goals, players, balls, regions and beacons; and ***AbstractObjects*** such as colour, shape, heading and position. The robot's beliefs and perceptions about the transient state of the soccer field are maintained and shared as assertions (ABox), while the ontology (TBox) is used and shared for recognizing and classifying those objects as specific and defined concepts.

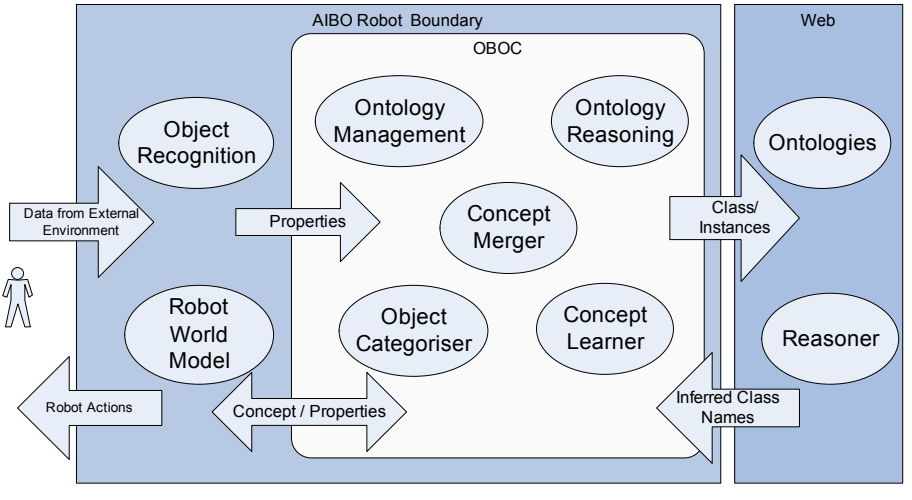


Fig. 2. System boundary model of the OBOC system

6.2 System Design

Figure 2 illustrates the foundation for the design model. The entities of the system are described below:

Object Categorizer: responsible for categorizing objects based on properties and context.

Ontology Management: manages existing classes and creates new classes in the ontology. In addition it services queries from the Concept Learner and communicates with the Ontology Reasoning entity to categorise a class based on the properties and/or context.

Concept Learner: queries the ontology for categorised concepts to infer identifiable properties and features of the object that it represents.

Concept Merger: is responsible for identifying semantic relationships between concepts.

Ontology Reasoning: performs queries on the ontology using the reasoning services of a third party component—Racer, located on a server.

6.3 Ontology Based Categorisation Tool

A proof-of-concept tool has been developed and evaluated. The Protégé API was used to access and manage the ontologies, while the Racer server was used for ontological reasoning. The tool allows users to interact with the RoboCup Soccer ontology by enabling properties to be entered via a graphical user interface and immediately respond with categorized objects. Although this tool has been developed for a specific platform, the methods and techniques could be integrated into any robot infrastructure where high level object recognition capabilities is required for communication and collaboration.

7 Results

We conducted a comparative evaluation of OBOC against an internationally competitive and typical RoboCup system called Unleashed!, with extraordinarily positive results. The grounding framework (Williams *et al.* 2005) provides an analysis tool and template for evaluating and comparing the grounding capabilities of specific systems. OBOC reaps all the benefits that ontologies offer and clearly outstrips any system based on a standard programming design whereby key object concepts reside in low level computer code (e.g., classes or files in C++ or Python).

In terms of performance, OBOC successfully identified beacons, balls and other objects via the ontology. The superiority of OBOC over Unleashed! was evident when properties of objects changed. For example, changing the color of the ball from *orange* to *pink* was straightforward in OBOC but problematic for Unleashed! because the concept *ball* is clearly defined in OBOC but its definition in Unleashed! is complex and distributed. In addition, the necessary heuristics that need to be developed to compensate for the poor sensors on the AIBO (a robot originally designed for entertainment as opposed to high performance soccer), are significantly more transparent in OBOC, when compared to Unleashed!.

Given the major role that human designers play in the grounding capabilities of current state-of-the-art robots, it is easy to appreciate that OBOC allows human developers to richly query the state of the perceived world and rapidly understand the particular consequences of ontology modification.

Despite the high level representations and logics of OWL, OBOC responds in a timely and efficient manner. The theoretical upper bounds of ontological reasoning were not a practical problem in our experiments.

While the current approach would violate of the rules of RoboCup (the use of an external reasoning server is prohibited), these results are very promising.

Furthermore, Unleashed! has no capacity to interact with other systems beyond the ones that it is hard-coded to share information with (i.e., team members), OBOC-enabled robots can, in contrast, interact with Semantic Web systems for the purpose of learning from knowledge resources, communication and collaboration with a wide range of other systems.

8 Benefits for Robots Using Ontology-Based Categorization

The OBOC system is a starting point for a holistic approach to grounding robots. In this section, we focus on the feasibility of incorporating other extensions to OBOC. We claim that it is possible for OBOC to support tasks including reasoning, planning, ontology evolution, multi-agent knowledge sharing and coordination.

8.1 Reasoning Support

OBOC already incorporates advanced reasoning capabilities as a consequence of the use of logical formalisms. OWL is based on description logics (DL)—these have well-defined model-theoretic semantics and decisions procedures that provide a sound and complete support for concept satisfiability, concept subsumption and instance checking. In fact, efficient reasoners such as FaCT, Racer or KOAN2 offer highly practical performance on standard case problems.

Nonmonotonic reasoning can simplify the expression of domain knowledge, and allows an agent to deduce intuitive results in the case of uncertainty and incompleteness using defaults. While OBOC currently does not support nonmonotonic reasoning, extensions of description logics have been proposed based on defaults (Baader and Hollunder 1995) and circumscription (Bonatti, Lutz and Wolter 2006), or may be layered with a nonmonotonic logic providing higher-level reasoning. Furthermore, integration of description logics with logic programming rules (Eiter et al. 2004; Motik and Rosati 2007; Yang and Chen 2007) can enrich the expressivity of an ontology and introduce closed world reasoning into otherwise open-world DL ontologies.

8.2 Planning

OBOC currently only supports categorization, but by extending the robot ontology to incorporate descriptions of actions, it would be possible to improve the groundedness of the robot's planning capabilities. Implementation need not be overly sophisticated: in DL-based ontologies, the existence of a plan can be reduced to reasoning on satisfiability of DL (Milicic 2007). Planning with DLs has been very important in the composition and automatic execution of Semantic Web services. These very same techniques can be applied to autonomous robots for planning their goals.

In the longer term, future research will explore the representation of even broader ranges of concepts. Ideally, it would be possible to represent not just a plan of the current robot, but also the intentions and likely plans of other robots (both team-mates and opposition). This level of sophistication would be cued by the opposition's behaviours (i.e., inferring intention from behaviour) and is necessary to anticipate behaviour and perform actions such as deliberate (as opposed to opportunistic) intercepts.

8.3 Ontology Evolution

While the ontology-based approach used in OBOC allows for rapidly evolving ontologies and sharing knowledge, in richly dynamic environments the underlying ontologies may be subject to on-line revision: humans may supply new knowledge, robots may merge knowledge bases and rules of the game may change. In classical formalisms, such changes introduce inconsistencies and the effect is explosive: any conclusion can be obtained from reasoning, and the ontology turns out to be trivial. The agent must either support nonmonotonic reasoning or, in the case of more dramatic change, revise its own ontology to allow the consistent addition of new knowledge.

There has been much recent work on inconsistency handling in DL ontologies. Although standard belief revision applies for the propositional case, generalizing these results to description logics can present a challenge. Flouris, Plexousakis and Antoniou (2005) proposes a DL version of AGM postulates that serve as rational postulates for ontology contraction and revision operators, and Qi *et al.* (2006) offer a model-theoretic version of AGM postulates. These works lay the foundation for ontology revision. Revision methods include those of Qi (2006), Meyer (2006) and Scholbach (2003). These methods allow robots to consistently revise their own knowledge in the face of new information.

Furthermore, ontology debugging and diagnosis may be useful for identifying and correcting (possibly with human assistance) the axioms that 'cause' inconsistency/incoherence (Parsia, Sirin and Kalyanpur 2005; Schlobach and Cornet 2003). Ontology debugging systems such as Swoop (Parsia, Sirin and Kalyanpur 2005) can be readily employed in OBOC for improving the testability and transparency (and therefore groundedness) of the system.

8.4 Multi-agent Communication

While OBOC currently assumes that the ontologies shared between robots are in some way consistent or readily merged, future versions will support more loosely coupled coordination through techniques such as ontology mapping (Shvaiko and Euzenat 2005) to create translations or bridge axioms (van Eijk *et al.* 2001). Our solution will have an ontology alignment protocol that can be interleaved with any other agent interaction protocol and would be triggered upon the receipt of an unrecognized message from a foreign ontology. Agents meeting each other for the first time would be able to negotiate the matching of terms in their respective ontologies and to translate the content of the message they exchange with the help of the alignment.

9 Conclusions

Building on Harnad's (1990) 'solution' to the symbol grounding problem, we have designed an architecture for constructing robot systems that is attentive to the richer understanding of grounding and groundedness. In OBOC, very simple iconic and categorical representations are causally connected to the robot's sensory subsystems—providing an elementary grounding upon which Semantic Web technologies are applied. The result is a richer, more flexible, more adaptive and therefore more grounded robotic system. While OBOC is currently limited to object categorization on the basis of features and context, the OBOC architecture and the ontology-based reasoning systems can be readily extended with greater capabilities.

There is much scope for future work. Outside the robot soccer field, we obviously do not have the ability to mark or colour code every object. Of course, the key point in OBOC is that the grounding still takes place via the sensors, even if those sensors are in-fact unable to distinguish or identify complex objects outside of the soccer field. Aside from improving the reasoning capabilities of the system, there is also therefore much scope for improving the richness of the elementary grounded symbols of the system. The long term objective is to gradually enrich this elementary grounding, improve the reasoning capabilities, tool-set and detail of the system's ontology—that is, to improve the system's grounding—so that it can respond to new objects and learn from other systems in unknown situations outside of the crafted soccer field.

References

Baader, F., Hollunder, B.: Embedding defaults into terminological knowledge represen¬tation formalisms. J. Autom. Reasoning 14(1), 149–180 (1995)

Bonatti, P., Lutz, C., Wolter, F.: Description logics with circumscription. In: Proc. of KR 2006 (2006)

Breen, C., Khan, L., Ponnusamy, A.: Image classification using neural networks and on¬tologies. In: The 13th Int. Wksp. on Database and Expert Systems Applications, pp. 98–102 (2002)

Coradeschi, S., Saffiotti, A.: Special issue on perceptual anchoring symbols to sensor data in single and multiple robot systems. Robotics and Autonomous Systems 43(2-3) (2003)

Eiter, T., Lukasiewicz, T., Schindlauer, R., Tompits, H.: Combining answer set program¬ming with description logics for the semantic web. In: Proc. of KR 2004 (2004)

Flouris, G., Plexousakis, D., Antoniou, G.: On applying the AGM theory to DLs and OWL. In: Gil, Y., Motta, E., Benjamins, V.R., Musen, M.A. (eds.) ISWC 2005. LNCS, vol. 3729, pp. 216–231. Springer, Heidelberg (2005)

Gärdenfors, P., Williams, M.-A.: Building rich and grounded robot world models from sensors and knowledge resources: a conceptual spaces approach. In: Proc. of the Int. Symp. on Autonomous Mini-robots for Research and Edutainment (2003)

Harnad, S.: The symbol grounding problem. Physica D: Nonlinear Phenomena 42, 335–346 (1990)

Kent, W.: Data and Reality. North Holland, Amsterdam (1978)

Limketkai, B., Liao, L., Fox, D.: Relational object maps for mobile robots. In: Proc. of the IJCAI 2005, pp. 1471–1476 (2005)

Meyer, T., Lee, K., Booth, R.: Knowledge integration for description logics. In: Proc. of AAAI 2005 (2006)

Milicic, M.: Planning in action formalisms based on DLs: first results. In: Proc. of DL 2007 (2007)

Modayil, J., Kuipers, B.: Bootstrap learning for object discovery. In: IEEE/RSJ Int. Conf. on Intelligent Robots and Systems, pp. 742–747 (2004)

Motik, B., Rosati, R.: A faithful integration of description logics with logic program¬ming. In: Proc. of IJCAI 2007 (2007)

Parsia, B., Sirin, E., Kalyanpur, A.: Debugging OWL ontologies. In: Proc. of WWW 2005 (2005)

Qi, G., Liu, W., Bell, D.A.: Knowledge base revision in description logics. In: Fisher, M., van der Hoek, W., Konev, B., Lisitsa, A. (eds.) JELIA 2006. LNCS (LNAI), vol. 4160, pp. 386–398. Springer, Heidelberg (2006)

Rosati, R.: DL+log: tight integration of description logics and disjunctive datalog. In: Proc. of KR 2006 (2006)

Schlenoff, C.: Linking sensed images to an ontology of obstacles to aid in autonomous driving. In: Proc. of the 18th Nat. Conf. on Artificial Intelligence: Wksp. on Ontologies for the Semantic Web (2002)

Schlobach, S., Cornet, R.: Non-standard reasoning services for the debugging of de¬scription logic terminologies. In: Proc. of IJCAI 2003 (2003)

Schober, J.P., Hermes, T., Herzog, O.: Content-based image retrieval by ontology-based object recognition. In: Proc. of the KI-2004 Wksp. on Applications of Description Logics (ADL 2004) (2004)

Searle, J.R.: Minds, brains and programs. Behavioural and Brain Sciences 3(3), 417–457 (1980)

Shvaiko, P., Euzenat, J.: A survey of schema-based matching approaches. J. of Data Semantics (JoDS), IV (2005)

Song, M.H.: Ontology based automatic classification of web pages. Int. J. of Lateral Computing 1(1) (2004)

Sowa, J.F.: Ontology, metadata, and semiotics, viewed 2/8/05 (2000), `http://www.isye.gatech.edu/faculty/Leon_McGinnis/8851/Sources/Ontology/SemioticsSowa.htm`

van Eijk, R., de Boer, F., van der Hoek, W., Meyer, J.: On dynamically generated ontology translators in agent communication. Int. J. of Intelligent Systems 16, 587–607 (2001)

Vogt, P.: Anchoring of semiotic symbols. Robotics and Autonomous Systems 43(2), 109–120 (2003)

Williams, M.-A., Gärdenfors, P., Karol, A., McCarthy, J., Stanton, C.: A framework for evaluating groundedness of representations in systems: from brains in vats to mobile ro¬bots. In: IJCAI 2005 Workshop on Agents in Real-Time and Dynamic Environments (2005)

Yang, F., Chen, X.: DLclog: a hybrid system integrating rules and description logics with circumscription. In: Proc. of DL 2007 (2007)

Ziemke, T.: Rethinking grounding', Austrian Soc. for Cognitive Science. In: Proc. of New Trends in Cognitive Science (1997)

Ontology-Based Expertise Finding

Maryam Fazel-Zarandi[1] and Eric Yu[2]

[1] Department of Computer Science, University of Toronto, Canada
mfazel@cs.toronto.edu
[2] Faculty of Information Sciences, University of Toronto, Canada
yu@fis.utoronto.ca

Abstract. To accomplish knowledge intensive tasks, people in organizations must be able to find the knowledge or information needed to solve complex problems. For this, people often rely on their past experiences, explicit documents, and others who have the needed expertise. Knowledge Management Systems that enhance and facilitate the process of finding the right expert in an organization have gained much attention in recent years. This paper explores the potential benefits and challenges of using ontologies for improving existing systems. A modeling technique from requirements engineering is used to evaluate the proposed system and analyze the impact it would have on the goals of the stakeholders. This paper also discusses the organizational settings required for the successful deployment of the system in practice.

Keywords: Expertise Finding Systems, Ontologies, Agent-Oriented Modeling.

1 Introduction

To accomplish knowledge intensive tasks, people in organizations must be able to find the knowledge or information needed to solve complex problems. In addition to relying on their own memory, past experiences, and information and knowledge stored in impersonal sources such as databases, people usually rely heavily on others with knowledge and information [1]. Finding an expert to acquire information or to learn from, however, might not be an easy task for many reasons. Expertise is highly dynamic [15], difficult to qualify [9], and varying in level [9]. It is also difficult to validate other people's expertise [15] and to distinguish a good expert from a bad one. Furthermore, due to the complexity of some problems, the assistance of multiple experts may be required [9]. The difficulty of locating experts increases in larger and more geographically distributed organizations [16].

In order to augment and assist the process of locating expertise within an organization, the study and development of special Knowledge Management Systems that suggest people who have some expertise in a given area, has received the attention of both researchers and organizations. The resulting recommender systems typically rely on individuals to provide accurate and comprehensive profiles of their competences and experiences [2], but some also use mechanisms to automatically discover up-to-date expertise information from secondary sources such as documents and forums [16]. In a review of these systems in [16], however, problems related to heterogeneous information sources, expertise analysis support, and interoperability were identified.

