Shlomo Argamon
Newton Howard (Eds.)

# Computational Methods for Counterterrorism

Springer

# Computational Methods for Counterterrorism

Shlomo Argamon · Newton Howard
Editors

# Computational Methods for Counterterrorism

Springer

*Editors*
Shlomo Argamon
Illinois Institute of Technology
Dept. of Computer Science
10 West 31st Street
Chicago IL 60616
USA
argamon@iit.edu

Newton Howard
Center for
Advanced Defense Studies
10 G St. NE, STE 610
Washington, DC 20002
USA
newton.howard@c4ads.org

ISBN 978-3-642-01140-5 e-ISBN 978-3-642-01141-2
DOI 10.1007/978-3-642-01141-2
Springer Dordrecht Heidelberg London New York

Library of Congress Control Number: 2009926862

ACM Computing Classification (1998): H.3, I.5, I.7, K.4, K.6.5



Printed on acid-free paper

Springer is part of Springer Science+Business Media (www.springer.com)

Dedicated to the victims of modern terrorism,
in the hope of saving others in the future.

# Foreword

In the summer of 2001, I was working at the Defense Advanced Research Projects Agency and exploring how software technologies could be used in the counterterror arena. At the time, DARPA did not have a primary role in counterterrorism, and most of the related work that was going on at the agency focused on sensing and countermeasures for chemical and biological threats, with little emphasis on software. Information Security was also certainly a concern, but the focus of that work was on detecting compromised systems and not on combating terrorism. As a result, my report was not getting a lot of interest, and when I left the agency at the end of my stint, on September 1st 2001, I was pretty sure that this work would get little or no attention. I was wrong.

Unfortunately, less than two weeks later, our country discovered that terrorism was a far greater threat than we thought, and that we were vulnerable to threats that we were not tracking. It was not until several years later that the 9/11 commission reported that information sharing had been a major contributor to our intelligence failure in the World Trade Tower attacks; however, to many in the intelligence community this fact had become clear much sooner. The phrase "connect the dots" was overused to the point of triteness, but for those who were looking at how to increase our nation's awareness of potential terror plots, it became a driving need. To push a metaphor, we realized we were not looking for a needle in a haystack, we were looking for many bits of needles hidden in haystacks full of metal by a dedicated adversary. It was clear that new techniques needed to be developed that would go well beyond those already deployed in the community.

One of the areas of emphasis of my DARPA work, unfortunately validated by the attack, was that the community needed not only their own sources of vetted information, but a much more comprehensive ability to track information from open sources. Furthermore, it was clear that no single "INT" (sigint, humint, etc.) was going to be as powerful as what could be learned if we could break down the silos and create capabilities to track information from multiple sources. New technologies beyond traditional keyword searches

had to be made available in the open source world, provenance and tracking information needed to be available both in open and protected systems, and the ability to visualize and analyze relationships needed to be significantly improved.

The book you are reading is a good example of how this research is proceeding. The papers focus on better search and text analytics, graphical modeling and group analysis, and the direct modeling of potential conflicts. Some of the papers report on systems that are ready for transition, others on the techniques that will power next generation applications. The importance of these systems continues to grow, and I welcome this book, and others that are sure to follow, in our ongoing effort to support the information analysis and sharing needs for counterterrorism efforts.

Tetherless World Senior Constellation Professor James A. Hendler
Rensselaer Polytechnic Institute
New York

# Preface

The attacks against the World Trade Center and the Pentagon on September 11, 2001 initiated a new phase in the global struggle against terrorist networks, often called the "War on Terror". Despite objections to this term, there is an important sense in which it is quite apt—the rise of well-trained religiously-motivated multinational terrorist networks and armed groups poses a novel and unprecedented threat to the modern international order (Arquilla and Ronfeldt 2001; Shultz Jr. 2005). Indeed, the nature of the evolving threat from these terrorist networks has profound implications for governmental action at all scales (Howard 2009). These developments are reshaping foreign and defense policies together with the military and intelligence doctrines intended to support these policies (National Commission on Terrorism 2000; DeRosa 2004). Importantly, they are also driving the development of new technologies to enable effective implementation of these policies in the field.

This book presents a range of current research on computational models and methods that can be used in the fight against modern multinational terrorist networks. The threat that they pose differs radically from previous threats to international security primarily due to several key factors:

- Modern terrorist organizations have no fixed territorial homes, are organized in complex non-hierarchical networks, and pursue highly adaptable goals (Dishman 2005; Howard 2009). The fluid nature of these networks makes it difficult to identify enemy agents and their possible targets effectively, as well as to track or predict enemy actions.
- These networks are fundamentally non-state actors, with religious and ideological, rather than territorial or economic, agendas. This makes it very hard to determine their strategic aims and hence to predict their likely behavior or to devise effective countermeasures.
- The high power of relatively inexpensive and easily available weaponry makes it possible for terrorist cells to operate effectively with a small operational footprint, and hence makes them even more difficult to track.

- The highly-interconnected nature of the modern communication infrastructure makes it easy today for widely distributed terrorist networks to exchange messages with little risk of detection, due to the extremely large volume of irrelevant information. Comparatively simple information-hiding schemes thus often suffice to protect terrorists' hidden communications.

For all of these reasons, information technology and computational modeling are now of central importance to national security doctrine and practice, primarily in intelligence (O'Connell 2005), but also for tactical and strategic assessment and planning. Accurate and timely intelligence is critical to fighting terrorism. The complexity and fluidity of the new threat environment, comprising multiple non-state actors organized in highly non-hierarchical networks and alliances, makes effective intelligence aggregation and analysis more difficult and more important than ever.

Thus, we require effective solutions to two fundamental problems: finding relevant information in truly vast collections of raw data (*information overload*) and discovering meaningful patterns made up of many data items, each meaningless on its own, but significant when taken together with the rest of the pattern (*data mining*).

Currently, analysts must laboriously sift through enormous amounts of structured and textual data to try and find meaningful connections between relationships, events, and activities to produce actionable intelligence. New theoretical and practical tools are needed to aid this process.

Rarely is it the case that an isolated piece of information is useful by itself. Usually, meaningful intelligence must be built from constellations of connected bits of information, each insignificant in itself, but together important. For example, in hindsight, we know that all nineteen September 11 hijackers were related, before the attack, to within 3 degrees of connection to various known individuals on the United States government terrorist watch list (DeRosa 2004). In several cases, multiple independent links connected different individuals. Thus, in principle, the information needed to find and stop the attackers was available beforehand.

However, even under the best of circumstances, doing so with existing tools would have been practically impossible. There are two main problems—first, efficiently searching for clusters of meaningful information within the enormous body of available data (both open source and classified), and second, distinguishing between those clusters that are indeed meaningful from the many that are not. It is certainly true that there are many individuals that were also linked to the 9/11 hijackers that were not involved in the attacks; such false positives, if not ruled out, would overwhelm any useful information found. Similar difficulties exist in finding useful information from enormous amounts of collected textual data; to the problem of filtering relevant from irrelevant information is added the difficulty of interpreting the meaning of free-form text, often in multiple languages.

In addition to aiding large-scale data analysis, computational models can help us reason more effectively in our engagement with the modern terrorist threat. Modern ideological/religious terrorist networks have goals that radically differ from the local political and economic ambitions of nation-states and the political terrorists of the 1970s and 1980s (Howard 2009). The "new terrorists" have the larger and more abstract goal of disrupting the international order, and thus a plethora of possible strategies and targets, which makes scenario prediction much more difficult. This problem is exacerbated by the central role of religious ideology in these networks, which makes it quite difficult for people outside such religious groups to understand and predict their actions (Cronin 2002). Computational models of adversarial planning and psychology can aid in exploring the implications of different models of enemy intentions. Furthermore, formal models of reasoning processes have the specific advantage of making explicit the assumptions and implications of the analytic process, and thus can greatly improve the quality of the final intelligence product.

The purpose of this book is to present current and far-reaching research on computational methods that can help solve these difficult problems, so that decision-makers and scientists can more effectively marshal efforts to develop new technologies to support counterterrorism. To this end, the work collected here is primarily basic research which will, it is hoped, soon lead to novel and useful applications.

The volume is an outgrowth of the Descartes Conference on Mathematical Models in Counterterrorism, held on September 28 and 29, 2006 at the United States Congress Rayburn House Office Building in Washington, DC. Chapters for the book were solicited from selected papers presented at the conference as well as from other researchers, and have been peer-reviewed. We have sought to include as wide a variety of relevant research as possible.

## Organization

*Computational Methods for Counterterrorism* is divided into four parts. The first part describes research on methods for providing effective access to relevant information buried in the enormous stores of textual and other data currently available online (both open source and classified). The second part of the book deals with the development of methods for analyzing and classifying digitized documents to extract useful information which can aid intelligence analysis. The third part of the book presents research on analyzing graphs and networks. These abstract mathematical methods offer new ways of processing intelligence information to discover hidden links and structures, as well as improving analysis of adversaries' goals and intentions. The fourth part of the book discusses models and software systems that allow for simulating and evaluating the implications of diverse real-world conflicts.

Part I, "Information Access," contains four chapters. Chapter 1, an invited chapter by keynote speaker Ophir Frieder, describes a prototype system for the novel problem of "complex document information processing." The problem is to effectively analyze and index information in real-world documents including text, graphics, handwritten markings, and so on. The author shows how an integrated approach to such an information processing problem can lead to a solution that is greater than the sum of its parts. Chapter 2, by Srinivasan and S. Srihari, discusses how document images can be retrieved by matching handwritten signatures in the documents. The method, based on applying conditional random fields to image-based features, is capable of effectively dealing with the presence of image noise and of irrelevant text overlapping signatures. Chapter 3, by Zhao, Santos, Nguyen, and Mohamed, discusses methods for text summarization, which can help analysts find and assimilate critical information quickly. The authors show how multi-document summarization can be improved by metrics that measure the diversity of the document set to be summarized. Chapter 4, by Knepper, Fox, and Frieder, describes a software toolkit that integrates multiple retrieval methods to enable adaptive retrieval, browsing, and visualization of search results. Such a tool can enable analysts to more easily find needed information and to visualize the relationships between retrieved data in a more useful fashion.

Part II, "Text Analysis," contains three chapters. Chapter 5, by R. K. Srihari, describes methods that can effectively discover hidden information in document collections by detecting links between concepts expressed in disparate texts. Such "unapparent information revelation" can help analysts find secret information about adversaries hidden in large open source document collections. Chapter 6, by Taghva, describes methods that automatically identify "sensitive unclassified" information in scanned documents, so that such information can be redacted before documents are made available to the public. Chapter 7, by Guidére, Howard, and Argamon, shows how textual search and analysis may be enhanced by proper understanding of certain semantic, pragmatic, and cultural aspects of language use by terrorists.

Part III, "Graphical Models," contains four chapters. Chapter 8, an invited chapter by keynote speaker Robert Haralick, describes the theory of dicliques, a network structure that can be interpreted as a sort of "functional module" in a network, such as a network of known associations between terrorists. Extracting dicliques from a given network can reveal its hidden structure, and suggest what unobserved connections between known entities may exist. Chapter 9, by Koester and Schmidt, demonstrates how a related method, formal concept analysis, can be used to find meaningful gaps in relational data sets such as those gathered in intelligence work. The authors demonstrate their approach on the analysis of the MIPT Terrorism Knowledge Base and on web mining. Chapter 10, by Lefebvre, develops an algebra of strategic choice within and among groups of interacting agents, based on the author's previous work on mathematically models of individual choice. The model extends game-theoretic constructs with psychological insights within a formal

graph-theoretical framework. Chapter 11, by Grice, Scavo, and McDaniel, reports on empirical validation of Lefebvre's algebraic psychological models, showing their validity in certain real-world situations.

Part IV, "Conflict Analysis," contains four chapters. Chapter 12, by Shearer and Marvin, presents methods for classifying instability patterns of nation-states that allow prediction of the development of significant conflicts or even state failure. The models applied include consideration of social, economic, and political features of the nation-states examined. Chapter 13, by Hendrickson, shows how reasoning about counterfactual questions involves several kinds of assumptions about what antecedent scenarios are possible and relevant—making these assumptions explicit is important for properly assessing analytic results. Chapter 14, by Braynov, discusses how extraction of a "coordination graph" from an integrated link analysis of an enemy network and its actions can be used to recognize and counter enemy plans. The formalism can also be used to distinguish between the roles of different enemy agents. Chapter 15, by Silverman, Bharathy, and Nye, describes a simulation game used for analyzing the development of ethno-political conflicts. The game may be played by human or software agents, and has been evaluated by correspondence testing against real-world conflict situations.

## Conclusions

The fight against multinational terrorism is not one that is likely to be won decisively any time soon. It is a long-term struggle in which the enemy is exceptionally adaptive and continually devises new tactics and strategies, and so we must constantly improve our methods of acquiring and analyzing intelligence. Methods such as those described in this volume promise to provide fundamentally new approaches to structuring, analyzing, and understanding information. Critical is the fact that these models help make explicit the assumptions necessary to draw conclusions, enabling analysts to better explore the effects of such assumptions on their analyses. As the role of computational models in counterterrorism will only grow in coming years, it is crucial that policymakers at all levels work to understand these methods, their potential, and their risks.

## Acknowledgments

There are many people whom I must thank for their roles in the preparation of this book. First, of course, are the scholars and researchers who contributed its chapters, without whom this volume would not have been possible. Special thanks are also due to the Center for Advanced Defense Studies for its support of this project, and to Dr. Gideon Frieder for chairing multiple sessions of the conference. Krista Butler's administrative work and Mark Atallah's linguistic

editing and formatting were also most helpful. Thanks also to Max Irishfrazin, whose highly professional and indefatiguable copyediting was indispensable to the quality of the finished volume. Finally, I owe a debt of gratitude to my wife, Stefanie, for her constant love and support, as well as her valuable help with this project.