T. Yamaguchi (Ed.): PAKM 2008, LNAI 5345, pp. 232–243, 2008.

The common solution to the problems related to heterogeneous information sources and interoperability is to formally specify the meaning of the terminology of each system and to define a mapping between these terminologies. In other words, using ontologies to provide a shared common understanding of the structure of information among systems and software agents [11]. In addition, because of their powerful knowledge representation formalism and associated inference mechanisms [12], ontologies can also be incorporated to address problems related to expertise analysis support.

Taking these facts into account, it would seem natural to expect that Expertise Finding Systems (EFS) can benefit from the use of ontologies. However, there are various potential difficulties and challenges associated with the use of ontologies that may cause the system to fail. In this paper, we are interested in investigating the circumstances under which an ontology-based EFS might or might not work. More specifically, we want to systematically explore and analyze how ontologies might be used in EFS, before creating a prototype and conducting case-studies. For this, we use a modeling technique from requirements engineering to evaluate the proposed ontology-based EFS and analyze the impact that it would have on the goals of the stakeholders.

The organization of this paper is as follows: Section 2 briefly presents the role that ontologies can play in Expertise Finding Systems. Section 3 is concerned with knowledge management analysis and discusses knowledge processes and knowledge markets between individuals and presents an analysis of the technology supported by goal models developed using the i* notation. The required organizational settings for a successful deployment of the technology are also elaborated on in this section. Finally, Section 4 concludes the paper.

2 Role of Ontologies

An ontology is a formal description of a set of objects, concepts, and other entities that are assumed to exist in a domain of interest along with their properties and relationships that hold among them [4] and the constraints that exist over them. Ontologies provide a shared common understanding of information among people or software agents [3] and enable the reuse of domain knowledge [5]. They can be used to capture and represent data semantics and by assuming deductive capability as provided by an inference engine, ontologies provide the means for deduction and automated reasoning in order to generate further knowledge [3] (i.e., knowledge that is not explicitly known but that can be deducted based on the general knowledge of the domain). In addition, ontologies can also be used for information integration, i.e., the merging of information from different sources despite differing conceptual and contextual representations [11].

In the context of expertise finding, different types of ontologies may prove to be beneficial for different purposes. A domain (expertise) ontology which would capture and represent the terminology and concepts used in the domain is of central importance. For example, EFS use different evidences as indicators of expertise [16], some of which include self declarations of expertise and artifacts created by experts [9]. A domain ontology can be used to automatically annotate existing information resources and to perform automated reasoning to improve the indicator detection and extraction mechanisms. Also, since sources can possibly be physically distributed across the

organization and stored in different formats, domain ontologies can be used to automate information integration. Furthermore, for self declarations of expertise, domain ontologies can be used to structure the different characteristics of users and their relationships, as well as help the users in specifying their goals and competences.

Another useful ontology is the organization ontology [3] which formalizes the organizational structure and can be used to infer expertise based on the roles that the agents play and the communications that occur among them. The knowledge provenance and trust ontologies presented in [8] are other examples of ontologies which can improve expertise finding. These ontologies can be used to formally define the semantics of information sources, information dependencies, relationships between information sources and experts, and trust relationships to improve expertise recognition and extraction.

In order to have a successful system, it is essential to take into account the potential difficulties and challenges in using formal ontologies for EFS. Developing a domain ontology that is agreed upon by the members of the organization is often a difficult task (for example, a term may have several widely accepted definitions or none at all). Achieving interoperability between different systems and integrating information from different sources are other highly challenging issues. Incompatibilities may arise due to different vocabularies and differences in the expressiveness power of the ontologies (or simply terminologies) used for different systems. This is usually rooted in the fact that different engineering teams and domain experts are involved in creating different systems. These issues are the focus of many on-going studies in the ontology engineering literature.

Once the required ontologies are developed and the desired level of interoperability is achieved, it is important to acknowledge that the ontologies should be incorporated in the daily activities and used for performing workflows. Otherwise, they will decay over time and would not be able to keep up with the dynamic nature of organizations. Specifications for ontologies often need to be changed to reflect changes in the real world; therefore, ontologies have to be maintained and modified frequently, and the maintenance process needs to be viewed as an organizational process [13]. A group of knowledge engineers should be responsible for ontology maintenance, and a set of rules for making changes to the ontologies must be present. When making changes to an ontology, ontology versioning must be taken seriously, and the impact of the changes to the overall architecture must be considered. Otherwise, changes may result in incompatibilities between different system components, and also may change the semantics of the data. Having to rely on a group of knowledge engineers for maintenance is by itself another potential difficulty.

Keeping the benefits and difficulties of using ontologies for improving EFS in mind, the question that is now raised is that under what circumstances would an ontology-based expertise finding system actually work and be effective in practice? One way to go about answering this question is to construct a prototype and conduct case-studies. However, we are more interested to see if it is possible to find a solution to this question by conducting a systematic analysis of the expertise finding problem. The final goal of such an analysis would be the determination of a set of steps for systematically guiding the design of EFS. The next section presents our attempts at answering this question using an agent-oriented modeling method.

3 KM Analysis Using Agent-Oriented Modeling

The knowledge management analysis in this section will help in better understanding the role that an ontology-based EFS can play in an organization. Some of the questions we intend to address include: Under which conditions can the ontology-based EFS fail?, and How does failure affect the goals of corresponding stakeholders?

To answer these questions, an agent-oriented modeling approach is used, in which the EFS can be regarded to represent an "intentional actor" [14] – *intentional* because it pursues assigned goals, and *actor* because it can exhibit active behavior to a certain extent. The benefits of using this approach are [14]: 1) Making intentions of EFS explicit aids in reasoning and arguing about it; 2) Reasoning about the EFS allows for the evaluation of different degrees of goal satisfaction among different actors, thus takes situational context of knowledge transfer into account; and 3) By making relations between different actors' goals and the EFS explicit, the *how* and *why* the EFS works or fails can be made visible.

For the construction of agent-oriented models, the i* framework [17] was chosen in this paper. This framework allows for clear and simple representation of actors' goals and dependencies among them by means of *strategic dependency (SD)* and *strategic rationale (SR)* models. In addition, i* also gives the ability to reason about modeled goals by means of goal evaluation algorithms [6]. In i*, actors are represented as agents, roles or positions and the framework has the ability to model common concepts such as goals, softgoals, tasks and resources. SD models consist of a set of nodes representing actors and depict goal, task, resource and softgoal dependencies between them. SR models contain goals, tasks, resources and softgoals of specific actors that are related to each other through task-decomposition and means-ends links. Fig. 1 shows some of the elements of the i* framework and their corresponding graphical representations. See [17] for a more comprehensive background information about the i* framework.

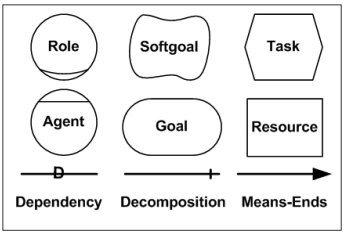


Fig. 1. Selected elements of the i* framework

Let us illustrate a simple example. Fig. 2 depicts a simplified SR model of the high-level goals, processes, and intentional dependencies of the roles of expertise seeker and provider. As can be seen in the figure, the expertise provider has the top level goals of Keep the job and Get promotion. In order to achieve the former goal s/he needs to Finish his/her own required functions on time. Satisfying management can also have a positive impact on both of the top level goals and to achieve this goal the expertise provider needs to Help expertise seekers in addition to Finish his/her own required

functions. The expertise seeker, on the other hand, depends on the expertise provider to provide him/her with the required information or knowledge and/or teach him/her the required skills needed for completing a project. Note that the means-ends link is used to indicate that the goal of Help expertise seeker can be achieved by Provide information or Teach expertise seeker.

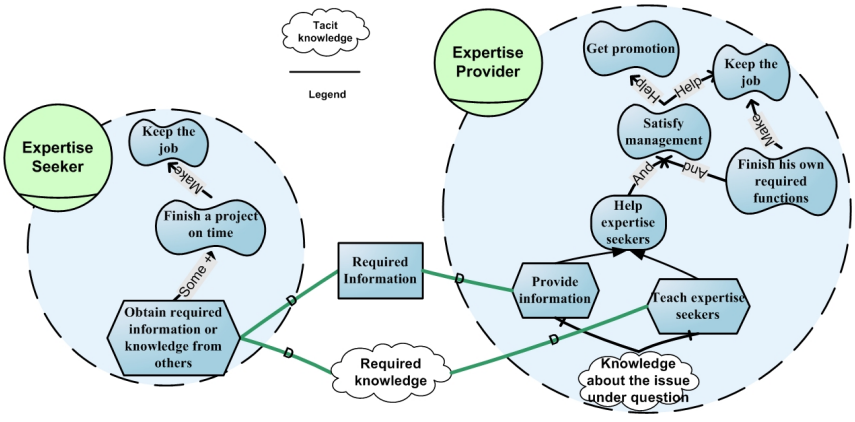


Fig. 2. Simplified model of intentional relationships between expertise seeker and provider

In the following subsections, we first analyze the expertise finding problem in a general setting using the i* modeling notation. Then, we add the EFS to the model and analyze the impact it would have on the goals of the stakeholders. Finally, we look at the development and maintenance of the ontologies for EFS and analyze the interactions of ontology engineers with other stakeholders.

3.1 Expertise Finding without EFS

To model and analyze the expertise finding problem we make use of existing studies on information and expertise seeking. In [10], McDonald and Ackerman report that participants in a medium-sized software firm use complex, iterative behaviors to minimize the number of possible expertise sources, while at the same time, provide a high possibility of garnering the necessary expertise [10]. They distinguished two steps in finding expertise within organizations: 1) Expertise identification: the problem of knowing what information or special skills other individuals have, and 2) Expertise selection: the problem of appropriately choosing among people with the required expertise.

Other studies also support this distinction. For example, the study by Cross and Borgatti [1] revealed that in addition to expertise seeker's awareness of a potential source's expertise, other factors such as timely access to the source, a degree of safety in the relationship, and willingness of an expertise provider to cognitively engage in problem solving, all play an important part in determining whom the expertise seeker chooses to go to. Other criteria for not selecting an expert include cultural differences, language problems, or a lack of experience in a related but necessary discipline [10].

Fig. 3 shows a SR model of the high-level goals, processes, and intentional dependencies of the roles of expertise seeker and provider and existing IT systems. For simplicity, other goals and tasks of these roles which are not related to expertise seeking and providing are omitted from the figure. For example, the Finish a project on time goal of the expertise seeker may involve finding information from explicit sources, but since this is not related to expertise finding, it is not shown in the figure.

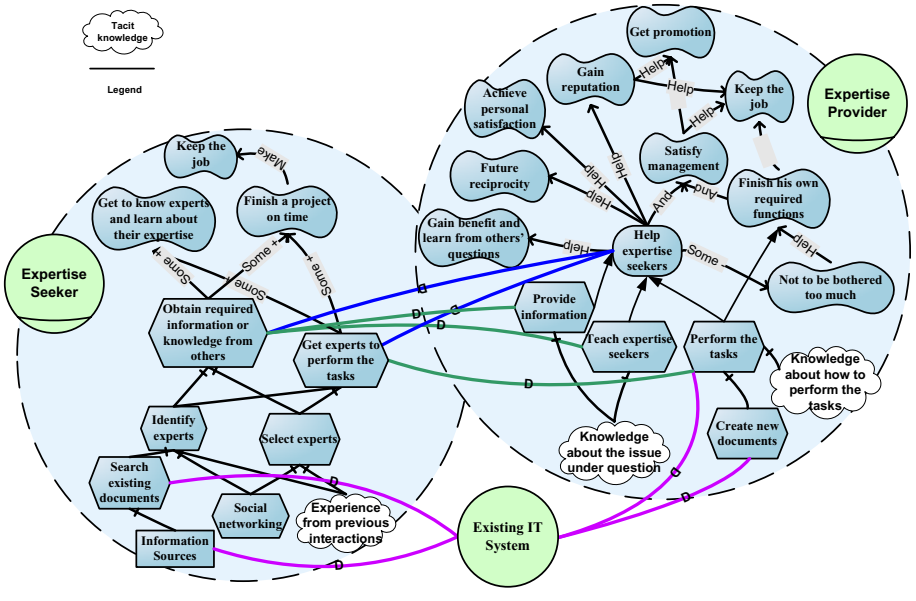


Fig. 3. Intentional relationships between expertise seeker and provider using the i* notation. For simplicity dependums are omitted.

As can be seen in the model, the Obtain required information or knowledge from others goal of the expertise seeker is decomposed into two tasks: Identifying experts and Selecting experts. The former task is further decomposed into Searching existing documents, Social networking, and using Experiences from previous interactions. The expertise selection task is performed by Social networking and using pervious experiences.

The model in Fig. 3 can also help in better understanding the knowledge market that may exist in an organization between expertise seekers and providers. The expertise seeker depends on the expert to provide him/her with adequate information or knowledge, teach necessary skills, or perform a task that the seeker is incapable of performing on his/her own. On the other hand, the expertise provider would want to Keep his/her job and Satisfy management in order to Get a promotion. Gaining reputation would also be helpful in Getting promotions, in addition to reducing the chances of being laid off and thus keeping the job. This goal can be (partially) achieved by helping expertise seekers. However, this task (Help expertise seekers) has a negative contribution to the goal of Not to be bothered too much which helps to get his/her own work done on time. Therefore, if the expertise provider is reputable enough, has a heavy

workload, or even in case of competition between members, s/he may not be motivated or willing to provide sufficient information, teach, or perform a task for the expertise seeker. Of course, there could also be personal satisfaction in helping others which could motivate the expertise provider to help the seeker. Considering the goals of the expertise provider, therefore, it is possible to see the importance of introducing proper incentives to encourage individuals to share their knowledge with others. Depending on the culture of the organization, incentives can include individual recognition, status upgrades, or even monetary prizes.

An organization may also have IT systems and technology-oriented interfaces that actors depend on for their daily tasks. These may include search and retrieval tools, storage mechanisms, and technologies for communication and collaboration such as groupware and Wiki. The technologies, on the other hand, depend on the expertise seeker and provider (amongst others) to contribute contents to the systems. Thus, human actors and IT systems can also have knowledge relationships, although the knowledge market between them would depend on the embedded knowledge and adjusted settings provided by the developers and maintainers.

3.2 EFS – Benefits and Pitfalls for Various Stakeholders

In this section, the analysis of the capabilities and functionalities of EFS is presented. The key requirements of an EFS, as pointed out in [9], typically include the ability to identify experts, classify the type and level of expertise, validate the breadth and depth of expertise, and recommend experts. Based on this and the material presented in the previous sections, the functionalities and goals of the EFS are shown in Fig. 4 without the consideration of the knowledge management needs that it can fulfill in an organization.

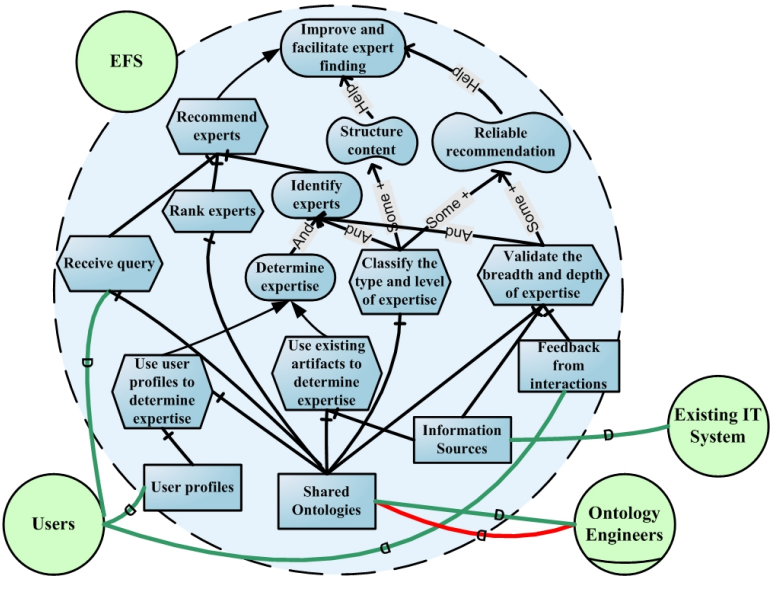


Fig. 4. Goals, capabilities, and functionalities of the EFS technology

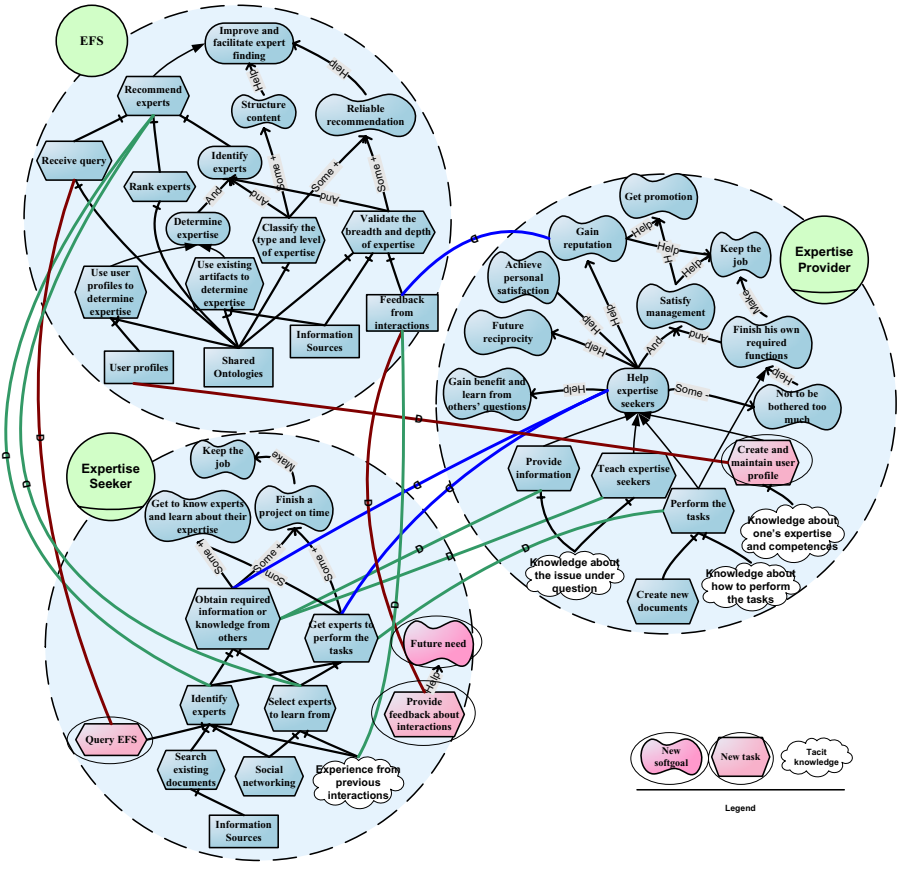


Fig. 5. Interactions of expertise seeker and provider with the EFS

The EFS should be able to interact and work with other existing IT systems in an organization in order to be more effective. For automatic detection of user expertise, the system depends on other IT systems to provide it with information sources that can be indexed and parsed to extract expertise. Such information sources may include documents generated by users, past projects individuals worked on, and contents of emails, forums, and bulletin boards.

In order to analyze the roles and knowledge processes required to make EFS successful in an organization, the interactions of the expertise seeker and provider with the EFS are illustrated in Fig. 5. This model expresses the tasks that the roles expect the EFS to do, as well as the required tasks that need to be performed by expertise seeker and provider in order to make the EFS successful.

As can be seen in Fig. 5, the introduction of the technology introduces some new tasks (shown with oval around them) for the expertise seeker and provider to do. For example, the new task of "Create and maintain user profiles" which is added to the responsibilities of the expertise provider and also "Query EFS" which is created for

the expertise seeker would not have existed before. However, the new task of "Provide feedback about interactions" may have existed before and done somewhat informally, but with the introduction of the new technology its nature changes.

Now that the EFS is conceptualized as an agent and strategic dependencies are made explicit, it is possible to do goal evaluation to see if stakeholder goals are achieved and to determine the viability of the proposed solution. For example, the goal evaluation algorithm of [6] can be used. This algorithm assigns qualitative evaluation labels to the elements of the i* model according to a six point scale that range from satisfied, partially satisfied, conflict, unknown, partially denied to denied. The algorithm starts by assigning initial evaluation values and then continues by propagating these initial values through the network of actors using a combination of guidelines and human judgment [14]. By conducting these kinds of analysis, it is possible to identify criteria for success and effects of failure of the EFS.

When, for example, the expertise provider does not perform the Create and maintain user profile task, then depending on whether s/he performs one or more of the other tasks satisfying Help expertise seekers or not[1], this may or may not result in the propagation of the negative effect to Help expertise seekers, and in return have an effect on Satisfy management and Get promotion goals. However, this initial assignment would result in denying the resource dependency User Profiles and in return result in denying the Use profiles to determine expertise task of the EFS. Now if the ontologies are properly maintained and the resource dependency Information sources is satisfied, then this would have a positive impact on Use existing artifacts to determine expertise and would stop the negative impact of the denial of Use profiles to determine expertise to propagate any further. However, if the ontology engineers fail to Maintain Ontologies, then the resource dependency Shared Ontologies is denied. This in turn would have a significant negative impact on the ability of the EFS to identify and recommend experts and would prevent it from achieving its goal Improve and facilitate expert finding.

On the other hand, if the expertise seeker fails to Provide feedback, in addition to having a negative impact on his/her own goal Future need, this action denies the resource dependency Feedback from interactions. This in turn has a negative impact on Validate the breadth and depth of expertise task of the EFS, and this negative impact propagates to the root and finally prevents the EFS from achieving its goal Improve and facilitate expert finding. In addition, the assignment of denied to the resource dependency Feedback from interactions, would also have a negative impact on the Gain reputation goal of the expertise provider. Thus, expertise seekers should be motivated to use the system and provide feedback. The ease of use, simplicity, and familiarity of the system, trust in that content is up-to-date, and the accuracy of expertise identification, are all factors that encourage the use of the system. Promoting the use of the system through the use of offline and online communications that highlight the features of the system can also be effective in encouraging individuals to use the EFS.

3.3 Role of Ontology Engineers

The shared ontologies (see Fig. 4) that are to be used to improve the EFS depend on ontology engineers to create and maintain them. This is an additional role that is

[1] The means-ends relationship will take the maximum value of its children with satisfied being the highest value and denied being the lowest value.

needed if the system is to benefit from the use of ontologies. These engineers rely on domain experts, end-users, and management to provide them with the knowledge about the domain and existing standards that are used within the organization. They may also require the specifications of exiting IT systems and their information sources in order to create ontologies that are compatible with those systems. These knowledge dependencies, along with high level goals, capabilities, and responsibilities of ontology engineers are shown in Fig. 6.

In this model we can see the important role that tacit knowledge about the domain plays in the design of high quality shared ontologies. Of course, depending on the expertise domain, the amount and quality of the tacit knowledge that is required to be transferred to ontology engineers varies. In situations where there is an ongoing need for interactions with various actors to gain the required knowledge, the development of the ontology may not be successful. Actors may not be motivated enough to transfer their knowledge, especially since what they would be gaining in return may seem vague and non-immediate. To create a standard ontology, thus, it is important that the contributing actors have a positive attitude towards the ontology.

In dynamic environments, the maintenance of ontologies may become an important concern. The changes may occur due to changes in the environment the system is operating in or changes to the requirements of the users. To have a successful and up-to-date system, the necessary changes should be applied to the ontologies frequently. Ontology engineers depend on domain experts, end-users, and management to inform them of the new requirements and possible changes in the environment. These actors should be aware of the importance of ontology maintenance and its implications for the whole system. Having a trained team of individuals in the organization can improve and facilitate the maintenance process.

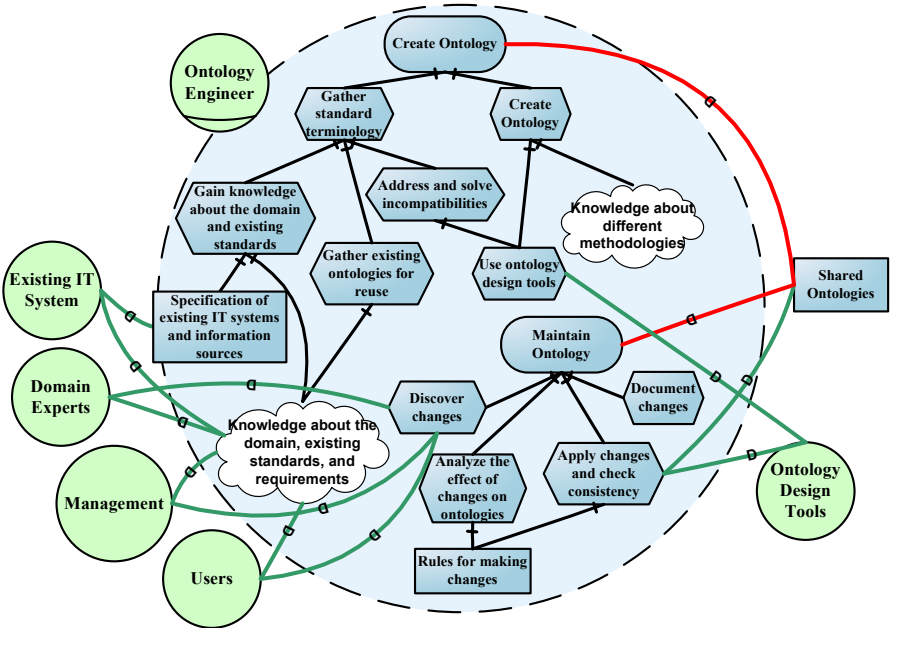


Fig. 6. Goals, capabilities, and responsibilities of Ontology Engineers

As already mentioned, designing large ontologies and insuring their consistency is an extremely complex and knowledge intensive task. Ontology engineers depend on ontology design tools for the creation and maintenance of ontologies. These tools typically use reasoners or theorem provers to provide feedback to the user about the logical implications of their design such as highlighting inconsistencies and redundancies [7]. If the ontology maintenance is to be partly supported by individuals within the organization, then it is necessary for these actors to become familiar with these tools. It is important to note that if the ontologies are designed as modular as possible – i.e., if the ontology is divided into an upper level which rarely changes and an organic lower level that can be modified and changed by individuals – then ontology maintenance becomes more manageable.

4 Conclusions

The most important contribution of this paper is the analysis of the expertise finding problem in a general setting using the i* modeling notation. This method allows for the systematic modeling of EFS and its effects on the goals of the stakeholders. Based on the findings, apart from the benefits and challenges of incorporating ontologies in systems, the social and organizational factors, such as having positive attitude towards the ontology, interactions between different actors, and having a trained team of individuals responsible for ontology maintenance, are important for the successful deployment of an EFS.

The role of the executive leadership in the application of these findings is critical. They can help motivate users to participate, establish realistic expectations, communicate progress and successes, and overcome barriers. They can encourage a culture of mutual support and knowledge sharing, and help change corporate culture from being competitive to being cooperative.

Future work will focus on the design and development of new systems based on the analysis in this study and the evaluation of these systems in real world settings.

Acknowledgements. The authors gratefully acknowledge the constructive comments and helpful advice of Professor Mark S. Fox.

References

1. Cross, R., Borgatti, S.P.: The Ties that Share: Relational Characteristics that Facilitate Information Seeking. In: Huysman, M.H., Wulf, V. (eds.) Social Capital and Information Technology, pp. 137–161. MIT Press, Cambridge (2004)
2. Earl, M.: Knowledge Management Strategies: Toward a Taxonomy. Journal of Management Information Systems 18(1), 215–233 (Summer 2001)
3. Fox, M.S., Barbuceanu, M., Gruninger, M., Lin, J.: An Organization Ontology for Enterprise Modelling. In: Prietula, M., Carley, K., Gasser, L. (eds.) Simulating Organizations: Computational Models of Institutions and Groups, pp. 131–152. AAAI/MIT Press, Menlo Park (1997)
4. Gruber, T.R.: Toward Principles for the Design of Ontologies Used for Knowledge Sharing, in KSL 93-04, Knowledge Systems Laboratory, Stanford University (1993)

5. Guarino, N.: Formal Ontology and Information Systems. In: Proceedings of FOIS 1998, Trento, Italy, pp. 3–15 (1998)
6. Horkoff, J.: Using i* Models for Evaluation, Master's Thesis, University of Toronto, Toronto, Canada (2006), http://www.cs.utoronto.ca/~jenhork/MScThesis/Thesis.pdf
7. Horrocks, I.: Semantic Web: The Story So Far. In: Proceedings of the 2007 International Cross-Disciplinary Conference on Web Accessibility (W4A), vol. 225, pp. 120–125. ACM Press, New York (2007)
8. Huang, J.: Knowledge Provenance: An Approach to Modeling and Maintaining the Evolution and Validity of Knowledge, PhD Thesis, Department of Mechanical and Industrial Engineering, University of Toronto, Canada (2008), https://tspace.library.utoronto.ca/bitstream/1807/11112/1/Huang_Jingwei_PhD_Thesis.2007-12.pdf
9. Maybury, M.T.: Expert Finding Systems, MITRE Technical Report, Bedford, MA (2006)
10. McDonald, D.W., Ackerman, M.S.: Just Talk to Me: A Field Study of Expertise Location. In: Proceedings of the 1998 ACM Conference on Computer Supported Cooperative Work (CSCW 1998), Seattle, WA (1998)
11. Noy, N., McGuiness, D.L.: Ontology Development 101: A Guide to Creating Your First Ontology, Stanford Knowledge Systems Laboratory Technical Report (2001)
12. Razmerita, L., Angehrn, A., Maedche, A.: Ontology-based User Modeling for Knowledge Management Systems. In: User Modeling 2003, Pennsylvania, pp. 213–217 (2003)
13. Staab, S., Studer, R., Schnurr, H.P., Sure, Y.: Knowledge Processes and Ontologies. IEEE Intelligent Systems (2001)
14. Strohmaier, M., Yu, E., Horkoff, J., Aranda, J., Easterbrook, S.: Analyzing Knowledge Transfer Effectiveness – An Agent-Oriented Modeling Approach. In: Proceedings of the 40th Annual Hawaii International Conference on System Sciences (HICSS 2007). IEEE Computer Society, HI (2007)
15. Vennesland, A.: Finding and Mapping Expertise Automatically Using Corporate Data, M.Sc. Thesis, Department of Computer and Information Science, Norwegian University of Science and Technology (2007)
16. Yimam-Seid, D., Kobsa, A.: Expert finding Systems for Organizations: Problem and Domain Analysis and the DEMOIR Approach. Journal of Organizational Computing and Electronic Commerce 13(1) (2003)
17. Yu, E.: Modeling Strategic Relationships for Process Reengineering, PhD thesis, Department of Computer Science, University of Toronto, Canada (1995), http://portal. acm.org/citation.cfm?id=922590

Design for Learning and Teaching: A Knowledge-Based Approach to Design Products

Mahmoud Moradi, Stéphane Brunel, Marc Zolghadri, and Bruno Vallespir

IMS-LAPS/GRAI, Bordeaux University, CNRS,
351 cours de la Libération, 33405 Talence cedex, France
{Mahmoud.Moradi,Stephane.Brunel,Marc.Zolghadri,
Bruno.Vallespir}@ims-bordeaux.fr

Abstract. In this paper, we will present some prospective ideas which should allow firms' strategic positioning of the market by using knowledge as a key strategic leverage. After presenting the three basic theories underlined design for learning and teaching, paper continues by describing the basic model based upon which we will develop our unique model; knowledge ingenition process. A generic framework is proposed containing a macro-model and a set of micro-models mapping knowledge elements and their dependencies. These models together are necessary to analyze the knowledge situation of the firm and to conceive a roadmap for future trainings of various employees of the firm during the lifecycle of a product. These concepts are illustrated through a part of a case study.

Keywords: Knowledge management, design theory, learning theory, ingenition process.

1 Introduction

The use of knowledge as an element of differentiation strategy is a quite complex challenge. "Knowledge" of a firm has several forms and contains elements from the firm's trade, structure, culture, and environment. The differentiation strategy offers obviously broad field for business competition and knowledge provides a fertile environment of differentiation. Knowledge must be understood as a vital source of competitive advantage. Following Nonaka *et al.* [1], we think that knowledge is continuously created in a dynamic system resulting from interactions amongst individuals and organizations in a specific context.

According to the OECD [2], the worldwide expenditure in educational systems in the next decade represents something about 2000b$. But products are often designed and industrialized without using a scientific approach toward learning and teaching dimensions. The whole knowledge, generated, stored and re-used in any firm, comes from its activities aiming at answering better and better final customers' needs. These needs should be collected, understood (more or less precisely) and translated into usable constraints for design and development team. In a "classical" product development project (*i.e.* usage-oriented products), these aspects are quite well understood. We call these projects *design for use*, because the main purpose of the product is to be used by customers. Laptops, cars and cell phones are all usage-oriented products. Nevertheless,

T. Yamaguchi (Ed.): PAKM 2008, LNAI 5345, pp. 244–255, 2008.

the research field does not pay much attention to learning/teaching-oriented products (the Sony™ Aibo robot or Lego for instance). As far as we know, there are neither methods nor tools in order to support deeply processes needed to understand direct and indirect customers needs for these products. Moreover, the way that these needs should be used by the firm as a design and development framework is not studied. In other words, the design-for-teaching/learning paradigm is not still developed. Our research addresses precisely this subject.

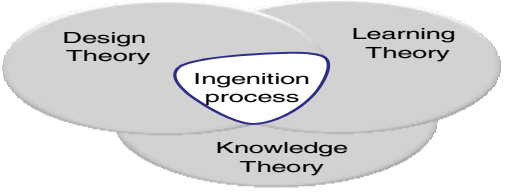


Fig. 1. Underline theories

Before giving more details, it is necessary to notice that two categories of customers are distinguished in learning and training contexts. *Direct* customers (an instructor for example) are those customers who use a product, so a learning/teaching-oriented product, to teach something. *Indirect* customers form the audience (students for instance). Therefore, the design of the learning/teaching –oriented products should take account of the needs of direct and indirect customers simultaneously. Our work stands at the intersection of three basic theories: knowledge theories, design theories and learning theories (see fig. 1). Based on these theories, we will propose an integrated framework to analyze the knowledge generation process under an ingenition model.