Associate Professor of Computer Science
Illinois Institute of Technology
Chicago, IL

Shlomo Argamon

## References

Arquilla, J. and Ronfeldt, D. F. 2001. *Networks and Netwars: The Future of Terror, Crime, and Militancy*, Rand Corporation.

Cronin, A. K. 2002. Behind the curve: Globalization and international terrorism. Reprinted in R. D. Howard, R. L. Sawyer, & N. E. Bajema (eds.), *Terrorism and Counterterrorism: Understanding the New Security Environment* (3rd ed.), New York: McGraw-Hill, 2009.

DeRosa, M. 2004. *Data Mining and Data Analysis for Counterterrorism*, Washington, DC: CSIS Press.

Dishman, C. 2005. The leaderless nexus: When crime and terror converge. Reprinted in R. D. Howard, R. L. Sawyer, & N. E. Bajema (eds.), *Terrorism and Counterterrorism: Understanding the New Security Environment* (3rd ed.), New York: McGraw-Hill, 2009.

Howard, R. D. 2009. The new terrorism. Reprinted in R. D. Howard, R. L. Sawyer, & N. E. Bajema (eds.), *Terrorism and Counterterrorism: Understanding the New Security Environment* (3rd ed.), New York: McGraw-Hill, 2009.

National Commission on Terrorism, 2000. *Countering the Changing Threat of International Terrorism: Report of the National Commission on Terrorism*, Pursuant to Public Law 277, 105th Congress. Available from `http://www.fas.org/irp/threat/commission.html`

O'Connell, K. M. 2005. The role of science and technology in transforming American intelligence. In P. Berkowitz (ed.), *The Future of American Intelligence* (pp. 139–174), Hoover Institution.

Shultz Jr., R. H. 2005. The era of armed groups. In P. Berkowitz (ed.), *The Future of American Intelligence* (pp. 1–39), Hoover Institution.

# Contents

# List of Contributors

Shlomo Argamon
Department of Computer Science
Illinois Institute of Technology
Chicago, IL 60616
argamon@iit.edu

Gnana K. Bharathy
Electrical and Systems Engineering Department
University of Pennsylvania
Philadelphia, PA 19104-6315
bharathy@seas.upenn.edu

Sviatoslav Braynov
Department of Computer Science
University of Illinois at Springfield
Springfield, IL 62703
sbray2@uis.edu

Kevin L. Fox
Harris Corporation
GCSD, Hanover
Melbourne, FL 32902
kevin.fox@harris.com

Ophir Frieder
Information Retrieval Laboratory
Department of Computer Science
Illinois Institute of Technology
Chicago, IL 60616
ophir@ir.iit.edu

James Grice
Department of Psychology
215 North Murray Hall
Oklahoma State University
Stillwater, OK 74078-3064
james.grice@okstate.edu

Mathieu Guidère
Descartes Institute
Center for Advanced Defense Studies
10 G Street NE, Suite 610
Washington, DC 20002
mathieu.guidere@c4ads.org

Robert M. Haralick
Computer Science, Graduate Center
City University of New York
365 Fifth Avenue
New York, NY 10016
haralick@ptah.gc.cuny.edu

Noel Hendrickson
Institute for National Security Analysis & Information Analysis Program
James Madison University
Harrisonburg, VA 22807
hendrinx@jmu.edu

Newton Howard
Descartes Institute
Center for Advanced Defense Studies
10 G Street NE, Suite 610
Washington, DC 20002
newton.howard@c4ads.org

Margaret M. Knepper
Harris Corporation
GCSD, Hanover
Melbourne, FL 32902
margaret.knepper@harris.com

Bjoern Koester
Technische Universität
Dresden, Germany
bjoern.koester@webstrategy.de

Vladimir A. Lefebvre
School of Social Sciences
University of California at Irvine
Irvine, CA 92697
valefebv@uci.edu

Brett Marvin
Sentia Group, Inc.
Washington, DC
blm@sentiagroup.com

Brenda L. McDaniel
Department of Psychology
215 North Murray Hall
Oklahoma State University
Stillwater, OK 74078-3064
bmcdani@ksu.edu

Ahmed Mohamed
Department of Computer Science and Engineering
University of Connecticut
Storrs, CT 06269
ahmed@engr.uconn.edu

Hien Nguyen
Department of Mathematical and Computer Sciences
University of Wisconsin
Whitewater, WI 53190
nguyenh@uww.edu

Benjamin D. Nye
Electrical and Systems Engineering Department
University of Pennsylvania
Philadelphia, PA 19104-6315
benjamid@seas.upenn.edu

Eugene Santos, Jr.
Thayer School of Engineering
Dartmouth College
Hanover, NH 03755
Eugene.Santos.Jr@Dartmouth.edu

Stefan E. Schmidt
Technische Universität
Dresden, Germany
stefan@phoenixmath.com

Lt. Col. Robert M. Shearer
Naval Postgraduate School
Monterey, CA
rlsheare@nps.edu

Barry G. Silverman
Electrical and Systems Engineering Department
University of Pennsylvania
Philadelphia, PA 19104-6315
basil@upenn.edu

Rohini K. Srihari
Department of Computer Science & Engineering
University at Buffalo
Buffalo, NY 14228
rohini@cedar.buffalo.edu

Sargur Srihari
Center of Excellence for Document Analysis and Recognition
State University of New York at Buffalo
Buffalo, NY 14260
srihari@cedar.buffalo.edu

Harish Srinivasan
Center of Excellence for Document Analysis and Recognition
State University of New York at Buffalo
Buffalo, NY 14260
hs32@cedar.buffalo.edu

Kazem Taghva
Information Science Research Institute
University of Nevada, Las Vegas
taghva@isri.unlv.edu

Qunhua Zhao
Thayer School of Engineering
Dartmouth College
Hanover, NH 03755
Qunhua.Zhao@Dartmouth.edu

# Part I

# Information Access

# 1

# On Searching in the "Real World"

Ophir Frieder

**Summary.** For many, "searching" is considered a solved problem. Indeed, for text processing, this belief is factually based. The problem is that most "real world" search applications involve "complex documents," and such applications are far from solved. Complex documents, or less formally, "real world documents," comprise a mixture of images, text, signatures, tables, etc., and are often available only in scanned hardcopy formats. Search systems for such document collections are currently unavailable.

We describe our complex document information-processing prototype. This prototype integrates "point solution" (mature) technologies, such as optical character recognition, signature matching and handwritten word spotting techniques, logo detection and recognition, and search and mining approaches, to yield a system capable of searching "real world documents." The described prototype validates the adage that "the whole is greater than the sum of its parts." Our complex document benchmark development efforts are likewise presented.

Having discussed the core approach, we describe some additional point solutions developed at the Illinois Institute of Technology (IIT) Information Retrieval (IR) Laboratory. These include an Arabic stemmer and a natural language source integration fabric called the IIT Intranet Mediator. In terms of stemming, we developed and licensed an Arabic stemmer and search system. Our approach was evaluated using the Arabic TREC collection and favorably compared against the state of the art.

We also focused on source integration and ease of user interaction. By integrating structured, semi-structured, and unstructured sources, we developed and licensed our mediator technology that provides a single, natural language interface to querying distributed sources. Rather than providing a set of links as possible answers, the described approach actually answers the posed questions.

## Introduction

Search systems are often viewed as "mature technology," and indeed, this perception is correct to a limited degree. For example, computer users, from

S. Argamon, N. Howard (eds.), *Computational Methods for Counterterrorism*, DOI 10.1007/978-3-642-01141-2_1,

novice to savvy, daily rely and use the Google™, Yahoo!™, and other search engines. These web search engines are easy to use, highly reliable, and maintain indexes, and hence links, to potential references to a wide range of topics. Furthermore, many mature information retrieval engines are available to search computer generated text documents. These observations support the maturity perception; the problem is that a large portion of "real world" documents is neither indexed by web search engines nor is it available in a computer readable text format.

"Real world" documents indeed consist of text, but they often also include numerous artifacts other than text. For example, documents may include graphs, charts, images, signatures, logos, hand-written annotations, watermarks, and stamps. Clearly, these components, like the text component, contain relevant information, and this information should be made searchable. Present day search engines simply ignore these other component types.

Another characteristic of a large portion of "real world" documents is that the documents are not computer readable. That is, these documents are often available only in hardcopy rather than in any common electronic format. To provide them in an electronic format, the documents are scanned. However, their derived image format is often of degraded quality as scanners introduce a level of distortion. Furthermore, this process also assumes that the hardcopy is of reasonable quality to begin with, which itself is often a false assumption.

In the legal domain, often in the "discovery phase" of litigation, the defendant is required to produce all relevant records; these records are often printed charts, handwritten notes, annotations on printed text, and crumpled sheets of design documents. Scanning these documents, besides being costly and time intensive, often yields poor quality images that are relatively useless in terms of search even after processing them using optical character recognition (OCR) software. The same situation exists in the intelligence domain, historical document realm, medical arena, and many other environments. Even with distorted image restoration techniques, such as those found in Agam et al. (2007), the image quality, and hence character recognition accuracy, remains poor.

To support the searching of "real world" documents, our lab, the Illinois Institute of Technology (IIT) Information Retrieval (IR) Laboratory, collaborating with researchers from the State University of New York at Buffalo (SUNYAB) and the University of Maryland (UMD) and system integrators from Clarabridge, Inc., developed, in less than a year, a complex document information processing (CDIP) prototype (Argamon et al. 2006). This prototype, the first of its kind to the best of our knowledge, is based on the integration of multiple mature technologies (Chen et al. 2005; Srihari et al. 2004, 2006) working in unison to significantly improve the state of the art. Via this integration, the preliminary results obtained already prove the old adage that indeed "the whole is greater than the sum of its parts."

The parts in the context of CDIP are mature technologies, namely, optical character recognition, table extraction, logo recognition, handwritten

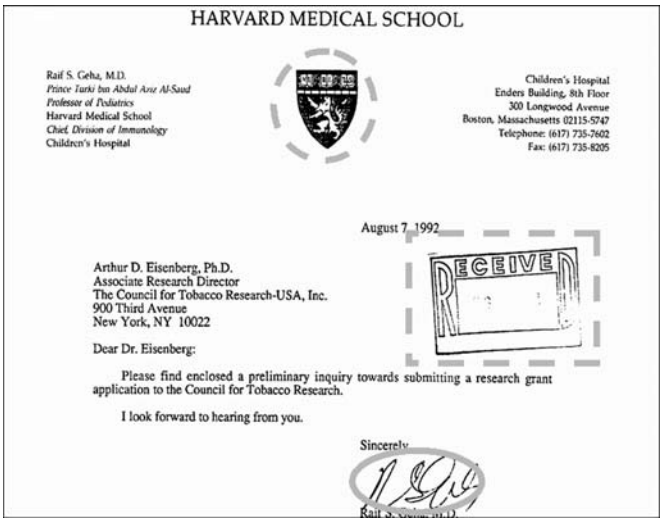
word spotting, signature matching, information retrieval, data mining, etc. Currently, the prototype consists of only the basic tools from within these mature "point solutions." In the near future, we plan to integrate additional functionality into our prototype. This additional functionality potentially includes commercialized versions of point solutions originally developed in the IIT IR Lab such as the Advanced Information Retrieval Engine – AIRE (licensed from IIT by Harris Corporation), the Arabic Stemmer (licensed from IIT by Blue Shoe Technologies), and the IIT Intranet Mediator (patented by and licensed from IIT by Intranet Mediator, Inc).

AIRE is a highly accurate search engine as shown by the independent industrial evaluation performed by Harris Corporation personnel prior to its licensing (Infantes-Morris et al. 2003). Due to space limitations, however, we forgo further discussion of AIRE. For a general text covering search techniques and optimizations such as those deployed in AIRE, see Grossman and Frieder (2004). As far as the Arabic stemmer and the intranet mediator, these efforts are unique. We therefore highlight their design and direct the interested reader to the corresponding original publications for greater detail.

## Complex document information processing

As stated earlier, complex or "real world" documents usually are available only in hardcopy. In addition to type-written text, they often also include logos, handwritten notes, signatures, and/or tables. At times, they have text in multiple languages. In Fig. 1.1, we illustrate a relatively clean complex document. This scanned image, part of the document collection made public through the Tobacco Master Settlement Agreement (Lewis et al. 2006), consists of printed text, a logo—outlined by a dashed oval, a signature—outlined by a solid oval, and a stamp—outlined by a dashed rectangle. As seen, this scanned image is relatively clean in that the text is type written, clear—that is, it is not distorted, and, for the most part, each component is separate—namely, there is no overlap.

In Fig. 1.2, we present another scanned image; this scanned document is part of the collection of documents seized during Operation Defensive Shield linking Arafat to terrorism (`http://www.mfa.gov.il`—available as of October 7, 2006). This document is significantly more complex than the one illustrated in Fig. 1.1. Here text is presented in multiple languages using multiple scripts—outlined by the solid squares. A logo is present—outlined, once again, by a dashed oval. There is also handwritten text, some of which is a signature, and some of which is adjacent to type written text—outlined by ovals. A table is present, and some of the handwritten notes are on top of the table structure. Further complicating the processing is that the image itself is somewhat distorted, making the processing and searching of this image even more complicated than the searching of the image presented in Fig. 1.1.