In this paper, we propose a new model, so-called *ingenition process. It* considers and then conceives Knowledge generation process based on the ultimate goal of any learning and teaching-oriented product. The rest of this paper is organized as follows. Section two gives a brief overview of the related works in design theories. In the third part we will develop a value chain for knowledge creation process based on semantic transformation. The paper continues by reviewing some literature in learning theories and positioning our work at the heart of these theories. The main concepts of our approach, incorporated in the ingenition process, are presented in section five. We illustrate micro-models through an example taken from the mechanical education. Some conclusions and perspectives will end the paper.

2 Design Theories in Relation to Knowledge

Almost always customers' requirements are defined in terms of usage of the target product. This is the very first set of needs of customers. That is the reason why design theories are mainly focused on usage-oriented services or products. Somehow, design is a process that covers various necessary steps going from the identification of market needs till the realization of the product.

Tollenaere [3] shows that it is necessary to model data and knowledge related to the product from the beginning of the design process. Grabowsky's approach [4] positions

the problem into the product lifecycle. Four modeling layers are then necessary: requirements modeling layer, functional layer, physical principles modeling layer and forms modeling layer. Umeda's «Function - Behavior - State» model [5] and the Shimomura's «Function - Evolution - Process» model [6] have similar characteristics by defining the designers' job according to three sequential steps. Andreasen's proposition [7] is focused on knowledge structuring of any product according to four fields, corresponding to the four sequential activities of design: physical phenomena, functions, organs and parts/items.

At the heart of these models, the Design Structure Matrix, (DSM), associated with a product description module is cornerstone of our work. DSM structures the product development phase by splitting it into several problems to solve. This matrix allows keeping track of past paths of design. Fagerstrom [8] uses it and structures the links between designers and sub-contractors in a design process. Lockledge [9] designs an Information System to facilitate communication between actors. Clarkson [10] explain the Visualization techniques to assist design process planning.

Our proposition is not against these theories of design but we consider them as approaches to be included in the practical aspect of our framework. That means we have been influenced by some various parts of several models. Closer to our research field, Norman [11], Maier and Fadel [12], Brangier [13] and later Brown and Blessing [14] works on the concept of "affordance". It refers to the capacity of a product to be understood and used without additional information. This concept is clearly related to classical products to be used by customers. However, it is possible to use this concept as well to qualify a learning/teaching product because even in this context products should be easily used by direct and indirect customers. To complete the qualification of a product, we propose the concept of "learnability" which refers to the capability of a product to support the process of knowledge transmission connecting direct customers to indirect customers. This concept includes the affordance. The "learnability" of a given product can be assessed, analyzed and improved by applying the ingenition process as it will be described hereafter.

3 Knowledge Theory: A Model Proposition

Holsapple and Jones [15] developed a value chain for KM process and activity. Lee and Yang [16] explained a knowledge value chain for all activities concerning KM. These endeavors are mainly theoretical and focus on activities in organization and a macro view of whole knowledge based view. We suppose that the problem of semantic view of KM, basically what happens and changes in the nature, content and context of concepts when transforming data to wisdom is unsolved.

In the value chain of knowledge creation, our goals are at first to valuate knowledge and the process that lead to organizational wisdom and in second to propose a conceptual framework that brings together basic semantic concepts in a universal way.

Knowledge value chain consists of the basic elements of semantic transformation, value processing activities, and output as final margin that here is knowledge performance. These processing components and activities are the building blocks by which a corporation creates a product or provides service valuable to its customers. Several authors tried to make a distinction between data, information, and wisdom;

some of them add a category as understanding [17]. In this paper we will try to make a distinction between data, information, knowledge, individual wisdom as competence or expertise, and collective wisdom as capability. This framework is drawn upon consciously and deliberate management of these activities as a global umbrella to bring together all transformation activities that lead to value creation for organizations. Fig. 2 depicts components of VCKC.

Based on the assumptions that each step and situation of this value chain is considered as in a unique time, unique place and context and with unique person, one of the results in this work is to open a window for information and engineering researchers to consider epistemological concepts in KM based upon process view.

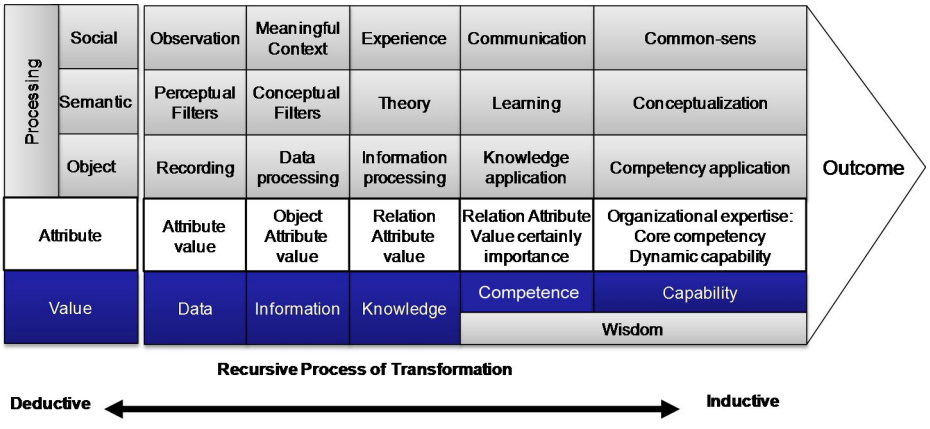


Fig. 2. Value Chain of Knowledge Creation (VCKC)

4 Learning Theory

In this field, we explore and compare our proposition with Giordan's proposition [18]. We will try to show that our model will be able to integrate all characteristics of this model. Research on knowledge and learning is currently converging towards several key findings. This specifically highlights the limits of both traditional teaching practices and of several innovations (active, non-directive, discovery methods).

The appropriation of knowledge results from design transformation processes in which the designer-learner must take the leading role. Knowledge acquisition is the result of elaboration activity during which designer-learner compares new information with mobilized knowledge and produce new meanings, which in turn are more appropriate for answering their questions. The main theories on learning are all quite limited in this respect. Therefore, to understand learning processes we must develop a new model that could integrate the parameters which challenge the mobilized designs. One attempt was initiated at the LDES in 1987 [19] and since 1989 it has been refined [20]. The model is now known as the Allosteric Learning Model (ALM). This model defines the issues, explain the main characteristics of learning, and allow predictions.

All of the theories require a detailed analysis in order to determine their overall potentials and limitations with regard to educational and cultural practices. Apart from certain cognitive approaches, learning is not the original focus of any of these theories but is considered, at best, as a potential side effect. However, when studying learning, we cannot just focus on learners and their conceptual mechanisms. Although they own self-organization capabilities, they are largely inter-dependant and related to conditions and to the successive environments through which they have emerged during each individual's history. To fill this gap we have tried to develop a new model, which combines 'interaction' and 'elaboration' but also 'integration' and 'interference' (fig. 3). In ALM, as explained above, learning is not dependant on a single factor, but on a network of conditions we call the 'didactic environment'; this is overwhelmingly important for teaching and for science popularization in general.

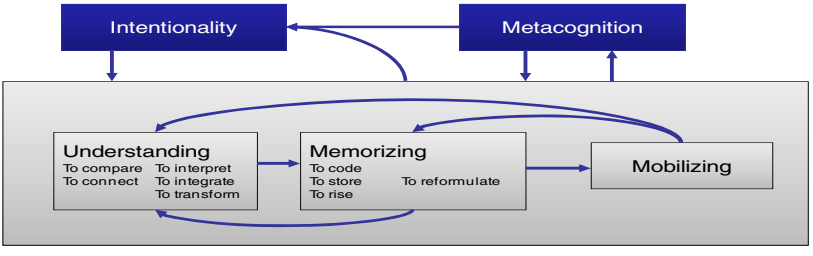


Fig. 3. Allosteric Learning Model (Giordan, 1994)

5 Ingenition: A Process to Improve Design for Learning/Teaching

The whole ingenition process is based on a cycle which studies two joint knowledge and competence fields (fig. 4). During the very first step of this process, a *macro-model* is defined (the grid on the top of the scheme) in order to determine the strategic orientations of the firm in terms of internal trainings for employees and external learning/teaching acquisition for customers. Then teaching/learning situations supported by learning/teaching-oriented product are studied in order to understand real customers' needs. These situations are modeled by AS IS "graph of knowledge elements" called also the ingenition micro-models. These graphs are obtained after observations of actions and reactions of trainees and their teacher during the teaching/learning processes.

After analyzing the graph, regarding various available references and target competences, several TO BE knowledge graphs may be established. Once assessed and analyzed, the most relevant graph is chosen and it will be used as a main framework from which specific constraints for design should be extracted. Moreover, from this new graph, protocols regarding training sessions could be identified in order to optimize customers' satisfaction by offering more efficient learning sequences to direct customers.

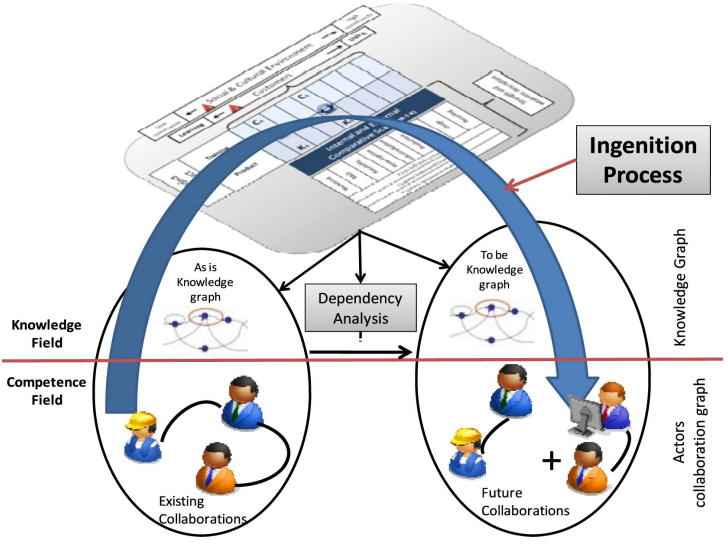


Fig. 4. General Ingenition framework

5.1 Ingenition Process Macro Model

When a project leader conducts its mission, (s)he should have a cross-functional (horizontal) view of it. By extending this notion to the whole product lifecycle, it can be seen that the learning process could cross the lifecycle phases. It means that there is a learning process parallel to the whole lifecycle. This is obviously not the case of the system managers who have very often a functional (or vertical) view of a given part of the system. This means that for a given product development project, it could be necessary to model all necessary learning activities within a global model covering both vertical and horizontal learning processes. We use a macro-model to represent these two complementary views of the same learning process (Fig. 5). Once established, this model makes possible the identification of most critical elements and of the life cycle of product development and to act consequently.

5.1.1 Environmental Descriptors

5.1.1.1 Product Characteristics. A product is used either for usage or learning and its characteristics (functional, structural and behavior) would not be the same. Obviously, there is a continuous scale going from the purely usage products to purely learning ones. The social and cultural constraints have to be also taken into account. By schematizing these constraints with a relative position of a cursor on a continuous scale, the socio cultural context of the product's environment is defined. Audits of experts of the firm will allow the definition of the relative position of these cursors.

5.1.1.2 Extended Product Analysis. A learning/teaching oriented product is an extended product containing at the same time not only the physical product but also some associated services. We consider training as one service associated to the physical product. By integrating knowledge elements accumulated in the firm (represented by KΔ), it is supposed that it will be possible to generate new competences (represented by CΔ) for indirect customers. The difference between knowledge and competence and the process which allows transforming knowledge to competence is based on Giordan's allosteric model [18].

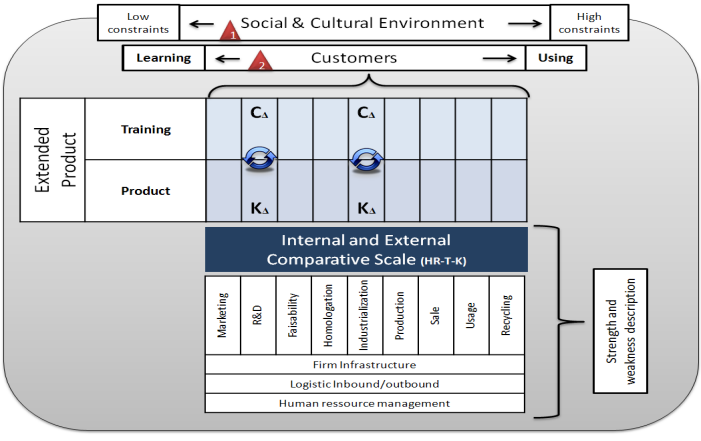


Fig. 5. Ingenition process macro-model

5.2 Ingenition Process Micro-models

The main purpose of the ingenition process is to focus on competences and the way to obtain C_Δ from knowledge K_Δ. Competences are those knowledge elements gathered, structured and usable by users and obtained during various trainings sequences. Formally we may model this main interaction by $K_\Delta \xrightarrow{i} \left| C_\Delta \right.$. The vertical line stands for training sequences. The transformation of knowledge to competences uses specific supports. One or several Learning Object (LO) may support this process, modeled by $K_\Delta \xrightarrow{i,\{LO\}} \left| C_\Delta \right.$. For example, both the software package implemented in a robot and the robot itself represent the two learning objects for a specific learning purpose; see for instance [21].

From a generic point of view, it is possible to decompose K_Δ into a sequence of dependent and more detailed knowledge elements. The scheme at the left side of Fig. 6 represents this graph. Ci and Cf correspond respectively to initial and final level of competence of indirect customers. Then C_Δ corresponds to the difference between these two levels of competence. Within the ellipse, we can see the graph of dependent knowledge elements. The ellipse models the frontier of the studied ingenition process. Each couple of knowledge elements is connected together by a dependency link. (K1, is

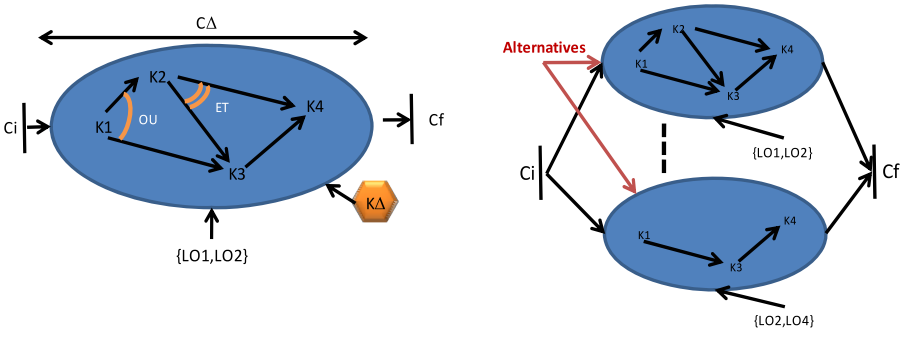


Fig. 6. Knowledge dependency graph

required for understanding of transistor, K2). This graph, which allows providing a competence to some trainees, is supported by a set of learning objects. The whole process is performed by actors but for simplicity they are omitted in the scheme.

The analysis of the knowledge graph can be based on various dependencies identified within the graph. Three basic logical dependency relationships may be identified within the graph: **Antecedence**: The understanding and description of Kj is not possible without explaining Ki. **Parallelism**: The understanding of Ki and Kj is independent. **Simultaneity**: K1 and K2 should be treated at the same time.

Consideration of these various dependencies could define directly the way that various learning objects are used: *As Is situation*. By analyzing these scenarios regarding target competencies of a learning situation, learning objects and their connections, and strategic learning decisions made within the ingenition macro-model, it should be possible to build learning alternatives (the schema at the right side of Fig. 4): *To Be situation*.

Each learning alternative defines target competences, necessary learning objects, and the graph of knowledge dependencies. The definition of these alternatives imposes structural and functional constraints on the learning objects design and realization.

The knowledge dependency graph obtained based on the analysis of target competence (Cf) is called *primary dependency graph*. Based on the choice of learning objects, complementary graphs may be added to the primary graph in order to let the indirect customers reach the target competence. These complementary graphs are called *auxiliary graphs* and their existence and complexity may give a clear indication of the learning/teaching efficiency. Using an engine to show fixed pivot liaison could require complementary knowledge to include in the learning/teaching sequence; auxiliary graph. In section 6, these two types of graphs will be illustrated more in detail.

5.3 Learning Objects, Learning Products and Design Process

A product may be seen from various points of view: design, realization and learning for example. All of these definitions should be connected to the design point of view. In fact, this view determines the rest of the product development and usage process. From the learning point of view, a product is decomposed into an arborescence of learning objects (defined in the past section). Every node of this decomposition corresponds to a

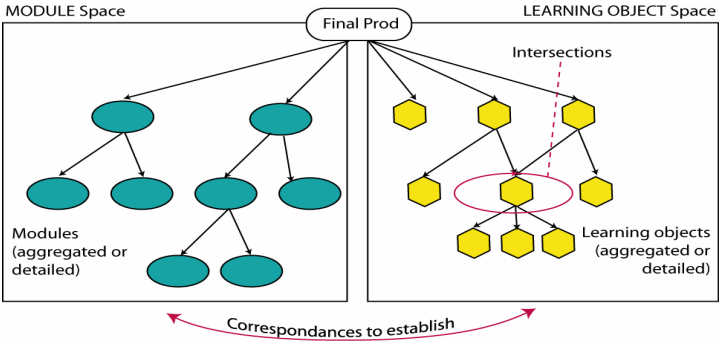


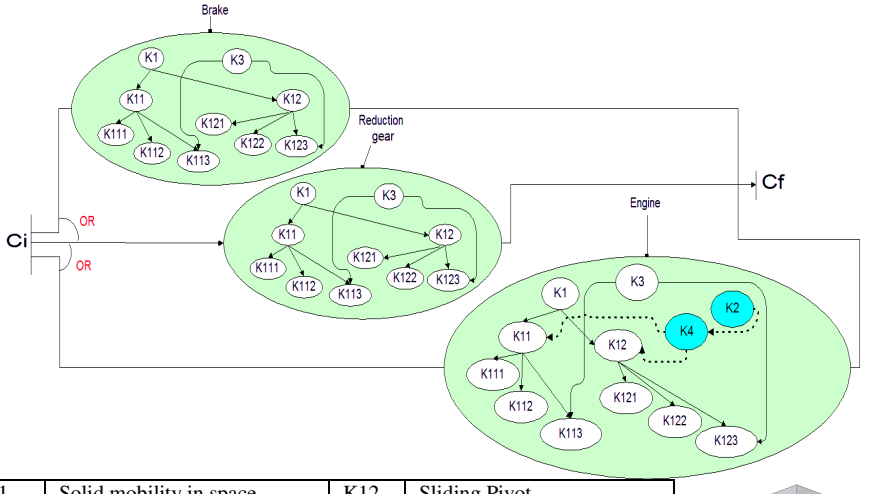
Fig. 7. Modular and Learning spaces

real learning object. A learning object is a collection of product modules which can be used as a support for knowledge transmission. Defined as such, it is clear that there could be intersections between two learning objects taken separately while there is no intersection between two modules. But, modules have mutual exchanges. These are the main differences between learning objects and product modules. Fig. 7 represents these two decompositions of a given product.

6 Description of Case Study

This model is applied to a French company called here *Innovia*. Innovia designs and produces products for primary, high schools and universities. Innovia's top management looks for a new product to support training mechanical devices in high school. The product is a small airplane. Its cultural and social environments do not impose hard constraints on the product (see cursor 1 on the top of the fig. 5) that is the reason why the cursor is positioned at the left end of the scale. Another cursor (triangle cursor 2 on Fig. 5) is positioned on the usage scale. This scale addresses the first usage of the product goes on a continuum from purely learning oriented to purely usage oriented. For this airplane, Innovia selects a highly learning oriented level. Direct implications of these two scales in terms of shape, materials, maneuverability, etc. of the product should be taken into account from the beginning of the New Product Development project.

The micro-models were developed to represent potential alternatives for teaching fixed pivot (K11) and sliding pivot (K12) connections. Each type of connection can be taught by at least one the following approaches: cinematic 2D diagram (K111 or K121), cinematic 3D diagram (K112 or K122) or by their torsor expressions (K113 or K123). Before teaching pivot liaisons, it is decided to teach also the solid mobility in space (K1). By representing all these knowledge elements, the primary dependency graph is obtained and is shown in the ellipses (Wight nodes and plain edges). Three potential learning objects may be used to illustrate the pivot liaison: a bicycle brake, a reduction gear or an engine (ellipses of fig. 8). Primary graph is the same for these three solutions. However, by using the engine, it is clear that additional knowledge elements should be included within the knowledge dependency graph. This is the auxiliary graph which contains: the main power transmission theory (K4) which requires a good understanding



K1	Solid mobility in space	K12	Sliding Pivot
K11	Fixed Pivot	K121	Cinematic 2D sliding Pivot
K111	Cinematic 2D fixed Pivot	K122	Cinematic 3D sliding Pivot
K112	Cinematic 3D fixed Pivot	K123	Torsor of a sliding pivot
K113	Torsor of a fixed pivot	K3	Torsor manipulation
K2	Power, Energy, Work, Force	K4	Power transmission

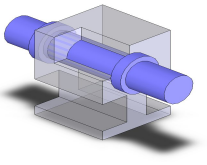


Fig. 8. Micro-model of the pivot linkage

of fundamental mechanical concepts such as energy, force, power and work. The auxiliary graph is represented by the blues nodes and dotted edges.

Some observations were made according to these micro-models: 1) Primary graph. It is necessary to identify clearly the primary knowledge dependency graph. A deep analysis of this graph helps to extract recommendations for designers and also for direct customers. 2) Auxiliary graph. In this example, it can be seen that including complex learning-object such as an engine in the learning-oriented product could diminish the learnability due to the increasing complexity of the graph, i.e. by including various supplementary elements. This is an important indicator of efficiency for final users. 3) By extracting primary and auxiliary graphs, new learning objects may be necessary to make them understandable. 4) It may be possible to analyze various alternatives in a performance assessment strategy in order to choose the most relevant learning object to the expected target and also to the trainees' specificities.

7 Conclusion

In this paper, we study the learning dimension of a product. It is argued that knowledge generation not only represents an important internal innovation source but also, the firm can use the learning and generated knowledge as a tool for positioning firm on the market. In this paper we present the basic underline theories with which we developed our unique framework in knowledge ingenition process. Based upon the model proposed as VCKC, the explanation of some related design theories and the positioning

of the learning theories context, we introduced an ingenition process as a global framework.

We propose the two models of the ingenition methodology. This methodology is built to ensure two goals: analysis and design of learning/teaching-oriented product and analysis and design of learning/teaching sequences. The ingenition is a methodology to engineer learning/teaching processes. An illustrative example is presented at the end.

In short, the main tool presented here, the learning grid allows: 1) To model the social and cultural environment regarding learning purpose of the firm. 2) To underline the purpose of the product, learning or using or something in between. 3) To keep track of Knowledge generated in relation with the activity considered. 4) To measure the variations between what the firms can do inside and what it should be outsource.

The micro-models represent the knowledge dependency graphs. These graphs should correspond at least to the map of necessary knowledge for a given purpose. Each graph is supported by one or several learning objects.

Our research is focused on a highly dynamic market: Education market. We are aware of the fact that a huge amount of works still has to be done in this field. Authors are working on a complete description of the ingenition methodology and would apply it in 20 schools of south-west of France. The results of this study will consolidate the ingenition approach.

References

1. Nonaka, Toyama, Konno: SECI, Ba and Leadership: a Unified Model of Dynamic Knowledge Creation. Long Range Planning 33, 5–34 (2000)
2. OECD. Analyse des politiques d'éducation (1998)
3. Tollenaere, M.: Quel modèle de produit pour concevoir? In: Symposium International La conception en l'an 2000 et au-delà. outils et technologies, Strasbourg, France (1992)
4. Grabowski, H.: Towards A Universal Design Theory. In: Kals, H. (ed.) Integration of Process Knowledge into Design Support Systems, pp. 47–56. Kluwer Academic Publishers, Dordrecht (1999)
5. Umeda, Y., Takeda, H., Tomiyama, T., Yoshikawa, H.: Function, Behavior and Structure. In: Applications of Artificial Intelligent in Engineering. Springer, Berlin (1990)
6. Shimomura, Y., Takeda, H., Yoshioka, M., Umeda, Y., Tomiyama, T.: Representation of Design Object Based on the Functional Evolution Process Model. In: Design Engineering Technical Conferences, ASME 1995, Boston, USA (1995)
7. Andreasen, M.M.: The Theory of Domains. In: Workshop on Understanding Function and Function to Form Evolution. Cambridge University, UK (1991)
8. Fagerstrom, B., Johannesson, H.: A product and process model supporting main and sub-supplier collaboration. In: 13th International Conference on Engineering Design, Glasgow, UK, August 21-23 (2001)
9. Lockledge, J.C., Salustri, F.A.: Design Communication using a Variation of the Design Structure Matrix. In: 13th International Conference on Engineering Design, Glasgow, UK, August 21-23 (2001)
10. Clarkson, P.J., Melo, A.F., Eckert, C.M.: Visualization Techniques to Assist Design Process Planning. In: 13th International Conference on Engineering Design, Glasgow, UK, August 21-23 (2001)
11. Norman, D.A.: The Psychology of Everyday Things. Basic Books, Inc. (1988)

12. Maier, J.R.A., Fadel, G.M.: Affordance: The Fundamental Concept in Engineering Design. In: Proceedings of ASME Design Theory and Methodology Conference, Pittsburgh, PA. Paper no. DETC2001/DTM-21700 (2001)
13. Brangier, E., Barcenilla, J.: Concevoir un produit facile à utiliser. Adapter les technologies à l'homme, Éditions d'Organisation (2003) ISBN: 2-7081-2900-7
14. Brown, B.: The relationship between function and affordance, Idetc/cie, Long Beach, Ca, DETC2005-85017, September 24-28 (2005)
15. Holsapple, C.W., Jones, K.: Exploring Primary Activities of the Knowledge Chain. Knowledge and Process Management 11(3), 155–174 (2004)
16. Lee, C.C., Yang, J.: Knowledge Value Chain. Journal of Management Development 19(9), 783–793 (2000)
17. Ackoff, R.L.: From Data to Wisdom. Journal of Applied Systems Analysis 16, 3–9 (1989)
18. Giordan, A., Girault, Y., Clement, P.: Conceptions et connaissances, Peter Lang (1994)
19. Giordian, A.: Les origines du savoir, Delachaux (1987)
20. De Vecchi, G., Giordan, A.: L'Enseignement scientifique, Comment faire pour que ça marche? Z'Editions (1989)
21. Bourda, Y., Hélier, M.: What Medata and XML can do for Learning Objects. Webnet Journal, 24–31 (January/March 2000)

Towards "Kiga-kiku" Services on Speculative Computation

Naoki Fukuta[1], Ken Satoh[2], and Takahira Yamaguchi[3]

[1] Shizuoka University, Hamamatsu Shizuoka, Japan
fukuta@cs.inf.shizuoka.ac.jp
[2] National Institute of Informatics, Chiyodaku, Tokyo, Japan
[3] Keio University, Hiyoshi, Kanagawa, Japan

Abstract. In this paper, we propose a concept for service called "kiga-kiku" service. In that, an agent proactively detect potential failures in providing a series of services for users, and prepare and execute follow-up plans for the failures automatically. The name of "kiga-kiku" is derived from a Japanese word meaning of proactive behavior to keep comfort of other people by using prediction of other people's behaviors, wishes, and preferences with shared social context. We show that a certain kind of "kiga-kiku" service can be realized as a combination of inference capability about preparation of possible failures and execution of follow-up plans in an acceptable cost. We also present a case study about "kiga-kiku" service and we show it is possible to implement such mechanism by simply adding a "kiga-kiku" service agent as a front-end to the existing service systems in a reasonable development cost.

1 Introduction

To enable highly intelligent systems and services, it is important to predict the users' behavior, intention, and wishes progressively and accurately. An approach for these predictions is log-based approach. For example, collaborative filtering is proposed to recommend a potential needs of pictures, music CDs, and other goods from past behaviors of similar users[1]. It is also used to predict missing data in the information retrieval researches[2]. More generally, association rule mining is very strong approach to find out relationships among items in a large data set[3,4]. Also applicability of association rule mining for infrequent but expensive items is recently investigated[5].

On the other hand, on Business Process Management(BPM), it is an important issue to build and maintain a process that is robust for small number of unexpected failures in the process[6]. Klein et, al. presented a systematic approach about handling exceptions in business process[7] and a domain-independent exception handling approach in multi-agent systems[8]. However, the shown exception handling is rather for reactive handling after exception has occurred but not intended for the purpose of proactive protection from potential failures. The above prediction approaches can also be used to prevent potential failures from users in a business workflow, when they predict potential failures and possible

T. Yamaguchi (Ed.): PAKM 2008, LNAI 5345, pp. 256–267, 2008.

alternative plans. However, those approaches do not consider contexts or underlying knowledge in their predictions. Therefore, for example, it needs additional functionality to predict a kind of failures that are *logically incorrect* but *happens rarely*. For this reason, logic-based approach is also investigated[9]. Such proactive support facility on software systems are also implemented in commercial office productivity software. However, when they are less intelligent, they sometimes do not help users but rather they make troubles. A good example to understand this issue is the automated editing support mechanism implemented in the past famous word-processor product. It wrongly edits and changes the users document by using too simple rules with very complicated operations so that the user sometimes could not control them effectively and just disable them to avoid harmful side effects. For this reason, typically we rather use explicit commanding for "safer" communication. Same approach is also used in many software applications and services. This is good when it is difficult to predict user's intentions and contexts behind them. For example, when there are a lot of users who have different intentions, contexts and cultural backgrounds, it is very difficult to predict their intentions, wishes, and possible behaviors accurately.

When there are some potential failures that cannot be recovered after they happen, it is very important to predict and avoid such failures proactively before explicit commands are given by users. In case a system is highly personalized/specialized for a specific user or group, and when it is also optimized for a specific context and purpose, it is possible to add an inference facility to predict user's implicit demands to prevent potential risks by the system proactively. Here, an important issue is to share the underlying context to predict the user's demands accurately. In this paper, we propose an inference mechanism based on "kiga-kiku" concept that is derived from Japanese traditional cultural behaviors. We apply our mechanism to find out possible improvements for an existing service built as a service composition system. Also we assure how the found improvement plans has been implemented to the existing system in a small cost.

The rest of the paper is organized as follows: Section 2 shows the backgrounds and definition of "kiga-kiku" concept. Also we show the underlying logic to implement the "kiga-kiku" concept to the real system. Section 3 describes a case study to extend an existing ordinary Web-based system to an *agentified* "kiga-kiku" system. Section 4 concludes the paper and mention about our future work.

2 Basic Concepts

2.1 Background and Related Work

When it is easy for all members of a society to predict the one's behavior, intentions, and wishes, we can implement a more intelligent automation on using systems and services. In Japan, it is considered as valuable thing to conjecture other people's intention or feeling and respond their implicit demands before receiving explicit request rather than responding people's explicit request at the requesting time. We call this concept "kiga-kiku", that is the word of a Japanese traditional custom. Furthermore, in the "kiga-kiku" concept, sometimes one may

predict the other's potential requests from underlying contexts even when it has not been noticed yet by the requester itself. In Japan, we think historically many people deeply share social contexts due to its geographical condition (e.g., Japan consists of small islands), political history(Japan had been closed for over 300 years until 1858) and other reasons. Therefore such custom is still found in Japan.

We believe the concept of Japanese "kiga-kiku" could be useful for a certain kind of applications that need to support users by intelligent and proactive behaviors. Here, an important issue is how to adjust the degree of the autonomy. In other words, an agent should know when the agent should do an autonomous behavior in which context. Tambe et, al. proposed a concept 'adjustable autonomy'[10] that realizes automated adjustment of autonomy level on agent systems. However, the adjustment is left for the users' feedbacks so therefore it is difficult when the user does not know whether the system should behave autonomously. In our 'kiga-kiku' approach, rather we try to adjust autonomy-level when users are not familiar about the cases, by using underlying shared contexts and risk assessments in the service scenarios.

In this paper, we propose a usage of "kiga-kiku agent" that realizes "kiga-kiku" services based on "kiga-kiku" concept presented above. As we discussed above, in "kiga-kiku" service, an agent predicts user's intention and then gives an appropriate service. However, as we said above, this prediction is successful only when the underlying context is shared between both sides. Therefore, a simple "kiga-kiku" application would fail for another task domains and their users since the application could not share contexts to the users.

As we mentioned, it is very important to share the context to be able to predict one's behavior, intentions, and wishes. Collaborative filtering or other data mining techniques could be possible solutions to learn and predict user's intention or preference from the log of past behaviors of users and others. However, we argue that it is not enough to prepare a mechanism to predict the users intentions and preferences from data since it is difficult to know the reason. Rather we seek an approach for this issue by logic-based approach that will make inference about users and predict users needs with reasons.

Of course a "kiga-kiku" system should have such prediction mechanism and the system should predict user's intention or preference properly to provide "kiga-kiku" services. Even so, this kind of prediction is only a prediction for preferences but it does not consider potential failures in a process of work. Furthermore, it is very difficult to learn such potential failures since it happens not so often. We argue that we need another facility of backing-up mechanism in order to handle this failure by logic-based approach.

2.2 A Natural Language Definition of 'Kiga-kiku' Concept

In this section, we try to present the definition of our "kiga-kiku" concept. At a starting point of this definition, we present a natural language definition of the concept instead of a formal definition.

A natural language definition of "kiga-kiku" agents is as follows:

- The user has initiative and agent's background behaviors are rarely recognized by the user.
- The agent predicts user's intention, preferences, and possible failures in the current goal.
- When the agent recognizes a need for proactive action, the agent does it with the best prediction value.
- When the agent recognizes a certain possibility of failure for the current goal with working on a current workflow, the agent suggests the users to switch another workflow to accomplish the goal by an alternative way. Note that in such case the workflow transition should be done seamlessly by using background reuse and re-run similar parts in the workflow.
- When inappropriate side-effects will appear by such proactive actions, the agent reverts such side-effects.

2.3 Formal Definition

We use a simple BDI agent[11] with plans and goals for our definition of agents.