HARVARD MEDICAL SCHOOL

Raif S. Geha, M.D.
*Prince Turki bin Abdul Aziz Al-Saud*
*Professor of Pediatrics*
Harvard Medical School
*Chief, Division of Immunology*
Children's Hospital

Children's Hospital
Enders Building, 8th Floor
300 Longwood Avenue
Boston, Massachusetts 02115-5747
Telephone: (617) 735-7602
Fax: (617) 735-8205

August 7, 1992

RECEIVED

Arthur D. Eisenberg, Ph.D.
Associate Research Director
The Council for Tobacco Research-USA, Inc.
900 Third Avenue
New York, NY 10022

Dear Dr. Eisenberg:

Please find enclosed a preliminary inquiry towards submitting a research grant application to the Council for Tobacco Research.

I look forward to hearing from you.

Sincerely,

**Fig. 1.1.** Relatively clean complex document example.

سيادة الأخ الرئيس القائد أبوعمار – حفظه الله

والأمر لسيادتكم

مكتب الرئيس

**Fig. 1.2.** Multi-language complex document example.

To search these and other complex, "real world" documents, we developed a CDIP prototype whose architecture is illustrated in Fig. 1.3 (Argamon et al. 2006). Initially, the complex documents are fed to the system as input, and in a pipelined manner, are processed by the layer extraction software, a point solution. This software extracts each component type, removing the

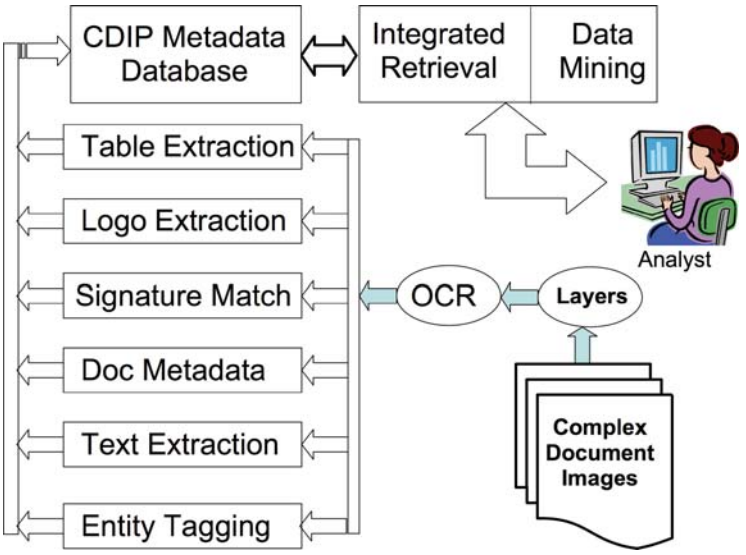



**Fig. 1.3.** The complex document information processing architecture.

surrounding artifacts. For example, tables, signatures, text, etc. are separated from each other. Individual components are sent to the optical character recognition module so as to recognize and correct the text portions of the artifact. Note that component types other than just the pure text will have text as part of them; hence they too need to be processed by the OCR software, a point solution. Our current implementation uses the ABBYY FineReader Engine; hence, many languages are already supported. However, neither Arabic nor Farsi script are among them. In the future, we expect to add an OCR module for these and other non-supported character sets. Having been processed by the OCR unit, the components are sent to the relevant additional point solutions for further processing. For example, a portion of text can be sent to the entity tagging point solution for noun identification. A signature can be sent to the signature matching point solution for identification from a select set of possible candidates.

Regardless which additional point solutions an artifact is routed to, each individual point solution, including the layering and OCR modules, generates metadata, that is, data about the data, describing the artifact content and stores these metadata in the metadata database. This process continues for all documents in the collection. Eventually the metadata are searched or mined for content in conjunction with a traditional text search of the textual components. Since metadata are generated for every component type within a document by their corresponding point solution and they are all stored in the same metadata associated with the same document (identifier), they can be queried and mined for content in a uniform and cohesive manner. Namely, one

can query about composite information that is not included in any individual component type but is present when multiple component types are viewed in the aggregate and possible correlations revealed.

Using our limited functionality prototype, we have already found documents that previously would not have been identified as relevant. For example, in one case we were looking for documents from a certain organization about a particular topic. Our prototype found not only those documents where the name of the organization and the topic were present in the text component but also those documents where their source (organization name) was not listed anywhere other than in their logo. Such latter documents would not have been found using conventional point solutions since a text processing point solution would not identify the source and a logo matching solution would not identify the topic. Our current activities involve incorporating a greater variety of point solutions, developing a better metadata fusion approach, and evaluating and improving the accuracy of retrieval of the prototype.

## Establishing a complex document benchmark

System evaluations are generally conducted via the use of a benchmark. In the search system realm, since 1992, the National Institute of Standards and Technology (NIST) has run an annual evaluation forum called TREC (see `trec.nist.gov`) where researchers are provided with benchmark data and queries to run on their respective systems. The outputs of the individual runs by the individual systems are submitted to the TREC organizers for evaluation. In mid November, the evaluation results are made publicly available. The problems with using TREC data and queries to evaluate a CDIP prototype is that all track data are "computer handy" and are text in nature; that is, they are not "real world." Thus, there did not exist a benchmark collection for CDIP evaluation, and obviously, one was needed.

Towards establishing a CDIP benchmark, several collection characteristics were selected to sustain the longevity of the applicability of the developed collection and query set. These included that the collection:

- Cover a richness of input in terms of a range of formats, lengths, and genres and variance in print and image quality.
- Include documents that contain handwritten text and notations, diverse fonts, multiple character sets, and graphical elements, namely graphs, tables, photos, logos, and diagrams.
- Contain a sufficiently high volume of documents.
- Contain documents in multiple languages including documents that have multiple languages within the same document.
- Contain a vast volume of redundant and irrelevant documents.
- Support diverse applications, thereby, include private communications within and between groups planning activities and deploying resources.
- Be publicly available at minimal cost and licensing.

After careful investigation of multiple potential sources, a document collection was selected, a wide range of simple and complex queries were developed, and corresponding relevant documents were determined. A complete description is provided in Lewis et al. (2006). Briefly, the collection chosen is a subset of the Master Settlement Agreement documents hosted by the University of California at San Francisco as the Legacy Tobacco Document Library (see `http://legacy.library.ucsf.edu`). These data were made public via legal proceedings against United States tobacco industries and research institutes. For the most part, the documents are distributed free of charge and are free of copyright restrictions. (The sued parties did not own a few of the Legacy Tobacco Document Library documents included; hence, some of them are potentially subject to copyright restrictions.) The collection consists of roughly 7 million documents or approximately 42 million scanned TIFF format pages (about 1.5 TB). These documents are predominantly in English; however, there are some documents in German, French, Japanese, and a few other languages included. A few of these documents also include multiple languages within a given document. As multiple companies at multiple sites using a diversity of scanners scanned the pages, the resulting image quality varies significantly. In access of 50 queries are available thus far, the scope of which and result set size varies significantly. For a complete discussion on the CDIP prototype and benchmark collection, see Agam et al. (2006).

## A light Arabic stemmer

English is obviously not the only language in which complex documents exist. As shown in Fig. 1.2 for example, at times, complex documents comprise of multiple languages on a single page. Furthermore, for counter-terrorism applications, most of the complex documents of interest are generally not in English. We now describe one potential future CDIP "point solution," a light Arabic stemmer developed at the IIT IR Lab.

The developed light stemmer (see Fig. 1.4, courtesy of Aljlayl and Frieder (2002)) consists of a sequence of rules derived based on the grammatical structure of the Arabic language. These rules locate and remove the most frequent prefixes and suffixes. As stated in Aljlayl and Frieder (2002):

> All Arabic words are based on tri-literal or quad-literal roots. Thus, choosing 3 letters as the minimum root preserves the integrity of the word-sense. Reducing the stem to less than 3 letters results in the loss of at least one of the original letters. Within each step, if an affix is matched to a word, then the condition that the stem be greater than or equal to 3 characters attached to that action are tested on what would be the resulting stem, if that affix was removed. Once an affix is matched in a word and the remaining characters satisfy the condition

Let $T$ denote the set of characters of the Arabic surface word
Let $L_i$ denote the position of letter $i$ in term $T$
Let Stem denote the term after stemming in each step
Let $D$ denote the set of definite articles
Let $S$ denote the set of suffixes
Let $P$ denote the set of prefixes
Let $n$ denote the total number of characters in the Arabic surface word

Step 1: Remove any diacritic in $T$

Step 2: Normalize أ, إ, آ in $L_1$ of $T$ to ا (plain alif)
Normalize ى in $L_n$ of $T$ to ي
Replace the sequence of ى in $L_{n-1}$ and ء in $L_n$ to ئ
Replace the sequence of ي in $L_{n-1}$ and ء in $L_n$ to ئ
Normalize ه in $L_n$ of $T$ to ة

Step 3: If the length of $T$ is greater than or equal to 3 characters then
Remove the prefix Waw "و" in position $L_1$

Step 4: For all variations of $D$ do,
Locate the definite article $D_i$ in $T$
If $D_i$ matches in $T$
$D_i = D_i +$ Characters in $T$ ahead of $D_i$
$Stem = T - D_i$
Normalize أ, إ, آ in $L_1$ of $S$ to ا (plain alif)

Step 5 If the length of *Stem* is greater than or equal to 3 characters then
For all variations of $S$, obtain the most frequent suffix,
Match the region of $S_i$ to *longest* suffix in *Stem*
If the length of $(Stem - S_i)$ greater than or equal to 3 characters then
$Stem = Stem - S_i$

Step 6 If the length of *Stem* is greater than 3 characters then,
For all variations of $P$ do
Match the region of $P_i$ in *Stem*
If the length of $(Stem - P_i)$ greater than 3 characters then
$Stem = Stem - P_i$

Step 7 Return the *Stem*

**Fig. 1.4.** Light Arabic stemming algorithm.

then that affix is removed and control moves to the next step; if the rule is not accepted, then the next affix is tested until either a rule from that step fires and control passes to the next step or there are no more affixes that satisfy the rules in that step, hence control moves to the next step.

Stemming is clearly not an end goal; rather, it is a means towards a goal. Our goal was the accurate search of Arabic documents. Thus, to evaluate the impact of our stemmer, we compared the search accuracy sustained against non-stemming (word based search), stemming using the state of the art stemmer, and our approach. The performance gains observed were statistically significant at the 99% level for our approach over the word based approach, and at least at the 95% level for our approach over the state of the art. Furthermore, we observed over a 70% improvement in accuracy derived by our approach over the conventional word based search approaches and a 20% improvement over the state of the art stemming based approaches. For additional details on the Arabic stemmer, see Aljlayl and Frieder (2002).

An optimized version (in terms of efficiency) of the above-described Arabic stemmer was licensed by IIT to Blue Shoe Technologies. For additional IIT IR Lab Arabic efforts, see Aljlayl et al. (2002) and Aqeel et al. (2006).

## IIT intranet mediator

In the past, the difficulty of the search process was finding the few relevant documents within the collection. Finding relevant documents is still difficult; however, with the ever increasing size of the available collections, another problem is introduced, namely, processing all the potentially relevant links to get the *answer* that one is looking for. In this context, we differentiate an *answer* from a *link* to a potential answer.

Users are familiar with search engines, and hence, they comfortably assume that a response to a user request should be a Google™-like style response, namely a set of links to potentially relevant resources. Such responses aid the users, however, they do not provide answers. Consider the request, "Who is Ophir Frieder?" A Google™-like style response would be links to resources describing Ophir Frieder. A better solution is an answer that includes pertinent information such as a title, department, e-mail address, and phone number. In addition to this answer, additional links could be returned for further information exploration should it be desired. Indeed, the IIT Intranet Mediator provides such an answer to a user a query; see Fig. 1.5. Simply stated, the returned results can be no worse than a Google™-like style response since links are still returned for further exploration. However, an answer is first and foremost provided, potentially reducing the burden of the user to find the desired information.

Evaluating the Intranet Mediator is difficult since benchmarks, in this case, are not applicable. Hence, as shown in Fig. 1.5, we provide an opportunity

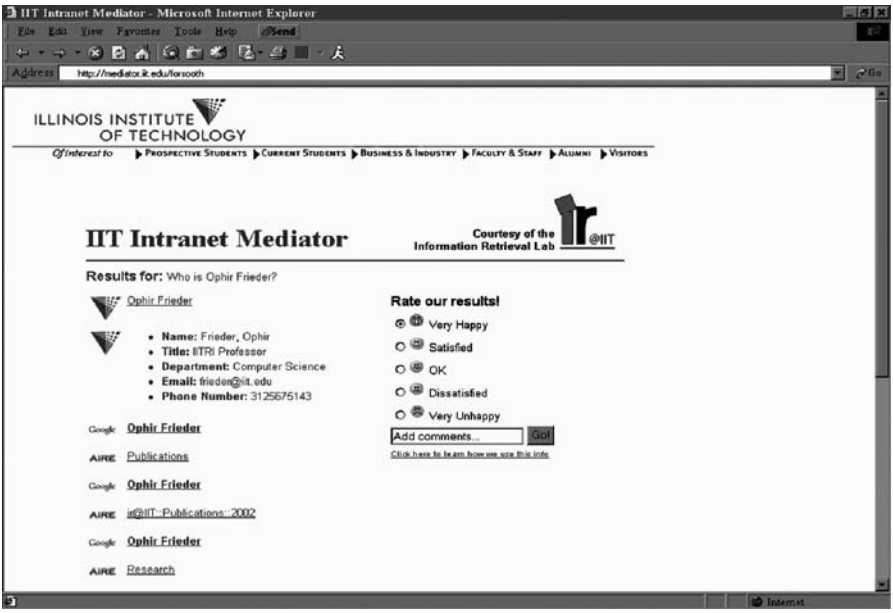



**Fig. 1.5.** Mediator answer to "Who is Ophir Frieder?"

for the user to evaluate their satisfaction as shown on the right hand side of the screen display. When the Intranet Mediator was made available to IIT undergraduate students, a total of 3662 queries were issued, and for 641 of the result sets returned, user satisfaction feedback was provided. From a 5-point satisfaction scale with 5 representing "very satisfied," 3 representing "OK," and 1 representing "unsatisfied," 60% of the users rated the system a 3 or better. We are now investigating means to improve user satisfaction. However, we are somewhat puzzled with the percentage of non-satisfied users as many of these same users are comfortable with Google™. Our surprise is based on that, in addition to an answer, we are still providing the same resource links that Google™ returns. In fact, we are actually including the top links returned by Google™ as part of the returned display.