Definition 1. Agents: *Let A be an agent, B be beliefs, G be a set of goals or constraints to be satisfied, P be a set of plans to accomplish G, agent A is defined as $A = \{B, G, P\}$.*

In this paper, we assume agents try to accomplish their goals by using external services. Therefore, possible actions of agents are only invoking external services. We define services as simple (atomic) services and complex services that are compositions of atomic services. This definition is borrowed from OWL-S[12].

Definition 2. Services: *Let S be a set of available services, $Proc$ be a set of processes to use services in S, and I_s, O_s, $Prec_S$, $Post_s$ be inputs, outputs, preconditions and postconditions(side-effects) for a service $s \in S$, respectively. A simple service s is defined as $s = \{I_s, O_s, Prec_s, Post_s, \phi\}$. A complex service $s' = \{I_{s'}, O_{s'}, Prec_{s'}, Post_{s'}, proc_{s'}\}$ is a combination of services $Sc \subseteq S$ with a process $proc_{s'} = proc(Sc) \in Proc$ such that $proc(Sc) = \{s_i \cdot op \cdot proc(Sc')\}, s_i \in Sc, Sc' = Sc - s_i, op \in OP$, and $I_{s'} = \sigma_{s_i \in Sc} I_{s_i} - \sigma_{s_i \in Sc} O_{s_i}$, $O_{s'} = \sigma_{s_i \in Sc} O_{s_i} + \sigma_{s_i \in Sc} I_{s_i}$, respectively.*

Goal is prepared as a skeleton of service. This approach is same as that of Web Service Modeling Ontology(WSMO)[13].

Definition 3. Goals and Plans: *A goal $g \in G$ can be defined as a skeleton of service $s_g = \{\phi, O_s, \phi, Post_s, \phi\}$. Also a plan $p \in P$ can be defined as a possible process $proc_{s_g}$ that implements a skeleton of service s_g. Therefore, for an agent A, the problem is to choose best $p \in P$ that can be implemented by a complex service s'_g that satisfies a goal $g \in G$.*

Here, we assume to have a good planning engine that decompose goals to a process for combinatory use of services. Also we assume to have a good quality

assessment mechanism for those service processes (e.g., [14]). In this paper, we use the word 'workflow' be the same meaning of a whole plan that satisfies the goal of the agent.

Additionally, we define two types of relationships among services, compatible and inverse, for the definition of robust plans.

Definition 4. Compatible Services: *Let S be a set of available services, Proc be a set of processes to use services in S, and I_s, O_s, $Prec_S$, $Post_s$ be inputs, outputs, preconditions and postconditions(side-effects) for a service $s \in S$, respectively. A service $s' = \{I_{s'}, O_{s'}, Prec_{s'}, Post_{s'}, proc_{s'}\}$ is a compatible service of $s = \{I_s, O_s, Prec_s, Post_s, \phi\}$ if $O_s \subseteq O\_s'$, $Post_s \subseteq Post_{s'}$, $Prec_{s'} \subseteq Prec_s$, and $I_{s'} \subseteq I_s$}.*

Definition 5. Inverse Services: *Let S be a set of available services, Proc be a set of processes to use services in S, and I_s, O_s, $Prec_S$, $Post_s$ be inputs, outputs, preconditions and postconditions(side-effects) for a service $s \in S$, respectively. A service $s' = \{I_{s'}, O_{s'}, Prec_{s'}, Post_{s'}, proc_{s'}\}$ is an inverse service of $s = \{I_s, O_s, Prec_s, Post_s, \phi\}$ if $Post_s \subseteq Prec_{s'}$, $Post_s + Post_{s'} = \phi$, and $I_{s'} \subseteq \{I_s \cup O_s\}$.*

We define robust plans, as follows:

Definition 6. Robust Plans: *Let $Sf_{s'}$ be a set of failures in a complex service s', a "kiga-kiku" plan could be a plan that is robust for failures of $Sf_{s'} \subseteq Sc'$ in a process $proc_{s'} = proc(Sc')$.*

To prepare robust plans for potential failures Sf_s, we add a speculative execution of services $S_{comp}(Sf_s)$ that are compatible to Sf_s with services $S_{inv}(S_{comp}(Sf_s))$ that are inverse services of $S_{comp}(Sf_s)$ to the plan be robust.

Our long-term research goal is to realize planning and execution mechanism for such robust plans for critical potential failures. Here, an important issue is how to find out which failures could be critical in the plan. Also it is important to confirm such service execution mechanism can be implemented in reasonable costs. In this paper, we focus on those two issues.

2.4 Implementing Kiga-kiku Agents Using Speculative Computation

To implement "kiga-kiku" agents, we need a certain background mechanism to realize them. For such purpose, speculative computation[15,16] is investigated on agent research domain. Recent years, we have been investigating "speculative computation" to realize a good reasoning mechanism for agents [17,18,19,20]. In our works, the term "speculative computation" means the following computation: When an agent needs to ask a question to other agents to proceed his reasoning, an agent just issues a query for the agent and continues its reasoning based on the assumed answer that predicted by the agent. If the actual answer from the other agent is consistent with the prediction, the reasoning continues. Otherwise, the reasoning is stopped and an alternative path of reasoning is started with the received answer.

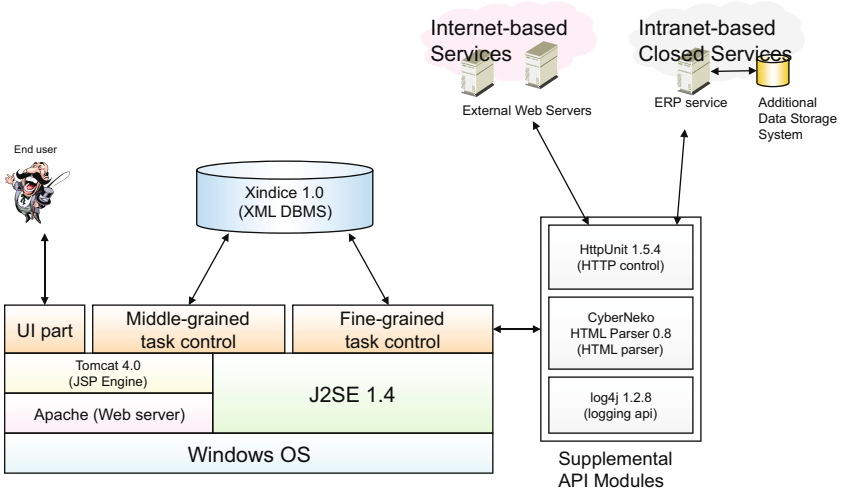


Fig. 1. Implementation Architecture of the BTSS System

We consider realizing "kiga-kiku" service by extending speculative computation of agents. Here, agents that implements "kiga-kiku" services invoke external services speculatively, instead of communicating to other agents in the above scenario. In this case, we should carefully treat the side-effects that are produced by services. Of course inter-agent communication may result side-effects to the environment, typically such side-effects can be managed by the communicating agent itself. Therefore, it is relatively easy to implement 'undo' mechanism for such side-effects. However, external services are just services but not agents, they may not have functions to revert their produced side-effects. In such case, we need to formalize and describe such effects / inverse effects relations carefully. Therefore, we presented definitions of services and agents in Section 2.3.

3 Case Study

3.1 Business Trip Support System

In our past project, we developed a service composition system for office business trip management tasks. The system integrates three core services and two dozens of additional services by using service integration mechanism. In the mechanism, ontology-based service matching has been implemented[21] for ad-hoc service composition. Also hand-made composed service processes have been developed that guarantees quality of composed services. Additionally, to support better data passing among services, an ontology-based data class design support method has been presented and used for the development[22].

The system provides "one-stop" integrated business trip support service that utilizes Internet-based services (train connection search service, hotel reservation service, train seat reservation service), and closed Intranet-based enterprise

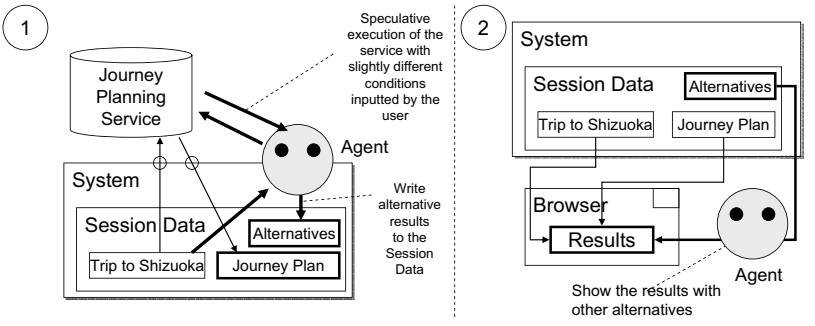


Fig. 2. Parallel Search for Alternatives by a Speculative Computing Agent

resource planning systems. For example, when a user enters a business meeting event on a location, the system automatically calculate a plan for reaching the location from the company, proposes a plan to use express trains and recommended hotels. Furthermore, the system has capability to issue actual reservation requests for train seats and hotel rooms via web-based interface on the background. Finally, plans and logs are sent to the enterprise system for further use.

The system has a web-based user interface that is build as a web portal site in a division of the company. Figure 1 shows the implementation architecture of the system. The system was implemented on Open Source Web Application Frameworks (JSP, Apache web server, etc.) with several useful APIs.[1] The system has been developed in cooperation with a Japanese company and primarily intended to be used in the company. For details about implementation, see [23].

3.2 Attaching Kiga-kiku Agent to Existing System

We developed a prototype system of kiga-kiku agents based on the business trip support system. This is done by attaching an agent which performs a user-prediction and handles exceptions of prediction into the existing Web-based service system with cancellation mechanism.

In this case study, we focused on examining difficulty of implementation and cost efficiency for realizing "kiga-kiku" services. We made a development team, it consists of two professional software developers who did not know the previous system, and a support person who are familiar with the implementation of the previous system.

First of all, we extracted several "kiga-kiku" service processes from our failure prediction mechanism. Since it is very difficult to set appropriate parameters for potential failures of services, the extracted service processes are chosen by our heuristic sense, rather than certain computational criteria. We asked the professional developer to extend the system to satisfy our specified functionalities

[1] At the time of implementation, not all services were available as SOAP-based Web services. Therefore, we used wrapper program to communicate ordinary Web-based applications.

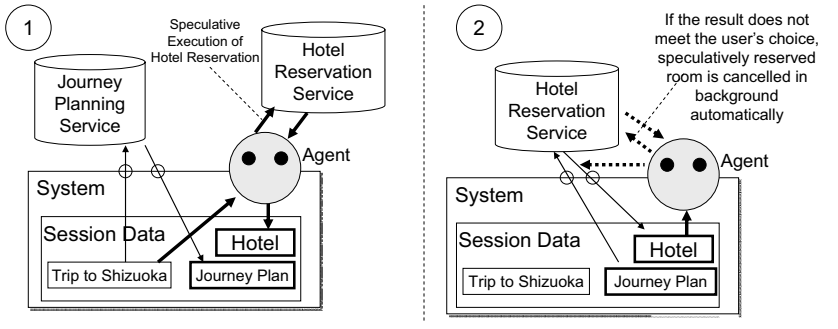


Fig. 3. Consistency Management of Speculative Computation and User Choice

that are derived from above "kiga-kiku" service processes. The specification of additional functionalities has been derived from the logic of "speculative computation".[2]

Figure 2 shows an example of our implemented functionality, that searches an alternative trip plan by train in parallel. This could be a backup plan when the user missed a train in a certain reason. Figure 4 shows an actual notification to the users on the user interface. This is a test case of implementing speculative invocation of same kind of services with different parameters.

Figure 3 shows another example. Here, the agent proactively book a hotel room before the user's explicit request, preventing a failure of booking a room in a reasonable price.[3] This is a test case of implementing speculative invocation of different kind of services with side-effects by considering underlying shared contexts.

Furthermore, it is also possible to indicate when room booking is not required in alternative plan. Figure 5 shows an actual notification to the users on the user interface. This is a test case of implementing speculative invocation with workflow switching (e.g., switch to *one day business trip* from *business trip with stay*, seamlessly).

In this case study, we implemented several test cases that covers possible scenarios produced by "kiga-kiku" agents. Here, we only spent less than 1/20 of cost compared to the building cost of the base system for this improvement. We think this is reasonable enough to implement our concepts on actual business process support systems.

Of course, from the viewpoint to realize each specific service, we know similar services have already been provided by some ASPs. Also some frameworks have been presented for specific application scenarios. For example, a generic framework for anytime scheduling has been proposed for dynamic planning of train routes[24]. However, their focus is not on preventing potential failures but

[2] Unfortunately, we could not complete all descriptions and underlying ontologies for the scenario. Therefore, the derived specification included a part that are based on our computation by hands.

[3] Since this operation is hidden from users, we omit the example of user interface.

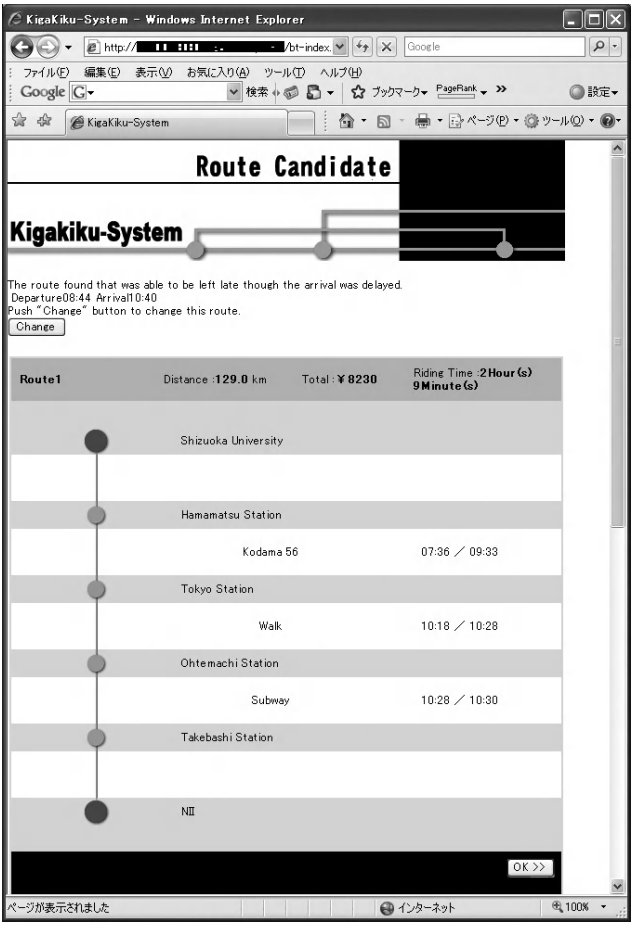


Fig. 4. Example of "kiga-kiku" notification on Transportation Search Process

rather seeking optimal ways in the specific scenario. We don't say our provided services themselves are new and sophisticated. Rather, we show this kind of service improvements can be derived from our concept, and it can be implemented within a reasonable cost. Our idea can also be applied to many workflows used in certain companies. That will effectively protect users from potential failures in their workflows.

Our primary goal is to use speculative computation for preventing potential failures that would be caused by 'late decisions'. For example, when we need to book a hotel room on high season, it is very important to book the room as soon as possible, or all reasonable rooms would be reserved by others. Here, when we only protect the user from potential failure, there is little need to prepare accurate prediction of the user's choice. However, when the prediction is highly accurate, it also improves response time when using services. For example, when a user tells a need to have a reservation of a hotel room, the agent could issue

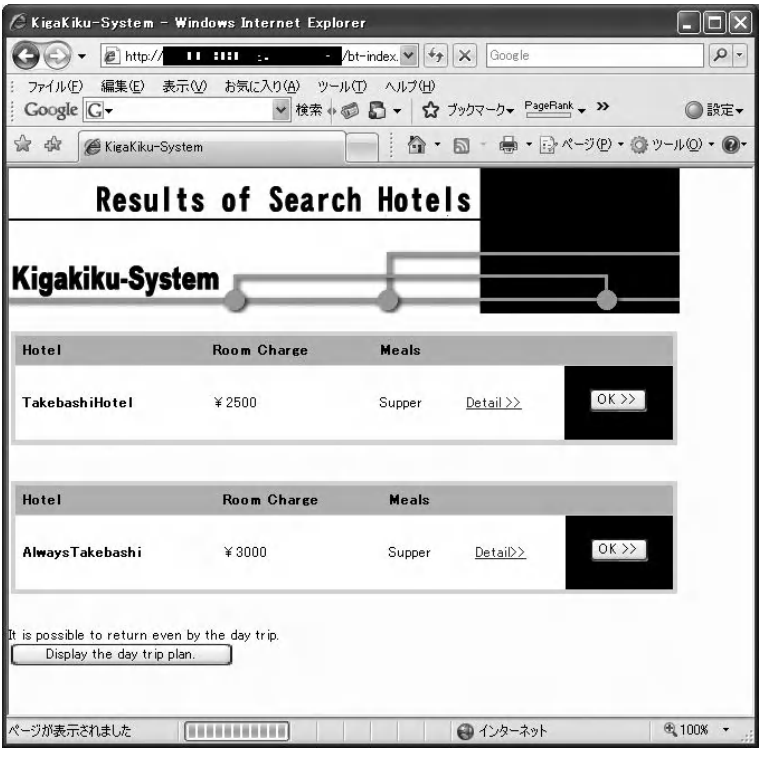


Fig. 5. Another Example on Hotel Reservation Process

reservation request before the user choose the hotel room. Therefore, when the prediction hits the user's decision, there is no need to wait for completion of the reservation request. In case the prediction is different from user's choice, cancellation procedure could be done in parallel. Therefore, when the computer has enough computing resource, there could be no actual penalty for response time of the "kiga-kiku" service.