To further demonstrate the potential of the Intranet Mediator, we likewise developed a prototype using some of the data sources available at the United States Holocaust Memorial Museum (USHMM). We harvested roughly six months, approximately 270,000 queries, of the USHMM query log and developed general rules for processing roughly 15% of the queries for demonstration purposes. Results from the query, "Which camp had the most deaths?" is shown in Fig. 1.6. As seen, an answer is produced, namely, the camp name (Auschwitz-Birkenau), the camp type (extermination and labor camp), location (Poland), date created (April 1940), date shut down (January 1945), number of deaths (1,500,000), number of prisoners (400,000), and camp liber-

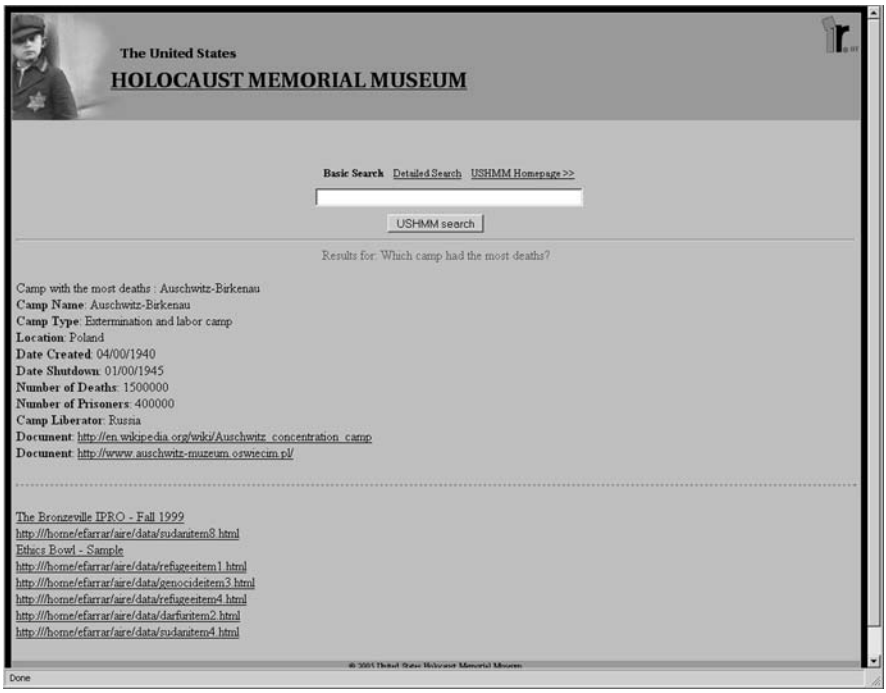



**Fig. 1.6.** Mediator answer to "Which camp had the most deaths?"

ator (Russia). Also listed are the sources of information along with additional links to relevant sources. Note that this answer is a composition of informational units obtained from multiple sources.

The conceptual architecture of the Intranet Mediator is illustrated in Fig. 1.7. It should be noted that the Mediator is not a search engine, but instead, it is a multiple search engine integration fabric. Initially, a user issues a query, namely a request, expressed as a question in English via a browser. Currently, we answer only factoid questions starting with "who," "what," "where," "which," and "when." A domain name server maps the request to a "lightly loaded" engine. Once the engine receives the query, a comparison is made against the recently asked queries (RAQ) cache to determine if the answer to the given query is already stored in the cache. Assuming that the query is not found in the cache, it is then tokenized and parsed using any efficient (simple) subject-verb-object parser and sent to the dispatcher.

The dispatcher consists of a set of rules. These rules define which source is relevant to the query and how to interface with the respective source. Sources are structured, e.g., a relational database, semi-structured, e.g., an XML database, or unstructured such as a collection of documents. Results from each of the respective search engines are sent to the results aggregator that

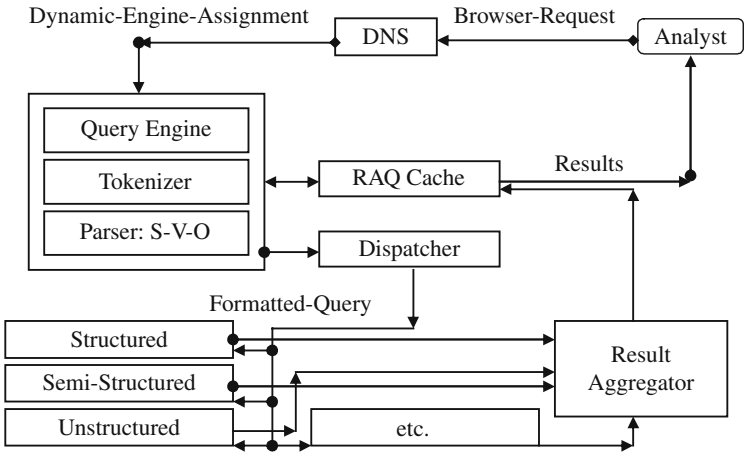



**Fig. 1.7.** The IIT Intranet mediator architecture.

combines them based on a weighting scheme. The weighting scheme takes into account the reliability factor of the site, the type of result, and the strength of conviction (scoring) of each of the returned results. The additional links returned are also recorded. The original query, combined answer, and associated links are then passed to the RAQ cache to register for potential future use and eventually an answer with additional links is sent back to the user. We are now in the process of developing and adapting more sophisticated source selection rules and answering schemes. For additional information related to the IIT Mediator, see Frieder and Grossman (2005), Grossman et al. (2002) and Heard et al. (2006).

## Conclusion

We started with the notion that "the search problem" is solved, and indeed, for conventional, computer generated and stored text documents, the field has significantly matured since its conception in the 1950s. The problem is that we are concerned with "real world documents," or more formally, complex documents, and those documents often are only available in hardcopy and contain handwritten annotations, signatures, tables, logos, and figures. For such documents, search technology is only in its infancy.

We described our approach to complex document information processing (search). By integrating multiple mature technologies, we developed a novel system that already has shown promise in finding relevant complex documents that previously could not be identified. We also developed a benchmark to evaluate future complex document search systems.

Having described our complex document information processing efforts to date, we then described two additional mature technologies developed at the IIT IR Lab, namely, an Arabic stemmer and an intranet mediator. The Arabic stemmer relied on a set of simple rules to remove prefixes and suffixes to yield 3 or 4 letter roots. In Arabic, most words roots are 3 or 4 letters. The accuracy of an Arabic search system was compared when using no stemming, prior state-of-the-art stemming techniques, and the described stemmer. The performance using the described stemmer was statistically significantly better than the other approaches evaluated.

Finally, we described the IIT Intranet Mediator. This mediator technology supports the unified, natural language querying of the integration of structured, semi-structured, and unstructured sources. We described its architecture and presented some sample results produced when using data from two different environments, a university system and the United States Holocaust Memorial Museum. We are presently extending the capability of the IIT Intranet Mediator to support a greater variety of natural language questions.

## Acknowledgments

Although I presented the talk, the work described was clearly a team effort. I appreciatively and wholeheartedly thank David Grossman, Gady Agam, Shlomo Argamon, David Lewis, David Doermann, Sargur Srihari, and the many people from Clarabridge for their respective contributions to the CDIP project; Mohammed Aljlayl for his work in Arabic stemming, David Grossman, Steven Beitzel, Eric Jensen, Jefferson Heard, and Jordan Wilberding for their Intranet Mediator efforts, and Abdur Chowdhury, Steven Beitzel, Eric Jensen and David Grossman for their involvement with the Advanced Information Retrieval Engine (AIRE). Finally, I also thank Nazli Goharian Gideon Frieder, and David Grossman for their constructive criticism of the efforts throughout and on their invaluable and repetitive comments and corrections on both this chapter and the keynote talk that predicated it.

## References

Agam, G., S. Argamon, O. Frieder, D. Grossman, and D. Lewis. 2006. Complex document information processing: Prototype, test collection, and evaluation. *Document Recognition and Retrieval XIII (Part of the IS&T/SPIE International Symposium on Electronic Imaging)*. San Jose, CA.

Agam, G., G. Bal, G. Frieder, and O. Frieder. 2007. Degraded document image enhancement. *Document Recognition and Retrieval XIV (Part of the IS&T/SPIE International Symposium on Electronic Imaging)*. San Jose, CA.

Aljlayl, M., and O. Frieder. 2002. On Arabic search: Improving the retrieval effectiveness via a light stemming approach. *ACM 11th Conference on Information and Knowledge Management (CIKM)*. Washington, DC.

Aljlayl, M., O. Frieder, and D. Grossman. 2002. On bi-directional English-Arabic search. *Journal of the American Society of Information Science and Technology* 53(13).

Aqeel, S., S. Beitzel, O. Frieder, D. Grossman, and E. Jensen. 2006. On the development of name search techniques for Arabic. *Journal of the American Society of Information Science and Technology* 57(6).

Argamon, S., G. Agam, O. Frieder, D. Grossman, D. Lewis, G. Sohn, and K. Voorhees. 2006. A complex document information processing prototype. *ACM 29th Conference on Research and Development in Information Retrieval (SIGIR).* Seattle, WA.

Chen, K., D. Doermann, T. Drayer, L. Hernandez, S. Jaeger, S. Sampat, and G. Zhu. 2005. DocLib: A document processing research tool. *Symposium Document Image Understanding Technology.* Adelphi, MA.

Frieder, O., and D. Grossman. 2005. *Intranet Mediator.* US Patent #6,904,428.

Grossman, D., S. Beitzel, O. Frieder, and E. Jensen. 2002. IIT Intranet Mediator: Bringing data together on a corporate intranet. *IEEE IT Professional* 4(1).

Grossman, D., and O. Frieder. 2004. *Information Retrieval: Algorithms and Heuristics*, 2*nd* ed. Springer Science and Business Media, New York.

Heard, J., G. Frieder, O. Frieder, D. Grossman, L. Kane, and J. Wilberding. 2006. On a mediated search of the United States Holocaust Memorial Museum data. *6th Next Generation Information Technology Systems (NGITS).* Sefayim, Israel.

Infantes-Morris, T., P. Bernhard, G. Faulkner, K. Fox, and K. Stripling. 2003. Industrial evaluation of a highly-accurate academic IR system. *ACM 12th Conference on Information and Knowledge Management (CIKM).* New Orleans, LA.

Lewis, D., G. Agam, S. Argamon, O. Frieder, D. Grossman, and J. Heard. 2006. Building a test collection for complex document information processing. *ACM 29th Conference on Research and Development in Information Retrieval (SIGIR).* Seattle, WA.

Srihari, S., G. Agam, S. Chen, O. Frieder, C. Huang, S. Shetty, and H. Srinivasan. 2006. Document image retrieval using signatures as queries. *IEEE 2nd International Conference on Document Image Analysis for Libraries (DIAL).* Lyon, France.

Srihari, S., M. Kalera, and A. Xu. 2004. Offline signature verification and identification using distance statistics. *International Journal of Pattern Recognition and Artificial Intelligence* 18(7).

# 2

# Signature-Based Retrieval of Scanned Documents Using Conditional Random Fields

Harish Srinivasan and Sargur Srihari

**Summary.** In searching a large repository of scanned documents, a task of interest is that of retrieving documents from a database using a signature image as a query. This chapter presents a signature retrieval strategy using document indexing and retrieval. Indexing is done using (i) a model based on Conditional Random Fields (CRF) to label extracted segments of scanned documents as Machine-Print, Signature and Noise, (ii) a technique using support vector machine to remove noise and printed text overlapping the signature images and (iii) a global shape-based feature extractor that is computed for each signature image. The documents are first segmented into patches using a region growing algorithm and the CRF based model is used to infer the labels of each of these patches. The robustness of the method is due to the inherent nature of modeling neighboring spatial dependencies in the labels as well as the observed data using CRF. The model parameters are learnt using conjugate gradient descent with line search optimization to maximize pseudo-likelihood estimates and the inference of labels is done by computing the probability of the labels under the model with Gibbs sampling. A further post processing of the labeled patches yields signature regions which are used to index the documents. Retrieval is performed using a matching algorithm to compare the query with the indexed documents. Signature matching is based on a normalized correlation similarity measure using global shape-based binary feature vectors. The end-to-end system is a content-based image retrieval system designed for signatures.

## Introduction

Retrieving relevant documents from a repository of scanned documents has many applications including the legal and forensic domains. In particular documents containing handwriting have a potentially useful role in counterterrorism operations, e.g., retrieving forms filled out by certain applicants for opening post-office boxes, identifying envelopes of interest in the mail stream, etc. In searching complex documents, a task of relevance is relating the signature in a given document to the closest matches within a database of documents; this is the signature retrieval task which is addressed in this chapter.

S. Argamon, N. Howard (eds.), *Computational Methods for Counterterrorism*, DOI 10.1007/978-3-642-01141-2_2,

Retrieval of handwritten words has been found to be more challenging than image matching due to the lack of low level distinguishing features like color and texture. Handwritten word retrieval has been discussed in Rath et al. (2004), Zhang et al. (2004), Kolz et al. (2000), Plamodon and Lorette (2000). The method of Kolz et al. (2000) extracts profile-based holistic shape features from a line or word image and uses dynamic time warping (DTW) to match words. A word shape based method was shown to perform better than the DTW method, in terms of efficiency and effectiveness (Zhang et al. 2004). Considering historical manuscripts, Rath et al. (2004) describe a method for retrieval based on text queries without recognition using a transcribed set of pages for training.

This chapter presents an effective signature extraction and retrieval technique. It is based on a statistical model for machine learning known as Conditional Random Fields (CRFs) (Lafferty et al. 2001; Kumar and Hebert 2003; Quattoni et al. 2005). CRFs are more general than Hidden Markov Models in that there are no implicit independence assumptions. The CRF model is used in extracting signatures from complex documents by isolating the different contents present in the documents. The motivation to use a CRF based model for this application arises from the spatial inter-dependencies of the different regions in documents. The problem is formulated as follows: Given a document: (i) Segment the document into a number of patches (approximately the size of a word), and (ii) Label each of the segments as one of Machine-Print, Handwriting or Noise. Then the region containing the signatures are identified from the labeled patches and isolated.

Given a database of signed documents, the retrieval task (Srihari et al. 2006) is to relate a query document to other documents in this database which have been signed by the same author. The documents under consideration are indexed by the features of the signatures extracted from the documents. The retrieval task would be to retrieve all the other documents signed by the same author. This involves extracting the features of the query signature and matching these features to those of the indexed documents. A technique based on query expansion using automatic relevance feedback (Salton and McGill 1983) has also been implemented, where the highest ranked result is used along with the original query to retrieve relevant documents. This retrieval technique can be extended to accept a text query of the authors name provided each author has been previously enrolled with at least 1 signature.

## Indexing

The steps involved in indexing the document images are described here.

### Signature block location

The first step in indexing a scanned document image is to extract the signature block. A signature block is defined as a rectangular image snippet

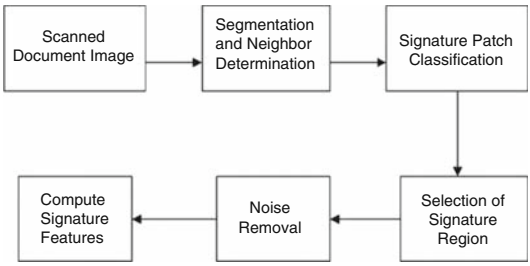



**Fig. 2.1.** Block diagram of indexing the documents.

containing the entire signature. The signature block is further processed to remove non-signature material, e.g., printed name of the signatory, portions of the accompanying text, spots, etc. The operational steps in signature extraction (Fig. 2.1) are: (i) segmentation into patches and neighbor determination, (ii) classification of patches into signature and non-signature classes, (iii) isolating the signature region (image snippet) from the rest of the image, (iv) removal of noise and printed text from the signature region and (v) extraction of features required for signature matching.

### Segmentation and neighbor determination

A patch is defined to be a region in a document such that, if a rectangular window (size determined dynamically for each document) is drawn with each foreground pixel within the patch at its center, then the window shall not contain any foreground pixel from another patch. The size of the patch was optimized in a way to represent approximately the size of a word. The algorithm for generating these patches is a region growing algorithm and a brief description is given below.

1. Initialize every pixel to be a separate patch.
2. Start with a foreground pixel that is not already marked.
3. With this pixel as the center, draw a rectangular window of size proportional to the height and width of the document being considered.
4. All foreground pixels of connected components with any pixel enclosed within this rectangular window are marked as belonging to the same patch as that of the center pixel.
5. Repeat steps 2 through 4 until all pixels are marked.
6. Patches with pixels lesser than a fixed threshold are ignored as noise and are not attempted to be labeled as one of machine-print, handwriting/signature, noise.

Once all the patches are obtained for a document, the neighboring patches are identified. A total of 6 neighbors are identified for each patch. These neighbors are the closest (top/bottom) and the two closest (left/right) in terms of the convex-hull distance between the patches considered. The reason for including more neighbors from the right and left, is the fact that scanned documents have greater dependency across the width of the document. The definitions of top, bottom, left and right are determined from the center of gravity of the patch being considered. However the convex-hull distance between two patches is measured taking the entirety of both the patches.

**Conditional Random Field model description**

A model based on Conditional Random Fields is used to label each of the patches identified using the labels of the neighboring patches. The probabilistic model of the Conditional Random Field used is given below.

$$P(\mathbf{y}|\mathbf{x},\theta) = \frac{e^{\psi(y,\mathbf{x};\theta)}}{\sum_{y'} e^{\psi(y',\mathbf{x};\theta)}} \tag{2.1}$$

where $\mathbf{y} \in \{\text{Machine-print, Handwriting, Noise}\}$ and $\mathbf{x}$ : Observed document and $\theta$ : CRF model parameters. It is assumed that a document is segmented into $m$ non-overlapping patches. Then

$$\psi(y,x;\theta) = \sum_{j=1}^{m} \left( A(j,y_j,\mathbf{x};\theta^s) + \sum_{(j,k)\in E} I(j,k,y_j,y_k,\mathbf{x};\theta^t) \right) \tag{2.2}$$

The first term in Eq. 2.2 is called the state term and it associates the characteristics of that patch with its corresponding label. $\theta^s$ are called the state parameters for the CRF model. Analogous to it, the second term, captures the neighbor/contextual dependencies by associating pair wise interaction of the neighboring labels and the observed data. $\theta^t$ are called the transition parameters of the CRF model. $E$ is a set of edges that represent the neighbors of a patch.

The association potential can be modeled as

$$A(j,y_j,\mathbf{x};\theta^s) = \sum_i (h_i \cdot \theta_{ij}^{s_2})$$

where $h_i$ is typically the state feature value associated with the patch being considered. In order to introduce a non-linear decision boundary we define $h_i$ to be a transformed state feature vector

$$h_i = tanh\left(\sum_l (f_l^{s_1}(j, y_j, \mathbf{x}) \cdot \theta_l^{s_1} i)\right)$$

where $f_l^s$ is the $l$th state features extracted for that patch. The state features that are used for this problem are defined later in Table 2.1. The state features, $f_l$ are transformed by the tanh function to give the feature vector $\mathbf{h}$. The state parameters $\theta^s$ are a union of the two sets of parameters $\theta^{s_1}$ and $\theta^{s_2}$.

The interaction potential $I(\cdot)$ is generally an inner product between the transition parameters $\theta^t$ and the transition features $f_t$. To introduce non-linearity, we use the idea of kernels, and the interaction potential is defined as follows:

$$I(j, k, y_j, y_k, \mathbf{x}; \theta^t) = \sum_l (\phi_l \cdot \theta_l^t)$$

where $\phi_l$ is the $l$th transition feature after applying a quadratic kernel on the original transition features as defined below.

$$\Phi_l = \left\langle f^t(j, k, y_j, y_k, \mathbf{x}) \cdot f^t(j, k, y_j, y_k, \mathbf{x}) \right\rangle$$

**Table 2.1.** Description of the 23 state features used.

| **State Feature** | **Description** |
|---|---|
| Height | Maximum height of the patch |
| Avg component width | The mean width of the connected components within a patch |
| Density | Density of foreground pixels within the patch |
| Aspect ratio | Width/Height of the patch |
| Gabor filter | 8 features capturing the different stroke orientations |
| Variation of height | Variation in height within a patch |
| Width variation | Variation in width within a patch |
| Overlap | Sum of overlap in area between the connected components within a patch |
| Percentage of text above | Relative location of the patch with respect to the entire document |
| Number of components | Count of the connected components within a patch |
| Maximum component size | Maximum size of a component within a patch |
| Points in convex hull | Number of points in the convex hull of the patch |
| Maximum run length | The maximum horizontal run length within a patch |
| Avg run length | The average horizontal run length within a patch |
| Horizontal Transitions | A count of the number of times the pixel value transitions from white to black horizontally |
| Vertical Transitions | A count of the number of times the pixel value transitions from white to black vertically |

**Parameter estimation**

There are numerous ways to estimate the parameters of this CRF model (Wallach 2002). In order to avoid the computation of the partition function we learn the parameters by maximizing the pseudo-likelihood of the documents, which is an approximation of the maximum likelihood value. We estimate the Maximum pseudo-likelihood parameters using conjugate gradient descent with line search optimization. The pseudo-likelihood estimate of the parameters $\theta$ are given by Eq. 2.3:

$$\hat{\theta_{ML}} \approx \arg\max_{\theta} \prod_{i=1}^{M} P(y_i|y_{\mathcal{N}_i}, \mathbf{x}, \theta) \tag{2.3}$$

where $P(y_i|y_{\mathcal{N}_i}, \mathbf{x}, \theta)$ (Probability of the label $y_i$ for a particular patch $i$ given the labels of its neighbors, $y_{\mathcal{N}_i}$), is given below.

$$P(y_i|y_{\mathcal{N}_i}, \mathbf{x}, \theta) = \frac{e^{\psi(y_i, \mathbf{x}; \theta)}}{\sum_a e^{\psi(y_i=a, \mathbf{x}; \theta)}} \tag{2.4}$$

where $\psi(y_i, x; \theta)$ is defined as before in Eq. 2.2.
Note that the Eq. 2.3 has an additional $y_{\mathcal{N}_i}$ in the conditioning set and hence the factorization into products is feasible as the set of neighbors for the patch form the minimal Markov blanket.

From Eqs. 2.3 and 2.4, the log pseudo-likelihood of the data is given by

$$\mathcal{L}(\theta) = \sum_{i=1}^{M} \left( \psi(y_i = a, x; \theta) - log \sum_a e^{\psi(y_i=a, x; \theta)} \right)$$

**Features for signature classification**

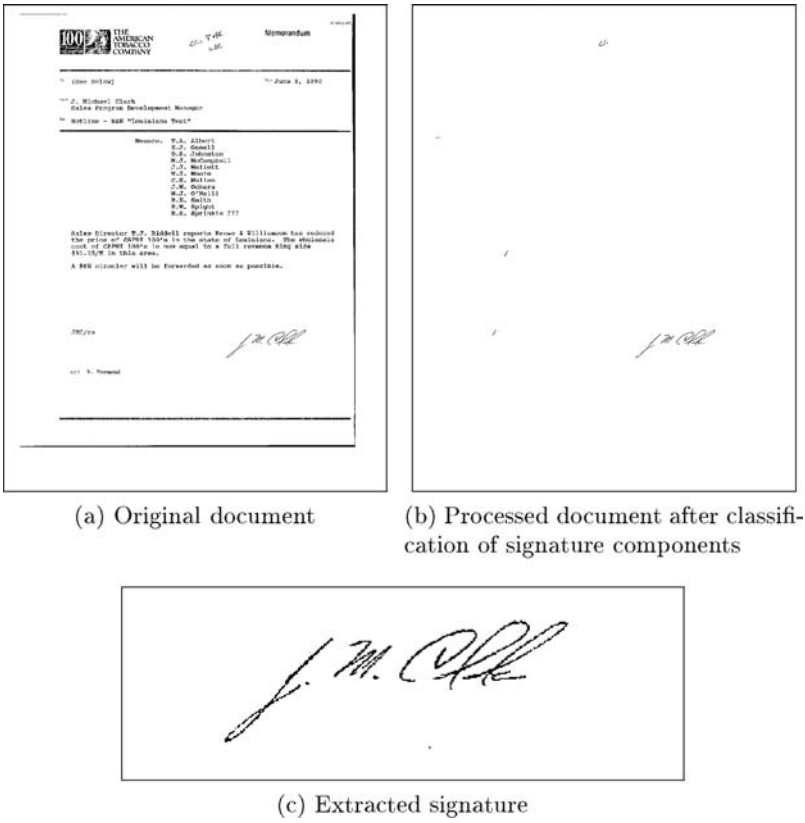
State features try to associate each patch to a label using characteristics of that patch alone. Analogous to these, transition features associate a patch to a label using information from the neighboring patches. Twenty-three state features are extracted for each patch, as described in Table 2.1. Then, the four transition features described in Table 2.2 are computed using the state features and neighbor information. Using these extracted features from each of the 3500 patches in the training set, the parameters of the CRF were estimated as described above. Figure 2.2a shows an example of a document used for feature extraction.

**Classification**

The goal of inference is to assign a label to each of the patches being considered. The algorithm for inference uses the idea of Gibb's sampling (Casella and George 1992).

**Table 2.2.** Description of the 4 transition features used. Transition features are computed for a patch and its neighbor.

| Transition Feature | Description |
|---|---|
| Relative location | Assigned weights based on the relative location - top/bottom or right/left |
| Convex hull distance | The convex hull distance between the 2 patches |
| Ratio of aspect ratio | The ratio of the aspect ratio values of the 2 patches |
| Ratio of number of components | The ratio of the number of components present in the 2 patches |



(a) Original document

(b) Processed document after classification of signature components

(c) Extracted signature

**Fig. 2.2.** Sample signature extraction results (a) Step 1: Feature extraction; (b) Step 2: Classification; (c) Step 3: Post-processing.

1. Randomly assign labels to each of the patches in a document based on an intuitive prior distribution of the labels.
2. Choose a patch at random and compute the probability of assigning each of the labels using the model from Eq. 2.4 to obtain a probability distribution $p$ for the labels.

3. Use Gibbs sampling to sample from this distribution $p$ to assign a probable label to the patch.
4. Repeat steps 2 and 3 until the assignments do not change. Store the set of label assignments along with the probability distribution $p$.
5. Repeat steps 1–4, for a sufficient number of iterations in order to eliminate the dependency on the initial random label assignments.
6. Consider the set of arrived assignments at step 4 in each of the iterations, and for all the patches pick the labels with the maximum probability as the final set of labels.