4 Conclusion

In this paper, we proposed the concept of "kiga-kiku" service that predict user's implicit demands and proactively protect the users from potential failures in the service processes. We proposed an idea to realize "kiga-kiku" agents based on our past work, speculative computation on agent systems. We implemented the mechanism on existing Web-based business workflow support system. We confirmed the implementation cost is reasonable enough to apply real business support systems that are built as Web-based services.

Currently, we assume we could have appropriate services, service descriptions, a good semantic service composition mechanism, and an appropriate risk assessment mechanism for service processes. Also we assume we have a reliable prediction

mechanism for user's preferences possibly implemented by collaborative filtering or data mining techniques. Presenting feasibility study and discussions about realization and preparation of such mechanisms and descriptions is our future work. Furthermore, in this paper, we only captured the concept of "kiga-kiku" in a limited aspect. Incorporating other aspects of "kiga-kiku" concept to our work is another future work.

Acknowledgements

We thank Prof. Randy Goebel and Prof. Sandra Zilles for their constructive comments.

References

1. Herlocker, J.L., Konstan, J.A., Terveen, L.G., Riedl, J.T.: Evaluating collaborative filtering recommender systems. ACM Transactions on Information Systems 22(1), 5–53 (2004)
2. Ma, H., King, I., Lyu, M.R.: Effective missing data prediction for collaborative filtering. In: Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval(SIGIR 2007), pp. 39–46. ACM, New York (2007)
3. Hipp, J., Güntzer, U., Nakhaeizadeh, G.: Algorithms for association rule mining — a general survey and comparison. SIGKDD Explorations Newsletter 2(1), 58–64 (2000)
4. Ceglar, A., Roddick, J.F.: Association mining. ACM Computing Surveys 38(2), 5 (2006)
5. Zhou, L., Yau, S.: Association rule and quantitative association rule mining among infrequent items. In: Proceedings of the 8th International Workshop on Multimedia data mining(MDM 2007), pp. 1–9. ACM Press, New York (2007)
6. Mutschler, B., Weber, B., Reichert, M.: Workflow management versus case handling: results from a controlled software experiment. In: Proceedings of the 2008 ACM symposium on Applied computing(SAC 2008), pp. 82–89 (2008)
7. Klein, M., Dellarocas, C.: A systematic repository of knowledge about handling exceptions in business processes. In: ASES Working Report ASES-WP-2000-2003. Center for Coordination Science, Massachusetts Institute of Technology (2000)
8. Klein, M., Rodriguez-Aquilar, J.A., Dellarocas, C.: Using domain-independent exception handling services to enable robust open multi-agent systems: The case of agent death. The Journal of Autonomous Agents and Multi-Agent Systems 7(1/2), 179–189 (2003)
9. Gmach, D., Krompass, S., Scholz, A., Wimmer, M., Kemper, A.: Adaptive quality of service management for enterprise services. ACM Transactions on the Web 2(1), 1–46 (2008)
10. Scerri, P., Pynadath, D.V., Tambe, M.: Towards adjustable autonomy for the real world. Journal of Artificial Intelligence Research 17 (2003)
11. Müller, J.P.: The Design of Intelligent Agents: A Layered Approach. Springer, Heidelberg (1996)

12. Martin, D., Brustein, M., Hobbs, J., Lassila, O., McDermott, D., McIlraith, S., Narayanan, S., Paolucci, M., Parsia, B., Payne, T., Sirin, E., Srinivasan, N., Sycara, K.: OWL-S: Semantic markup for web services. W3C Member Submission, November 22 (2004)
13. de Brujin, J., Bussler, C., Dommingue, J., Fensel, D., Hepp, M., Keller, U., Kifer, M., KonigRies, B., Kopecky, J., Lara, R., Oren, H.L.E., Polleres, A., Roman, D., Scicluna, J., Stollberg, M.: Web service modeling ontology (WSMO). W3C Member Submission, June 3 (2005)
14. Zeng, L., Benatallah, B., Ngu, A.H., Dumas, M., Kalagnanam, J., Chang, H.: Qos-aware middleware for web services composition. IEEE Transactions on Software Engineering 30(5), 311–327 (2004)
15. Hayashi, H., Cho, K., Ohsuga, A.: Speculative computation and action execution in multi-agent systems. In: Proc. of Computational Logic in Multiagent Systems (CLIMA-III), pp. 136–148 (2002)
16. Barish, G., Knoblock, C.A.: Speculative execution for information gathering plans. In: Proc. of AIPS 2002, pp. 184–193 (2002)
17. Satoh, K., Inoue, K., Iwanuma, K., Sakama, C.: Speculative computation by abduction under incomplete communication environments. In: Proc. of the 1st International Conference on MultiAgent Systems(ICMAS 2000), pp. 263–270 (2000)
18. Satoh, K., Yamamoto, K.: Speculative computation with multi-agent belief revision. In: Proc. of the 1st International Conference on Autonomous Agents and Multi-Agent Systems (AAMAS 2002), pp. 897–904 (2002)
19. Satoh, K., Codognet, P., Hosobe, H.: Speculative constraint processing in multi-agent systems. In: Lee, J.-H., Barley, M.W. (eds.) PRIMA 2003. LNCS (LNAI), vol. 2891, pp. 133–144. Springer, Heidelberg (2003)
20. Satoh, K.: Speculative computation and abduction fro an autonomous agent. IEICE Transactions on Information and Systems E88-D(9), 2031–2038 (2005)
21. Takabayashi, Y., Niwa, H., Taneda, M., Fukuta, N., Yamaguchi, T.: Managing many web service compositions by task decomposition and service quality evaluation. In: Reimer, U., Karagiannis, D. (eds.) PAKM 2006. LNCS (LNAI), vol. 4333, pp. 291–302. Springer, Heidelberg (2006)
22. Minegishi, S., Fukuta, N., Iijima, T., Yamaguchi, T.: Acquiring and refining class hierarchy design of web application integration software. In: Karagiannis, D., Reimer, U. (eds.) PAKM 2004. LNCS (LNAI), vol. 3336, pp. 463–474. Springer, Heidelberg (2004)
23. Fukuta, N., Osawa, T., Iijima, T., Yamaguchi, T.: Semantic service integration support for web portal. In: Proc. of IEEE/WIC/ACM International Conference on Web Intelligence (WI 2005), pp. 161–164 (2005)
24. Havens, B., Goebel, R., Berger, J., Proulx, R.: A constraint optimization framework for multi-agent anytime scheduling. In: Proc. of AAAI 1999 (1999)

Context Model Based CF Using HMM for Improved Recommendation

Jong-Hun Kim[1], Chang-Woo Song[1], Kyung-Yong Chung[2], Un-Gu Kang[3], Kee-Wook Rim[4], and Jung-Hyun Lee[1]

[1] Dept. of Computer Science Engineering, Inha University
[2] School of Computer Information Engineering, Sangji University
[3] Dept. of Information Technology, Gachon University
[4] Dept. of Computer and Information Science, Sunmoon University, South Korea
ddcome@nate.com, up3125@hotmail.com, dragonhci@hanmail.net, ugkang@gachon.ac.kr, jhlee@inha.ac.kr, rim@sunmoon.ac.kr

Abstract. Users in ubiquitous environments can use dynamic services whenever and wherever they are located because these environments connect objects and users through wire and wireless networks. Also, there are many devices and services in these environments. However, it is difficult to effectively use conventional filtering method of the recommendation system in future ubiquitous environments because it does not reflect context information well in these environments. This paper attempt to define context model and propose new Collaborative Filtering (CF) based on Hidden Markov Models (HMMs) that are trained by context information. The Collaborative Filtering using HMMs (CFH) is suited to a user's interests and preferences. The Ubiquitous Recommendation System (URS) used in this study based on CFH uses an Open Service Gateway Initiative (OSGi) framework to recognize context information and connect device in smart home.

Keywords: Recommendation System, Collaborative Filtering, Ubiquitous Computing.

1 Introduction

A ubiquitous computing environment is a world that obtains required information without restrictions to time and location by connecting a large number of intelligent computers to various wire and wireless networks. Within this ubiquitous environment, it is easy to listen to music regardless of location using a variety of devices. Furthermore, the necessity for a service that recommends appropriate music requested by users has increased because users cannot afford to investigate every file among such a vast number of music files [1].

In recent years, a recommendation system has been actively studied that predicts and recommends only information requested by a user. Various filtering methods [2][3] are used in this recommendation system in order to estimate a user's preferences.

In this recommendation system, the similarity between the content of an item and user information was measured to recommend information desired by the user, and a

T. Yamaguchi (Ed.): PAKM 2008, LNAI 5345, pp. 268–279, 2008.

content-based filtering method that based the rank on this measurement was also used. However, the recommendation of multimedia data is still limited [4] and is not highly reliable due to filtering only being based on static information.

This study proposes a Collaborative Filtering (CF) method that is able to effectively recommend items in smart home environments. It determines a context model and establishes item Hidden Markov Models (HMMs) using context information. In this Collaborative Filtering using HMMs (CFH), the preference of items can be estimated using item HMMs in the case of the absence of item preferences. In addition, the collaborative filtering is performed using the estimated preference and that recommends items to users. Thus, it is possible to recommend items, which represent a high satisfaction level for users, using this filtering method because it well reflects the dynamic information of users in ubiquitous environments rather than that of the conventional filtering method.

Also, this paper attempted to design a Ubiquitous Recommendation System (URS) that is able to recommend music according to a user's interests and conditions by applying context information to a Hidden Markov Models (HMMs) and using Collaborative Filtering (CF). The recommendation system used in this study uses an Open Service Gateway Initiative (OSGi) framework to recognize context information and connect device in smart home.

2 Personalized Recommendation System

A personalization process selects content based on a user's profiles and new and unexpected items of the user. It can be treated as a process that finds certain knowledge and rules to personalize it using various statistics, analysis, comparison, and data mining methods for the particular characteristics of users. A recommendation system is required to perform this personalization.

The recommendation system used at the present time has been largely used in electronic commerce sites. In addition, the filtering methods used in this recommendation system are rule-based filtering, learning agent, content-based filtering, and collaborative filtering methods.

A rule-based filtering method identifies the preference of users and provides recommendations by connecting items to the characteristics of that item. The rule used in this method is predefined by experts. This method is effective with an item that has a high unit price and complicated characteristics.

A content-based filtering method keeps the information related to items and recommends these items to the user who inputs a keyword, which is related to the property of the information. This method has been largely used in the field of information search. This method includes some of the following disadvantages. First, the recommendation presents beneficial results when there is enough information unless the reliability of the recommendation is lowered. Second, the recommended item tends to show a certain specialized manner regardless of user preferences. Lastly, the quality of the recommendation is varied by the number of a user's opinion in which the collection of a user's opinion becomes a burden to the user.

A collaborative filtering method establishes a database according to the item preference of users and searches neighbors that have similar preferences when a new

customer appears. In addition, the item that a neighbor is interested in is recommended to the customer under the assumption that the customer will be interested in the item. Thus, this method can compensate the problem that exists in a content-based filtering method. However, a recommendation system that uses a collaborative filtering method and was evaluated as a successful system in certain studies and actual fields presented low quality in recommendations when users who have different preferences compared to a customer were few. In addition, this method had problems with recommendations for newly given recommendations before anyone could evaluate that recommendation.

3 Configuration of Context Model

This system defines contexts to recommend item by considering surrounding contexts and user information. The definition of contexts created by Brown [5], defined information for a users' location or surroundings. Brown's definition is an accurate method to develop application services, used to configure and determine proper context for an item recommendation service in this system.

The configuration of context Model for the Ubiquitous Recommendation System (URS) consists of user information (sex, age), time, pulse, weather, temperature, and location information.

Information on user, time, and location can be used to provide active services in smart home environments and analyze users' action patterns. Also, information on temperature, weather, and pulse can be assumed as context information that affects item selections.

Table 1 presents the definition of context model as different spaces, such as class 5 for pulse, class 4 for temperature, class 7 for weather, and class 6 for location information.

Table 1. Configuration of Context model

Context	Class(num.)
Sex	Male, Female
Age	Infant(0~7), Child(8~11), Young Adult(12~17), Adult(18~61), Old Adult(62~)
Time	Time(1~24)
Location	Balcony(10), Bathroom(20), Bedroom(30), Guestroom(40), Kitchen(50), Living-room(60)
Weather	Clear(10), Sunny(20), Cloudy(30), Shower(40), Rain(50), Snow(60), Storm(70)
Temperature	Cold(0~30), Cool(32~68), Warm(69~86), Hot(87~)
Pulse	Danger(0~59), Low(60~79), Normal(80~130), High(131~180), Danger(181~)

Weather data can be established as a database retrieved from the Internet. Temperature data can be transferred using an OSGi framework and communication from a sensor used in wireless communication. User information and location information can be traced using an RFID-Tag which is attached to a user's watch, and pulse data can be obtained from a pulse sensor which is attached to a user's watch through real-time Zigbee communication.

4 Hidden Markov Model for Item

Hidden Markov Model (HMM) [6] is a statistical pattern recognition method. HMM is an algorithm that was founded on mathematics. It was introduced in the field of speech signal processing in 1975 and widely applied from isolated word recognition to the spontaneous speech recognition. This algorithm can be classified as a learning and recognition process under the assumption that feature vectors is modeled after the Markov process.

This study applies an HMM as a context pattern recognition method of item according to user's context information. The Baum-welch method is used as a learning method for the HMM. In addition, probability for the HMM is calculated using a Vitervi algorithm.

The parameters used in the HMM consist of the transition probability between states, output probability subordinated to states, and initial presence probability of states. The parameter of the HMM can be simply expressed as Equation (1).

$$\lambda =< A, B, \pi > \tag{1}$$

A : State-transition probability distribution
B : Observation symbol probability distribution
(Trained to context information)
π : Initial state distribution

This study applied a modified Bakis model that included five different states as illustrated in Fig. 1 in order to express (1).

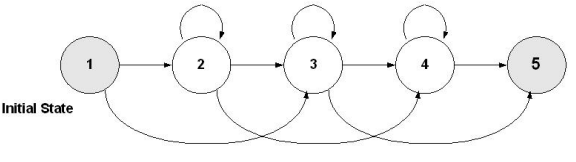


Fig. 1. HMM Topology for a Item

The major characteristics of the HMM topology can be determined as five STATEs, First-Order-Markov Chain model, and self migration potential and next state potential probability for each STATE.

Each item consists of HMMs that are named by item(number).hmm trained by context information. The Training Context Sequence (TCS) to practice models consists of the information of {location, time, weather, temperature, pulse}. The

Baum-Welch algorithm obtains optimum parameter values by repeating specific steps that estimate parameters and maximize the value of P(TCS|λ), which is the likelihood value of the given sequence (TCS). The computation of P(OCS|λ) value can be applied by using forward-backward processes. An observation context sequence (OCS) make according to context of user.

Fig. 2 illustrates the internal probability of the "Item32" HMM, a type of item.

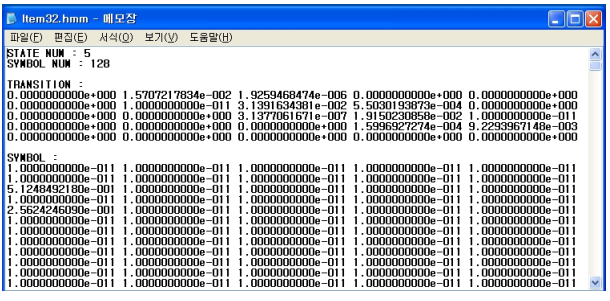


Fig. 2. Five States and 128 Observation Symbols of HMM

The accuracy of the HMM and recognition algorithm was tested on 25 items used in computer applications. Fig. 3 shows the selection of the item that exhibited the highest probability among 25 items by applying a context sequence from the context sequence used to training the HMM.

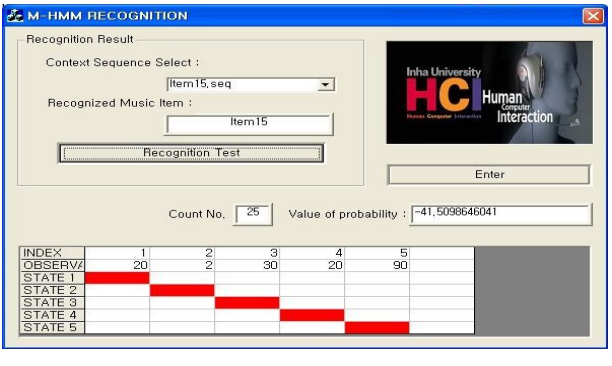


Fig. 3. Model Recognition Application for Selection of the HMM

5 Collaborative Filtering Using Hidden Markov Model

The Collaborative Filtering using Hidden Markov Model (CFH) selects items for a user based in the opinions of other users that have similar environment. CFH considers every user as an expert for his taste, so that personalized recommendations can be provided based on the expertise of taste-related users. CFH system collects the user opinions as ratings on a numerical scale, leading to {*user*, *item*} item rating

matrix, $r_{u,i}$. Table 2 shows the item rating matrix of {user, item}. Users evaluate the degree of preferences for items in a collaborative filtering process used to recommend items. The rating of preferences can be classified as six steps ranged from 0 to 1.0 with a 0.2 interval.