Figure 2.2b shows an example of a document image obtained as a result of the classification of the signature labels on the document in Fig. 2.2a.

**Post-processing**

In this step, only the patches labeled as possible signatures are considered. Each of these patches is merged with other neighboring possible signature patches, the components on the right and left side being weighed more than those on the top and bottom. A region growing algorithm like the one described above but with a larger window size is used to merge the patches. Other small components which were left out initially are inserted back into the signature blocks being considered. Figure 2.2c shows the result of the post-processing step on the image in Fig. 2.2b.

**Noise removal**

Noise removal is carried out to get rid of any noise or printed text overlapping the extracted signature region. We use Support Vector Machines (SVM) (Burges 1998) to classify each connected component as either a part of a signature or a noise component, comprising of printed text, small handwritten text, logos, noise, etc. The SVM is previously trained on the connected components extracted from 10 sample signatures with noise. At the end of the classification step we obtain the signature image with only the signature components remaining. The features used include directional features, height, perimeter and aspect ratio. An example of the results obtained by this noise removal procedure is shown in Fig. 2.3.

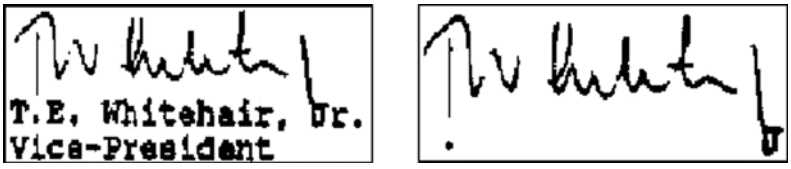



**Fig. 2.3.** Example of noise removal.

**Signature feature extraction**

The next step involves indexing each document by converting the signature image extracted from the document into a set of binary feature vectors. The features used here are the Gradient, Structural and Concavity (GSC) features which measure the image characteristics at local, intermediate and large scales and hence approximate a heterogeneous multi resolution paradigm to feature extraction. The features for the signature images which are extracted under a 4 × 8 division, contain 384 bits of gradient features, 384 bits of structural features and 256 bits of concavity features, giving us a binary feature vector of length 1024 (Zhang and Srihari 2003b). Each of these sets of binary features uniquely represents a given sample signature. Figure 2.4 shows an example of a signature image under this 4 × 8 division and the corresponding binary feature vector obtained.

The gradient features capture the stroke flow orientation and its variations using the frequency of gradient directions, as obtained by convolving the image with a Sobel edge operator, in each of 12 directions and then thresholding the resultant values to yield a 384-bit vector. The structural features represent the coarser shape of the word and capture the presence of corners, diagonal lines, and vertical and horizontal lines in the gradient image, as determined by

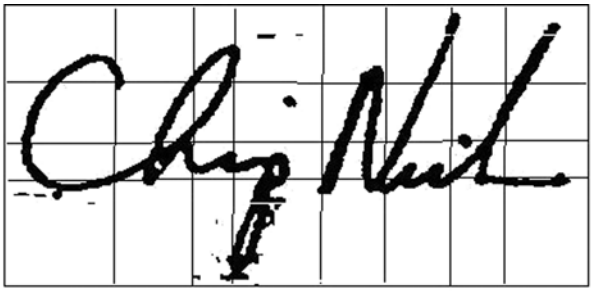



**Fig. 2.4.** Feature extraction (a) Signature image under a 4 × 8 division; (b) 1024 bit binary feature vector extracted.

12 rules (Favata and Srikantan 1996). The concavity features capture the major topological and geometrical features including direction of bays, presence of holes, and large vertical and horizontal strokes.

## Retrieval

The document retrieval is performed using a matching algorithm to compare the query with the signature. Figure 2.5 shows the various operational steps in the retrieval process: (i) noise removal from the query signature; (ii) feature extraction from the query signature after noise removal; (iii) matching the query signature features to each of the indexed documents; and (iv) ranking the documents in accordance with the results from the matching algorithm.

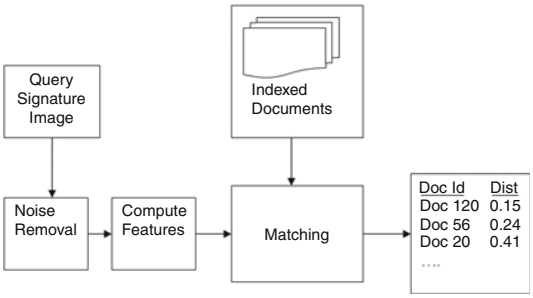



**Fig. 2.5.** Block diagram of document retrieval.

### Matching algorithm

Given a query signature image, the relevant documents are retrieved using a matching algorithm. The GSC binary feature vectors are extracted for the query, and the matching algorithm's task is to compare these features with the indexed features of the signatures present in the database of documents. Figure 2.6 shows a query signature image being matched against a few extracted signatures and the resulting dissimilarity measures obtained using the matching algorithm.

The distance between the queried signature and each of the indexed documents in the database is calculated using a normalized correlation similarity measure (Zhang and Srihari 2003a, b). Given the two binary feature vectors $X \in \Omega$ and $Y \in \Omega$, each similarity score S(X, Y) uses all or some of the four possible values, i.e. $S_{00}$; $S_{01}$; $S_{10}$; $S_{11}$. Here $S_{ij}$, (i,j) $\in \{0,1\}$, is the number of occurrences where pattern i occurs in the first binary vector and pattern j occurs in the second vector in the same position. The similarity distance S(X, Y ) between two feature vectors $X$ and $Y$ is given by Eq. 2.5.


**Fig. 2.6.** Subset of retrieval results with the query image on the *left* and the signatures matched against and their corresponding dissimilarity distances on the *right*.

$$S(X,Y) = \frac{1}{2} + \frac{S_{11}S_{00} - S_{10}S_{01}}{2((S_{10}+S_{11})(S_{01}+S_{00})(S_{11}+S_{01})(S_{00}+S_{10}))^{1/2}} \tag{2.5}$$

where

$S_{00}$ = the first binary vector has a 0 and the second vector too has a 0 in the corresponding positions.
$S_{11}$ = the first binary vector has a 1 and the second vector too has a 1 in the corresponding positions.
$S_{01}$ = the first binary vector has a 0 while the second vector has a 1 in the corresponding positions.
$S_{10}$ = the first binary vector has a 1 while the second vector has a 0 in the corresponding positions.

When constructing the similarity distance measure all possible matches $S_{ij} \in 0{,}1$ are considered for better classification. Also $S_{00}$ has been weighted with a beta value of 0.5 to boost classification. The results are ranked in the increasing order of this dissimilarity distance which varies between 0 and 1, a value of 0 indicating an exact match. In the signature retrieval process there is no prior knowledge of the writers signature, the goal is to identify the closest

matching signatures and to identify all the documents containing signatures by the writer of the queried signature. Each of the retrieved signature images is also linked with its corresponding document ID, which allows the user to easily retrieve its location and the document it belongs to.

Before the matching algorithm is applied, the query signature image is processed to remove any overlapping printed or noisy components as mentioned above. Following this, the GSC features for this component are extracted.

### Query expansion using automatic relevance feedback

A query expansion is done using the feedback (retrieval results) of the matching algorithm. The matching score $S_i$ for a query $q$, matched against a document $D_i$, given by Eq. 2.6, is computed for each document and sorted in ascending order. The document with the lowest $S_i$ being the most relevant document retrieved.

$$S_i = S(f(q), f(D_i)) \tag{2.6}$$

where $f(q)$ is the binary feature vector of the image q, $f(D_i)$ is binary feature vector indexed in $D_i$, and $S(f(q), f(D_i))$ is given by Eq. 2.5.

Let document $D_i$ correspond to the document with the lowest $S_i$. The signature image extracted from the document $D_i$ is used as a new query $q_{new}$, and added to the existing query to formulate an expanded query consisting of the 2 images, q and $q_{new}$.

The retrieval is performed using the matching algorithm with this new query $\{q, q_{new}\}$. The new score for each document, $S_i(\{q, q_{new}\}, D_i)$, is computed by the minimum distance obtained from the 2 queries as given by Eq. 2.7.

$$S_i(\{q, q_{new}\}, D_i) = min\{S(q, D_i)\ ,\ S(q_{new}, D_i)\} \tag{2.7}$$

This technique improves the accuracy of the retrieved results as the matching algorithm consistently returns relevant documents in the top results.

## Dataset

The dataset used for this experiment was taken from a set of 744 document images signed by 67 different authors. This set of documents consists of a variety of documents, a majority of which have printed text with a signature at the bottom. There are also documents with handwritten text around this printed text, only handwritten documents, documents with images like tables, graphs, etc and multiple signatures per document or no signatures at all. Many of these documents also have logos, other symbolic text and noisy components like words circled or scratched or handwritten text overlapping the printed text or printed text overlapping the signatures. There are also documents with lines and black borders and noise. Some of the writers have several

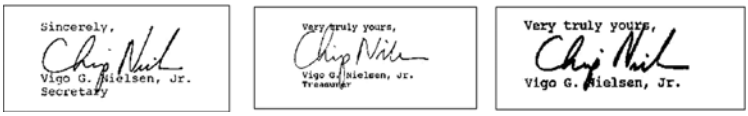



**Fig. 2.7.** All the automatically extracted samples for writer "10".

types of signatures like the writer's full name, initials, only first name, etc. Documents with multiple signatures per document and purely handwritten documents with signatures have also been considered here. For this experiment we randomly picked several different authors and picked 2–5 documents per author making a total of 101 documents containing a total of 114 signatures. Figure 2.7 shows all the signature samples automatically extracted from the documents belonging to one of the writers.

## Experiments and results

In this section, the test setup and the experimental results obtained for the signature retrieval task are described. In the test setup for Signature Retrieval, the images were divided into 2 groups per writer. One group consisting of known document images and the second group consisting of the questioned signatures for testing. The image formats supported are png, jpeg and tiff. The database of documents with known signatures are first processed to index each document. Out of the 101 documents from which the signatures were extracted, in 91.2% (= 104) of the cases the extracted region contained the entire signature image correctly extracted. Following this, the signature image in question is selected and this queried image is preprocessed to remove any overlapping printed text or noise. The set of indexed documents are selected and the signature retrieval process is carried out against this set of known documents. In each case, the precision and recall measures are calculated. The precision and recall measures (Salton and McGill 1983; van Rijsbergen 1979) for a rank "R" where the author of the questioned signature is represented by "A" are defined as follows

$$Recall\ of\ label\ 'a' = \frac{Amount\ of\ correctly\ classified\ data\ of\ label\ 'a'}{Total\ amount\ of\ data\ of\ label\ 'a'}$$

$$Precision\ of\ label\ 'a' = \frac{Amount\ of\ correctly\ classified\ text\ of\ label\ 'a'}{Total\ amount\ of\ text\ classified\ to\ be\ of\ label\ 'a'}$$

The testing was done for 1–2 extracted signature images per writer which were randomly selected from the entire set. Each of these signatures was queried against the entire set of 114 indexed signature images in the database. The ranks of the retrieved documents which were signed by the author of the questioned signature were noted in each case and the average precision and recall values were estimated for different ranks.

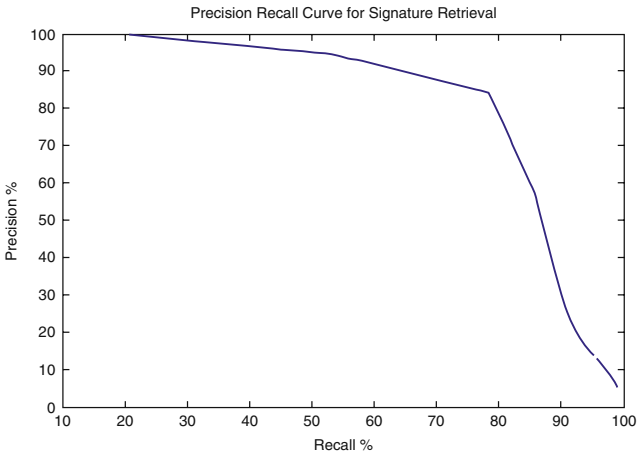



**Fig. 2.8.** Precision-recall curves for signature retrieval results: Precision of 84.2% at recall of 78.4% after query expansion.

The experiments were conducted using query expansion, where the top results from the retrieval results for the initial query were used along with the initial query to retrieve relevant documents. Figure 2.8 shows the precision recall curves obtained in this experiment. In the top 5 results a recall of 78.4% is obtained, the precision at this point is 84.2%. Table 2.3 shows the results at the end of this phase. There is an increase in the retrieval accuracy on using query expansion, this shows that the system consistently retrieved a relevant document as the top choice. And the usage of this top choice result along with the original query strengthened the retrieval accuracy.

The retrieval accuracy also has been impacted by several factors like: the signature extraction was effective in 91.2% of the cases, so some of the indexed documents contained spurious signature images; the noise removal technique has led to the removal of some components belonging to the signature in a small number of cases; and the poor quality of some of the documents.

**Table 2.3.** Recall measures for signature retrieval from entire database.

| No of Results Considered | Recall Measure(%) |
|---|---|
| Rank 1 | 18.6 |
| < Rank 3 | 52.9 |
| < Rank 5 | 78.4 |
| < Rank 10 | 87.3 |
| < Rank 15 | 89.7 |
| < Rank 25 | 92.6 |
| < Rank 50 | 97.0 |

## Conclusions

Here the set of experiments done for the problem of document retrieval using signatures and its results were presented. The tests were conducted on a variety of document and signature samples including those with noise, logos, figures, printed and handwritten text. Although the presence of noise and text overlapping the signatures make retrieval a challenging task, our technique returned a relatively high precision and recall accuracy of 84.2% and 78.4% respectively when considering the top 5 results. This can be attributed to the usage of conditional random fields for the removal of printed and noisy data from the documents leading to an accurate signature extraction in most cases, followed by the usage of an effective matching algorithm using global shape-based features.

## References

Burges, C. 1998. A tutorial on support vector machines for pattern recognition. *Data Mining and Knowledge Discovery*, 2(2):121–167.

Casella, G. and E. George. 1992. Explaining the Gibbs sampler. The *American Statistician*, 46:167–174.

Favata, J. T. and G. Srikantan. 1996. A multiple feature resolution approach for handprinted digit and character recognition. *International Journal of Imaging Systems and Technology*, 7:304–311.

Kolz, A., J. Alspector, M. Augusteijn, R. Carlson, and G. V. Popescu. 2000. A line-oriented approach to word spotting in handwritten documents. *Pattern Analysis and Applications*, 2(3):153–168.

Kumar, S. and M. Hebert. 2003. Discriminative fields for modeling spatial dependencies in natural images. *Advances in Neural Information Processing Systems (NIPS-2003)*.

Lafferty, J., A. Macullum, and F. Pereira. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequential data. *Eighteenth International Conference on Machine Learning (ICML-2001)*.

Plamondon, R. and G. Lorette. 2000. On-line and offline handwriting recognition: A comprehensive survey. *IEEE Transactions on Pattern Recognition and Machine Intelligence*, 22(1):63–84.

Quattoni, A., M. Collins, and T. Darrel. 2005. Conditional random fields for object recognition. *Advances in Neural Information Processing Systems 17 (NIPS 2004)*.

Rath, T., R. Manmatha, and V. Lavrenko. 2004. A search engine for historical manuscript images. *Proceedings of the 27th Annual Int'l SIGIR Conference*.

Salton, G. and M. J. McGill. 1983. *Introduction to Modern Information Retrieval*. McGraw-Hill, New York.

Srihari, S., S. Shetty, S. Chen, H. Srinivasan, and C. Huang. 2006. Document image retrieval using signatures as queries. *Document Image Analysis for Libraries (DIAL'06)*.

van Rijsbergen, C. J. 1979. *Information Retrieval*. Butterworths, London.

Wallach, H. 2002. Efficient training of conditional random fields. *Proceedings of 6th Annual CLUK Research Colloquium*.

Zhang, B. and S. Srihari. 2003a. Binary vector dissimilarity measures for handwriting identification. *SPIE, Document Recognition and Retrieval X*, pp. 155–166.

Zhang, B. and S. Srihari. 2003b. Properties of binary vector dissimilarity measures. *Cary, North Carolina*, September.

Zhang, B., S. N. Srihari, and C. Huang. 2004. Word image retrieval using binary features. *Document Recognition and Retrieval XI, SPIE, San Jose, CA*.

# 3

# What Makes a Good Summary?

Qunhua Zhao, Eugene Santos, Jr., Hien Nguyen, and Ahmed Mohamed

**Summary.** One of the biggest challenges for intelligence analysts who participate in prevention or response to a terrorism act is to quickly find relevant information from massive amounts of data. Along with research on information retrieval and filtering, text summarization is an effective technique to help intelligence analysts shorten their time to find critical information and make timely decisions. Multi-document summarization is particularly useful as it serves to quickly describe a collection of information. The obvious shortcoming lies in what it cannot capture especially in more diverse collections. Thus, the question lies in the adequacy and/or usefulness of such summarizations to the target analyst. In this chapter, we report our experimental study on the sensitivity of users to the quality and content of multi-document summarization. We used the DUC 2002 collection for multi-document summarization as our testbed. Two groups of document sets were considered: (I) the sets consisting of closely correlated documents with highly overlapped content; and (II) the sets consisting of diverse documents covering a wide scope of topics. Intuitively, this suggests that creating a quality summary would be more difficult for the latter case. However, human evaluators were discovered to be fairly insensitive to this difference. This occurred when they were asked to rank the performance of various automated summarizers. In this chapter, we examine and analyze our experiments in order to better understand this phenomenon and how we might address it to improve summarization quality. In particular, we present a new metric based on document graphs that can distinguish between the two types of document sets.

## Introduction

To prevent or quickly respond to a terrorism act, every intelligence analyst needs to gather critical information and makes decisions based on retrieved information under time pressure. He or she must deal with a huge volume of online and offline information resources on a daily basis. It has been estimated that an analyst needs to process 1500–2000 messages classified at various levels every day (according to CIA analyst Larry Johnson; Goldstein 2006). Along with research on search engines, automatic text summarization has been

S. Argamon, N. Howard (eds.), *Computational Methods for Counterterrorism*, DOI 10.1007/978-3-642-01141-2_3,

proposed as one natural approach to coping with this problem (Elhadad 2004; Mani and Maybury 1999). Usually, summaries can be classified as *informative*, which provide readers with the key information contained in the original texts, and *indicative*, which are used to help analysts judge relevancy of the texts to the tasks at hand so that they can decide quickly whether it is worth going through the full texts. Both types of summaries are critical during information gathering for intelligence analysis.

Text summarization has been defined as "the process of distilling the most important information from a source (or sources) to produce an abridged version for a particular user (or users) and task (or tasks)" (Mani and Maybury 1999). This definition actually includes three elements: (I) a summary is a condensed version that contains important information; (II) a summary can be generated for a single document or a set of documents (the latter is called *multi-document summarization*); and (III) a summary should be user- and task-oriented. Taking into account the possible uses that a summary may serve, it can be indicative by pointing out the topics which have been addressed in the text, or informative by covering as much as possible, important content and/or critical information that offers a critique of the source. Hence, the intention and coverage of summaries can be different depending on tasks. Even for the same tasks, various people may have different opinions on what pieces of information are interesting and important. However, the prevalent form of summaries that we usually encounter is the generic summary (for both informative and indicative ones) that targets a wide range of readers. Recently, user-sensitive (or user-centered) summarization has become an increasingly active research area which focuses on how individual differences affect a user's judgment on what should be included in a summary (Elhadad 2004).

Intuitively, multi-document summarization seems to be a more difficult task than creating a summary from a single text, given that a set of documents typically cover a variety of topics. For example, assuming that there is a document set containing news reports on the Mumbai commuter rail explosion in India on July 11, 2006, the topics included may vary from what has happened in nearby locations, the injuries or deaths, and the terrorism organizations involved. As has been pointed out, it is extremely difficult to determine what should be covered in a good summary because of the breadth of the document set (McKeown et al. 2002a). At the same time, multi-document summarization has great potential in assisting intelligence analysts in their daily work, where they are likely to receive related messages/reports/documents in groups.

In this chapter, we work with a standard testbed from the Document Understanding Conference (DUC) 2002 data collection (Over and Liggett 2002) to find an answer to the question: What is needed for a good summary? Within the DUC 2002 data collection for multi-document summarization, there are basically two groups of document sets: (I) document sets which consist of closely related documents; and (II) those of highly diverse texts. Intuitively, it should be much more difficult to create a good summary for the document sets in the latter case. We conducted a user study to examine if the difference

between these two groups of document sets has any impacts on judgments about the qualities of the summaries included in the corpus. It was expected that human evaluators could identify this difference easily. Surprisingly, this is not the case. The impact of this difference on human judgments of summarization quality can only be identified by further detailed analysis. As such, we try to determine if differences between two groups are quantifiable. We find that our document graph ($DG$) approach to measure content is capable of doing so (please see the following sections for details of $DG$ and $DG$ generation and their applications). Thus, it allows us to alert a user to not take the summary at face value, i.e., there is a high risk of missing important information in the summary when it is created from a diverse document set.

For the insensitivity exhibited by human evaluators, we believe that this arises from the fact that there were no specific guidelines or tasks required during the summary evaluations. In this situation, we believe that human readers tend to accept summaries which simply contain general information as quality ones. In this chapter, we present and analyze our experiments in order to better understand this phenomenon, how we address it to improve summarization, and better satisfy users' needs.

In the following text, we first introduce the data collection we used in the experiment. Next, we describe the document graph approach that we used to identify the two groups of document sets. Our experiments on evaluating the performance of various automatic summarization systems (summarizers) are followed and a comparison between different summarization ranking approaches currently in use is provided. Finally, we conclude with a discussion of our results.

## DUC 2002 data collection for multi-document summarization

The National Institute of Standards and Technology launched a study on automatic text summarization and evaluation called the Document Understanding Conference (DUC). Since 2001, different types of summarization tasks have been studied, such as single document summarization, multi-document summarization, extract and abstract generation, and headline generation. A data collection is provided to serve as a testbed for state-of-the-art algorithms and systems.

DUC 2002 data collection for multi-document summarization was used in our experiments. It has 59 document sets. Each set has from 5 to 15 documents (with an average of 10). This collection contains articles from the Wall Street Journal, AP newswire, San Jose Mercury News, Financial Times, LA Times, and FBIS records (Over and Liggett 2002). The document sets have been classified into four categories:

- Category 1: Documents about a single natural disaster and created within at most a seven day window (***one event, disaster domain, limited time*** category).
- Category 2: Documents about a single event in any domain created within at most a seven day window (***one event, any domain, limited time*** category).
- Category 3: Documents about multiple distinct events of a single type (no limit on the time window) (***multiple events, unlimited time*** category).
- Category 4: Documents that present biographical information mainly about a single individual (***individual biography*** category).

There were ten teams that submitted summarization outputs generated by their automatic summarizers. For each document set, two model summaries were also created by human assessors. These summaries are extracts. Two different sizes of extracts, 200 words or 400 words, have been generated for each document set. For this chapter, we worked with 200 word extracts in our experiments in order to reduce the time for our user study.

## Using document graph approach to identify different needs for summarization

Generating a summary is actually a process of extracting important relevant information and then presenting it to the user. Accordingly, we use an approach called document graph (*DG*) generation for information extraction and representation, which is described below.

We have been developing the concept, algorithms, and implementation for *DG*s and *DG* generation for several years now. Our *DG* approach was first applied to an intelligent information retrieval application called "Kavanah," which has been used on data collections from the medical domain (Santos et al. 2001, 2003a). It has also been successfully evaluated on standard intelligent information retrieval testbeds including CACM (Nguyen et al. 2004a) and CRANFIELD (Nguyen et al. 2004b). More recently, *DG*s were tested to support intelligence analysts through user-centered intelligent information retrieval on a data collection for weapons of mass destruction gathered from public web sites (Santos et al. 2003b; Cheng et al. 2005) as well as a data collection from the Center for Nonproliferation Studies (CNS, Sept. 2003 distribution) (Santos et al. 2005). Lastly, it has also been applied to automatic summarization evaluation (Santos et al. 2004) and expanded into an intelligence analysis framework (Santos et al. 2005). In summary, we applied *DG*s to various systems for the purposes of capturing and representing the content of documents, queries, and context information in our user model, as well as for similarity comparison. In this study, *DG*s are used to analyze the contents of the summaries.

**Document graph (*DG*)**

A *DG* is a directed graph of concepts/entities and the relations between them. It contains two kinds of nodes, concept/entity nodes and relation nodes. Currently, only two kinds of relations, "*isa*" and "*related to*," are captured for simplicity. The construction of a *DG* is an automated process, which contains following steps: (I) tokenizing a document in the plain text format into sentences (a summary is treated the same as a document); (II) parsing each sentence by using Link Parser (Sleator and Temperley 1993); (III) extracting noun phrases (NPs) from the parsing results; and (IV) generating relations between concepts/entities based on heuristic rules, and put them into the graph format. The most computationally costly step is parsing the sentence, with a complexity of $O(m^3)$ where $m$ is the number of words in a sentence (Sleator and Temperley 1993). For graph generation and comparison, we note that we are working strictly with labeled graphs as opposed to general graph isomorphism. We also greatly improve efficiency by using hashing methods.

We employ three heuristic rules for relation generation which are (Nguyen et al. 2004a; Santos et al. 2004):

- The NP-heuristic: It helps set up the hierarchical relations. For example, from a NP "*folk hero stature*," we generate relations "*folk hero stature - isa - stature*," "*folk hero stature - related to - folk hero*," and "*folk hero - isa - hero*."
- The NP-PP-heuristic: It attaches all prepositional phrases to adjacent NPs. For example, from "*workers at a coal mine*," we generate a relation, "*worker - related to - coal mine*."
- The sentence-heuristic: It relates all the concepts/entities contained within one sentence. The relations created by sentence-heuristic are then sensitive to verbs, since the interval between two noun phrases usually contains a verb. For example, from a sentence "*workers at a coal mine went on strike*," we generate a relation "*worker - related to - strike*." Another example, from "*The usual cause of heart attacks is a blockage of the coronary arteries*," we generate "*heart attack cause - related to -coronary artery blockage*." Fig. 3.1 shows an example of a partial *DG*.