Table 2. The Item Rating Matrix of {user, item}, $r_{u.i}$

	U_1	U_2	….	U_m	….	U_u
$Item_1$	$r_{1,1}$	$r_{2,1}$	….	$r_{m,1}$	….	$r_{u,1}$
$Item_2$	$r_{1,2}$	$r_{2,2}$	….	$r_{m,2}$	….	$r_{u,2}$
….	….	….	….	….	….	….
$Item_n$	$r_{1,n}$	$r_{2,n}$	….	$r_{m,n}$	….	$r_{u,n}$
….	….	….	….	….	….	….
$Item_i$	$r_{1,i}$	$r_{2,i}$	….	$r_{m,i}$	….	$r_{u,i}$

Table 3 represents the example of the preference of items evaluated by users in which the preference of the item, i, which is not evaluated by users, is marked by using "*".

Table 3. The Item Rating Matrix of {user, item}, Which is Not Evaluated by Users

	U_1	U_2	….	U_m	….	U_u
$Item_1$	0.6	0.2	….	0.8	….	0.4
$Item_2$	1.0	0.6	….	*	….	*
….	….	….	….	….	….	….
$Item_n$	0.2	*	….	0.8	….	0.2
….	….	….	….	….	….	….
$Item_i$	0.6	0.4	….	*	….	0.4

CFH then uses this {*user*, *item*} item rating matrix and {*user*, *item*} item HMM matrix, $h_{u,i}$ in order to derive predictions. Table 4 shows Hidden Markov Model that trained according to context information of user.

Table 4. The Item HMM Matrix of {user, item}, $h_{u,i}$

	U_1	U_2	….	U_m	….	U_u
$Item_1$	$h_{1,1}$	$h_{2,1}$	….	Not make	….	$h_{u,1}$
$Item_2$	$h_{1,2}$	Not make	….	$h_{m,2}$	….	Not make
….	….	….	….	….	….	….
$Item_n$	$h_{1,n}$	Not make	….	$h_{m,n}$	….	$h_{u,n}$
….	….	….	….		….	….
$Item_i$	$h_{1,i}$	$h_{2,i}$	….	Not make	….	$h_{u,i}$

The conventional collaborative filtering method uses the items that are commonly evaluated by users, whereas {*user, item*} merge matrix, $m_{u,i}$ are used to evaluate the items for recommend items.

In the case of the item recommendation using a CFH method, it is possible to recommend items by considering the context of users because it produces not only item rating matrix of {*user, item*} but also {*user, item*} merge matrix using the item HMM matrix of {*user, item*}.

It is necessary to generate merge vectors in order to produce {user, item} merge matrix. The merge vector $m_{u,\mathrm{i}}$ can be defined as Equation (2). The merge vector consists of the estimated rating using {*user*, *item*} item HMM matrix for the rating, which is actually evaluated by users, and the items, which are not evaluated by users in which $r_{u.i}$ is the rating actually evaluated by users and $hw_{u,i}$ is the estimated rating using {*user*, *item*} item HMM matrix and {*user*, *item*} item rating matrix when there are no items evaluated by users.

$$m_{u,i} = \begin{cases} r_{u,i} & (r_{u,i}\, is\, value) \\ hw_{u,i} & (else\, undefined) \end{cases} \tag{2}$$

CF algorithms predict the preference based on the rating of similar users. When Pearson correlation coefficient is used, similarity is determined from the correlation of the rating vectors of user u and the other users a by $w(u,a)$. It can be noted that $w(u,a) \in [-1,+1]$. The value of $w(u,a)$ measures the similarity between the two users' rating vectors.

$hw_{u,\mathrm{i}}$ is the rating estimated by using the item HMM when a user, u, did not evaluate the item, i. The estimation can be performed using {user, item} item HMM matrix, $h_{u,\mathrm{i}}$. If there is no $r_{u.i}$, the $\{P(OCS|\lambda)_{1(>average\ of\ w(u,a)),i} \cdots\cdots P(OCS|\lambda)_{m(>average\ of\ w(u,a)),i}\}$ of $\{h_{1(>average\ of\ w(u,a)),i} \cdots\cdots h_{m(>average\ of\ w(u,a)),i}\}$ in a {user, item} item HMM matrix is to be calculated. Also, the within 30% HMMs which have high value in the values of P(OCS|λ) are selected. Then, the preference of the $hw_{u,\mathrm{i}}$ can be estimated using the average value of $r_{selected\ users,i}$ that corresponds to $h_{selected\ users,i}$.

The {*user, item*} merge matrix are also used to calculate the new weight of similarity in a CFH method. Table 5 shows the merge matrix of {user, item}

Table 5. The Merge Matrix of {user, item}, $m_{u,\mathrm{i}}$

	U_1	U_2	….	U_m	….	U_u
$Item_1$	0.6	0.2	….	0.8	….	0.4
$Item_2$	1.0	0.6	….	0.44	….	0.87
….	….	….	….	….	….	….
$Item_n$	0.2	0.67	….	0.8	….	0.2
….	….	….	….	….	….	….
$Item_i$	0.6	0.4	….	0.59	….	0.4

The general predict formula is based in the assumption that the prediction is a weighted average of the other users rating. The weights refer to the amount of similarity between the user u and the other users by Equation (3). $nw(u,a)$ is new weight of similarity using {*user, item*} merge matrix.

$$p^{collab}(u,i) = \overline{m_u} + k\{(m_{u,i} - \overline{m_u}) + \sum_{a \in U_i} nw(u,a)(m_{a,i} - \overline{m_a})\}$$

$$k = \frac{1}{\sum_{a \in U_i} nw(u,a)} \tag{3}$$

6 Design of Ubiquitous Recommendation System

This section designed and implemented the Ubiquitous Recommendation System (URS) that was able to recommend proper item by using Collaborative Filtering based on Hidden Markov Model (CFH) in a Java-based OSGi framework. Fig. 4 presents the diagram of the overall system. The system used in this study consisted of three large sections; a Context Agent section that recognizes context information, Service Agent section that continuously supports service even though the user moves to another location or a user device changes, and Item Recommendation Agent that recommends an item list using a CFH method based on the context information obtained in the Context Agent.

The Context Agent searches and collects the data sensed from the Data Crawler in various device environments, which are connected by OSGi. The Context Interpreter that received the data from the Date Crawler performs a context-aware process and transfers the context information for various home services. In the Item Context Information Finder, it transforms the transferred context information into information sequences, which are required in the Item Recognition Agent, and transfers it to the Service Agent.

The Service Agent consisted of a bundle service that provided item recommendation service as a bundle in a Simple Object Access Protocol (SOAP) Service, OSGi framework installed device in order to transfer information received from the Item Context Information Finder to the Item Recommendation Agent, and a Smart Home Application and Agent Manager Service that supported the management of the mobility of bundles. Communication between the Context Agent, Item Recommendation Agent, and a Smart Home Application can be performed using the SOAP Service. This service makes possible a real-time process in item recommendation service using the context information between different systems. The Agent Manager Service automatically manages bundles and supports service mobility. Service mobility in a URS means that service is not interrupted by a different device that has an OSGi framework even though a user's location has changed.

The Item Recommendation Agent produces optimum probability values for the context observation sequence for each item row of HMM matrix using the context observation sequence transferred from the Service Agent and makes {user, item} merge matrix for CFH. The viterbi algorithm is used to produce the optimum probability value. The CFH method based the preference for a neighbored item with similar context. The CFH used for recommendation employed a Content Information DB, Rating Matrix, and Item HMM Matrix.

The system proposed in this study developed an OSGi framework using the Knopflerfish 1.3.3, an open architecture source project which implemented a service framework.

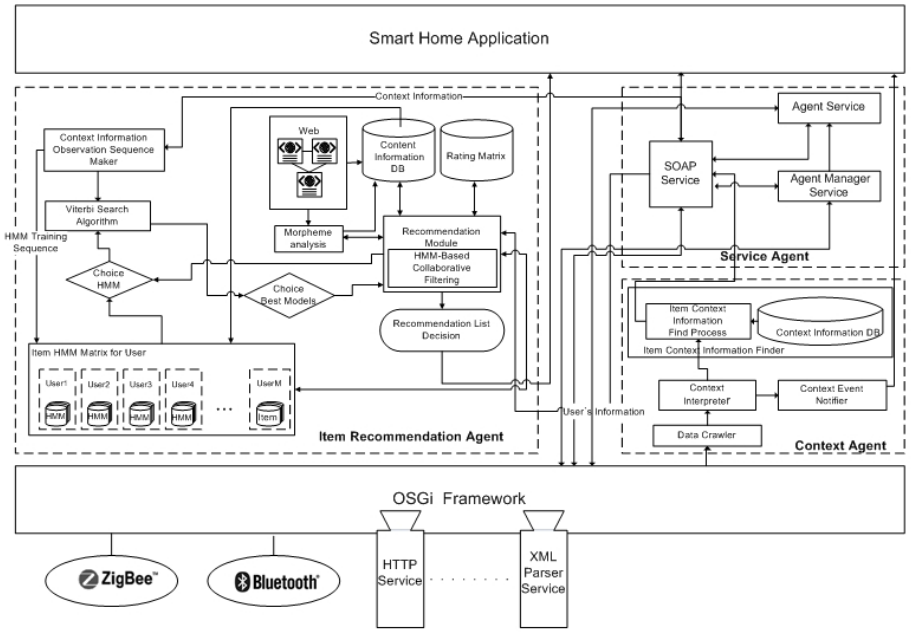


Fig. 4. The URS based on CF using HMM

7 System Evaluation

This study established evaluation data based on the results of the survey performed through off- and on-line in order to verify the effectiveness and validity of the proposed system. By 25 users, the evaluation data was used to evaluate the specific satisfaction for the recommended music item list ranged from 1 to 5 with the interval of 1 using the proposed system. Whereas, in the number ranged from 1 to 5, 1 means very negative evaluation and 5 shows very positive evaluation. The survey collected 300 evaluation data for six days.

The performance evaluation analyzes the Ubiquitous Music Recommendation System (UMRS) using Collaborative Filtering based on HMM (CFH) and the distribution of the service evaluation data applied in the existing systems that apply the collaborative filtering. It evaluates the applicability of the UMRS using a T-test.

The T-test is a method that uses t-distribution in a statistical verification process. The t-distribution shows bilateral symmetry like normal distribution and changes in the peak of the distribution according to the number of cases. Also, the T-test can be used to verify a possible difference in average values between two target groups. In addition, it classifies the groups as the case of independent sampling and dependent sampling.

This study applied a dependent sampling T-test to verify the satisfaction of services as a statistical way for the difference between the Ubiquitous Music Recommendation System (UMRS) and the Music Recommendation System using Collaborative

Table 6. UMRS / CFMRS Paired Samples Statistics

System	Mean	N	Std. Deviation	Std. Error Mean
UMRS	3.3800	150	1.00114	.08174
CFMRS	2.9533	150	.91473	.07469

Table 7. UMRS / CFMRS Paired Samples Test

Paired Variables	Paired Samples Test				t	df	Sig. (2-tailed)
	Mean	Std. Deviation	95% Confidence Interval of the Difference				
			Lower	Upper			
UMRS / CFMRS	.42667	.49625	.34660	.50673	10.530	149	.000

Filtering (CFMRS) for 25 users in the survey. The hypothesis, H_0, is that "there are no statistical differences in the satisfaction of the UMRS and CFMRS services", and the hypothesis, H_a, is that "there are certain statistical differences in the satisfaction of the UMRS and CFMRS services". Table 6 demonstrates the mean and standard deviation of the evaluation data for the satisfaction of the UMRS and CFMRS services. The difference of the average values was 0.4267.

Table 6, 7 shows the results of the T-test for the paired samples of UMRS and CFMRS. As the significant level of α was 0.05, the critical value for making a decision was $t < 0.34660$ or $t > 0.50673$. Because the values of t was presented as $10.530 > 0.50673$ in the T-test of the evaluation data in this paper, Thus, "there are certain statistical differences in the satisfaction of the UMRS and CFMRS" was accepted. Also, it was verified that the satisfaction of the UMRS service showed a high level of 0.4267, which is the difference in the average value of the evaluation data, compared to the CFMRS service.

8 Conclusions and Future Work

In recent years, life patterns have been changed by ubiquitous infrastructures due to the development of ubiquitous computing technologies and that lead to a new culture. The ubiquitous technologies support not only the conventional methods that provide services based on customers' requests, but a new system that provides services as an active manner by investigating the intention and circumstance of users. This new system can be considered as a ubiquitous service.

This paper proposes a new Collaborative Filtering method using Hidden Markov Models (CFH) that can be used to such ubiquitous environments by reflecting users'

context information and designs Ubiquitous Recommendation System (URS) based on CFH.

It is possible to provide dynamic and active services to users because this system was designed based on agents in OSGi framework.

This study determined a context model of users and modeled the relationship between the context of users and items using the HMM. It was possible to recommend items, which reflect the preferences of users, with common contexts by using the HMM to the estimation of the preferences that are not evaluated by users.

In the results of the system evaluation, the method proposed in this study increased the satisfaction of users compared to that of the existing methods, such as the Ubiquitous Music Recommendation System (UMRS) and the Music Recommendation System using Collaborative Filtering (CFMRS).

An URS that considers the context patterns of users including users' interests will be developed for future studies. In order to achieve the system, it is necessary to determine a proper context model for ubiquitous environments and perform context inferences for recommending items.

Acknowledgment

"This research was supported by the MKE(Ministry of Knowledge Economy), Korea, under the ITRC(Information Technology Research Center) Support program supervised by the IITA(Institute of Information Technology Advancement)" (IITA-2008-C1090-0801-0020).

References

1. Kim, J.H., Jung, K.J., Lee, J.H.: Hybrid Music Filtering for Recommendation Based Ubiquitous Computing Environment. In: Greco, S., Hata, Y., Hirano, S., Inuiguchi, M., Miyamoto, S., Nguyen, H.S., Słowiński, R. (eds.) RSCTC 2006. LNCS (LNAI), vol. 4259, pp. 796–805. Springer, Heidelberg (2006)
2. Balabanovic, M., Shoham, Y.: Fab: Content-based, Collaborative Recommendation. Communication of the Association of Computing Machinery 40(3), 66–72 (1997)
3. Jung, K.Y., Lee, J.H.: User Preference Mining through Hybrid Collaborative Filtering and Content-based Filtering in Recommendation System. IEICE Transaction on Information and Systems E87-D(12), 2781–2790 (2004)
4. Chen, H.–C., Chen, A.L.P.: A music recommendation system based on music data grouping and user interests. In: Proc. of the CIKM 2001, pp. 231–238 (2001)
5. Brown, P.J., Bovey, J.D., Chen, X.: Context-Aware Application: From the Laboratory to the Marketplace. IEEE Personal Communication, 58–64 (1997)
6. Rabiner, L.R.: A Tutorial on Hidden Markov Models and Selected Application in Speech Recognition. Proc. IEEE 77(2), 257–286 (1989)
7. Breese, J.S., Heckerman, D., Kadie, C.: Empirical Analysis of Predictive Algorithms for Collaborative Filtering. In: Proc. of the 14th Conference on Uncertainty in AI (1998)
8. Herlocker, J., et al.: An Algorithm Framework for Performing Collaborative Filtering. In: Proc. of ACM SIGIR 1999 (1999)

9. Resnick, P., et al.: GroupLens: An Open Architecture for Collaborative Filtering of Netnews. In: Proc. of ACM CSCW 1994, pp. 175–186 (1994)
10. Jung, K.Y., Lee, J.H.: Prediction of User Preference in Recommendation System using Association User Clustering and Bayesian Estimated Value. In: McKay, B., Slaney, J.K. (eds.) Canadian AI 2002. LNCS (LNAI), vol. 2557, pp. 284–296. Springer, Heidelberg (2002)
11. Dobrev, P., Famolari, D., Kurzke, C., Miller, B.A.: Device and Service Discovery in Home Networks with OSGi. IEEE Communications Magazine 40(8), 86–92 (2002)
12. Bellavista, P., Corradi, A., Stefanelli, C.: Mobile Agent Middleware for Mobile Computing. IEEE Computer 34(3) (2001)
13. Liu, T., Martonosi, M.: Impala: A Middleware System for Managing Autonomic, Parallel Sensor Systems. In: ACM SIGPLAN Symp. Principles and Practice of Parallel Programming (2003)
14. Gu, T., Pung, H.K., Zhang, D.Q.: An Ontology-based Context Model in Intelligent Environments. In: Proc. of Communication Networks and Distributed Systems Modeling and Simulation Conference, pp. 270–275 (2004)
15. Romer, K., Schoch, T., Mattern, F., Dubendorfer, T.: Smart Identification Frameworks for Ubiquitous Computing Application. In: IEEE International Conference on Pervasive Computing and Communication (2003)

Author Index