We evaluate the similarity between the two documents based on the DGs that are generated from them. The simple similarity of two *DG*s, $DG_1$ to $DG_2$, is given by Eq. (3.1).

$$Similarity(DG, DG) = \frac{n}{2N} + \frac{m}{2M} \tag{3.1}$$

which is modified from Montes-y-Gómez et al. (2000). In the equation, $N$ is the number of concept/entity nodes in $DG_1$, and $M$ stands for number of relations in $DG_1$; $n$ is the number of matched concept/entity nodes in two *DG*s, and $m$ is the number of matched relations. Two relation nodes

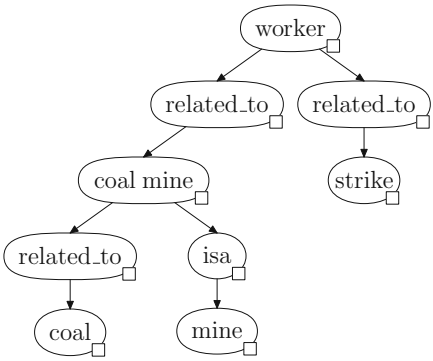



**Fig. 3.1.** A partial *DG*.

are matched only when its parent node and its child node are also matched. Since we might compare two *DG*s that are significantly different in size (for example, *DG*s representing an extract and its source document), we used the number of concept/entity nodes and relation nodes in the target *DG* as $N$ and $M$, instead of the total number of nodes in both *DG*s. Similarity($DG_1$, $DG_2$) provides a percentage of $DG_1$ that is contained in $DG_2$; at the same time, Similarity($DG_1$, $DG_2$) represents how much of $DG_2$ has been covered in $DG_1$. Next, an $F$-score can be calculated by Eq. (3.2) (Van Rijsbergen 1979).

$$F = \frac{2 \times P \times R}{(P + R)} \tag{3.2}$$

In the equation, $P$ is precision and $R$ is recall, where we define $P$ to be Similarity($DG_1$, $DG_2$) and $R$ as Similarity($DG_2$, $DG_1$). The $F$-score is then used as a direct measurement of the similarity between two documents and in ranking the performance of different summarizers. Currently, we weight all the concepts/entities and relations equally.

### Two groups of document sets

We generate a *DG* for each document in the DUC 2002 collection, and then calculate the value of similarity ($F$-score) between each pair of *DG*s within the same document set, and finally obtain the average similarity score. Using our similarity, we can naturally and automatically divide the document sets into two groups: similar document and diverse document sets. The similarity data demonstrated that most of the document sets in *one event, disaster domain, limited time* and *one event, any domain, limited time* categories are of the group of similar document sets, and most of the document sets in *multiple events, unlimited time* and *individual biography* categories are of diverse document group (as shown in Table 3.1).

**Table 3.1.** The average $F$-scores for the document sets in different categories.

| **Category** | One Event disaster Domain Limited Time | One Event any Domain Limited Time | Multiple Events Unlimited Time | Individual Biography |
|---|---|---|---|---|
| $F$-scores | 0.10 | 0.12 | 0.06 | 0.07 |

An example of the group of similar document sets is *D.79.E.200.A* in this collection, which has been classified by DUC as *one event, disaster domain, limited time* category. It contains 9 articles, which are news reports about Hurricane Gilbert: when and where did it happen, and what kind of damage it caused. Obviously, the content of these news articles are highly overlapped. When creating the model extracts for this set, one human assessor picked 7 sentences and another picked 8; between them, 4 sentences are actually exactly the same. The group of diverse document sets covers much broader topics. For example, the set *D.106.E.200.G* has been classified by DUC as *individual biography* category. It contains 10 articles consisting of stories such as: Northeastern University planning to award an honorary degree to Nelson Mandela; imprisoned African National Congress Leaders being allowed to visit Mandela; a family group visited Mandela; the Jewish-American group concerned about Mandela's attitude towards Israel and Palestine Liberation Organization; and news that he would to be released. For this document set, the two model extracts created by human judges are quite different, shown as follows:

Assessor A: ONE MAN'S STRUGGLE *From behind bars, Nelson Mandela has dominated the fight for black rights in South Africa. The following are some of the key events in his life.*

*Mandela joins the African National Congress at age 26, later becoming president of the group's Youth League. When the National Party comes to power in 1948, the ANC begins planning a campaign of civil disobedience to fight the party's apartheid policy. April-June, 1964 Mandela and seven others are sentenced to life in prison. December, 1989 President Frederik W. de Klerk meets Mandela for the first time. World leaders welcomed South Africa's announcement Saturday of Nelson R. Mandela's impending release from prison, and international rejoicing began to build for an event so long awaited by so many. Soweto and other black townships around Johannesburg have been hit by black factional fighting that has killed about 800 people since August. In the 15 months since he walked free from prison, Nelson Mandela has played the chief role in talks with President F. W. de Klerk and the white government. The radicals in the ANC may well use the trial and conviction of Mrs. Mandela to arouse members against Nelson Mandela's policy of peace and negotiation.*

*Assessor F: Mandela the 70-year-old leader of the ANC jailed 27 years ago, was sentenced to life in prison for conspiring to overthrow the South*

*African government. Mandela was allowed to meet last Friday with five other ANC members convicted with him in 1964 of trying to overthrow the government and still imprisoned in Cape Town, according to Dullah Omar, a family friend and attorney. President Pieter W. Botha offers to free Mandela if he renounces violence; Mandela refuses until the government takes steps to dismantle apartheid and grants full political rights to blacks. De Klerk lifts ANC ban and says Mandela will be released. Commonwealth Secretary General Shridath Ramphal said Mandela's safety must be guaranteed by the South African government and that until change there becomes irreversible, the Commonwealth will sustain its pressure for the end of apartheid. The arrival Wednesday of African National Congress leader Nelson Mandela prompted an outpouring of praise and encouragement from politicians and supporters. In the 15 months since he walked free from prison, Nelson Mandela has played the chief role in talks with President F. W. de Klerk and the white government. Mandela and other ANC leaders have been negotiating with President F. W. de Klerk to pave the way for full-scale talks on ending apartheid, the South African system of racial separation.*

Assessor A spent the first half of the summary listing important events in the life of Nelson Mandela, then continued with his role in talks with the white government in South Africa and the problems he was facing. While in Assessor F's summary, the information was more about what was happening at the time of the news concerning Mandela's release, after very briefly mentioning that he is an African National Congress (ANC) leader and has been jailed for 27 years. From the example, it is apparent that when facing such a complex task, people will try to pick up and organize information around their own biases. Just as central is what the judges ignored in their write-ups.

It would be very difficult for an automated summarizer to decide what should be covered in the summary. For the set $D.106.E.200.G$, two machine generated extracts are as follows:

*System 19: 1944 Mandela joins the African National Congress at age 26, later becoming president of the group's Youth League. When the National Party comes to power in 1948, the ANC begins planning a campaign of civil disobedience to fight the party's apartheid policy.*

*January, 1962*

*In August he is charged with incitement and leaving the country illegally; gets five-year jail term.*

*July, 1963*

*Mandela, seven others are charged with sabotage and conspiracy to overthrow the government.*

*April-June, 1964*

*Mandela and seven others are sentenced to life in prison.*

*February, 1990*

*De Klerk lifts ANC ban and says Mandela will be released.*

*On Feb. 11, the nation's best-known black leader becomes a free man at last.*

*World leaders welcomed South Africa's announcement Saturday of Nelson R. Mandela's impending release from prison, and international rejoicing began to build for an event so long awaited by so many.*

*Mandela has embraced Yasser Arafat, chairman of the Palestine Liberation Organization, in public and has compared the struggle of Palestinians to that of black South Africans.*

*Mayor David Dinkins: "Our honored guest...is a George Washington, a Bolivar, a King, a Herzl."*

*System 21: A 3-year campaign that has succeeded in getting Northeastern University to award an honorary degree to jailed South African nationalist leader Nelson Mandela has raised objections from some faculty and students.*

*A representative of the African National Congress said Saturday the South African government may release black nationalist leader Nelson Mandela as early as Tuesday.*

*Four imprisoned African National Congress leaders, convicted in 1964 with Nelson Mandela, were allowed to visit their ANC colleague at his isolated prison house, an activist said today.*

*The largest family group to visit black leader Nelson Mandela since he was jailed 27 years ago gathered in Cape Town for a meeting Tuesday with the prisoner on his 71$^{st}$ birthday.*

*World leaders welcomed South Africa's announcement Saturday of Nelson R. Mandela's impending release from prison, and international rejoicing began to build for an event so long awaited by so many.*

*Mayor David Dinkins warned Jewish groups against protesting the visit of anti-apartheid leader Nelson Mandela to New York City, saying the protests might insult the black community, a newspaper said today.*

*Unidentified assailants hurled a grenade and fired shots today at the home of relatives of Nelson Mandela, killing a baby girl and injuring her parents, police said.*

For System 19, it tried to include the important events in Mandela's life and his fight against apartheid. While, for System 21, it is more likely that it picked one sentence from every article in the document set without any preference or focus.

In general, the group of similar document sets consists of highly correlated documents forming a more coherent focus; and the group of diverse document sets contains more diverse documents covering a wide scope of topics. This would naturally suggest that, it would be more difficult to generate quality summaries for the group of diverse document sets, since summaries would need to cover more topics. Our experiments in the next section demonstrate otherwise.

## Comparing rankings obtained from different approaches for evaluation on the impact of document sets

### Hypothesis

As mentioned above, there are two general groups of document sets in the DUC 2002 collection, which suggests that obtaining good summaries for the group of diverse document sets is more difficult than for those documents belonging to the group of similar document sets. If we assume that some automated summarizers can generate quality summaries for both groups, while others performed worse for one group (more likely group of diverse document sets); then the ranking orders of summarizer performance by human assessors should be different, when they work with document sets that belong to different groups. There is a possibility that all summarizers performed similarly good or bad when working with certain document sets, however, from our examination of the automatically generated summaries, we believe this possibility to be low.

### Experimental procedure

Three different approaches have been used to rank the performance of summarizers.

(i) Exact sentence matching, where system generated extracts were compared to model summaries created by human (which are included in DUC 2002 data).
(ii) Document graph comparison, where automatically generated extracts are compared with the original documents based on the document graphs that generated, and average $F$-scores were calculated for comparison and ranking purposes.
(iii) Human judgments on the quality of the summaries, where every summary is assigned a score by participants. It was expected that human evaluators would clearly recognize the difference between the two types of document sets.

Five people participated in the experiment. They are all graduate students pursuing their PhDs in computer science. One of them is a native English speaker, while the other four participants are from Asia and North Africa. Although English is a second language for four participants, they have no problem understanding general news reports in English since they have been living in the U.S. and studying towards advanced science degrees for at least five years.

Each participant was given 4 document sets, which were randomly picked from the 59 document sets in the 2002 DUC data collection; together with extracts generated by 10 automated summarizers for each document set. The summarizers have been numbered by DUC as systems 16, 19, 20, 21, 22, 24,

**Table 3.2.** Ranking order obtained based on three different approaches (*DG*: document graph comparison, *S*: sentence matching, *H*: Human judger scoring).

| | Group of Similar Document Sets | | | Group of Similar Document Sets | | |
|---|---|---|---|---|---|---|
| System | DG | S | H | DG | S | H |
| 16 | 8 | 9 | 9 | 6 | 8 | 10 |
| 19 | 3 | 1 | 2 | 4 | 6 | 3.5 |
| 20 | 6 | 4 | 4 | 5 | 4 | 5 |
| 21 | 4 | 3 | 3 | 3 | 1 | 1.5 |
| 22 | 9 | 10 | 10 | 10 | 10 | 9 |
| 24 | 7 | 2 | 1 | 1 | 3 | 3.5 |
| 25 | 5 | 7 | 7.5 | 8 | 9 | 6 |
| 28 | 10 | 5 | 7.5 | 9 | 2 | 8 |
| 29 | 2 | 6 | 5 | 2 | 5 | 7 |
| 31 | 1 | 8 | 6 | 7 | 7 | 1.5 |

25, 28, 29, and 31 (Table 3.2). To avoid possible bias, the extracts have been renumbered. For example, in document set 61, summary number 1 is generated by System 22, but in document set 62, summary number 1 is generated by System 20. In total, 20 document sets have been evaluated in the experiment, within which 11 belong to the group of similar document sets and 9 belong to the group of diverse document sets.

The participants also received an instruction on how to evaluate the extracts. They were asked to read carefully through the original document sets and identify the information they think that should be covered in a summary. After reviewing extracts, they are asked to assign a score to each of the extracts using a five-point-scale scoring system, in which from 1 to 5, the quality of the summary would be from very poor to very good, respectively. They were instructed that for a good summary, its quality is based on how well it covers important information, while the order of the sentences and the transition between two sentences in the extracts are not important factors. The time spent on each document set was 53.2 min on average (ranging from 31 min to 257 min, depending on the length the documents).

The rankings obtained based on the different approaches have been compared by using the Spearman rank correlation coefficient ($r$) (Myers and Well 1995).

### Results

Note that three different ranking approaches have been applied and compared in our experiments. Sentence matching compares the machine generated summaries and the model summaries created by human; *DG* approach measures

the information coverage of the extracts; and human evaluation is based on direct human judgment.

The ranking results are shown in Table 3.2, and the correlation data for different ranking approaches are shown in Table 3.3. Different ranking approaches ended up with different results, only the ranking from sentence matching and human judgment for the group of similar document sets demonstrated that they are highly correlated.

For the group of similar document sets, sentence matching and human judgment gave very similar ranking results ($r = 0.92$, $p < 0.01$), which were different from the $DG$ approach results. This demonstrates to us the difference among the three ranking approaches. Both sentence matching and human judgment involve human opinion on what is important, while the $DG$ approach simply measures the information coverage. However, when applied to the group of diverse document sets, the correlation between sentence matching and human judgment was no longer statistically significant (Table 3.3), which may due to the reason that summarization for the group of diverse document sets is a more complex task and, hence, it is more difficult to reach agreement on which summarizers performed better from different ranking approaches.

Both sentence matching and human judgment involve human opinion on what is important, while the $DG$ approach simply measures the information coverage. However, when applied to the group of diverse document sets, the correlation between sentence matching and human judgment was no longer statistically significant (Table 3.3), which may due to the reason that summarization for the group of diverse document sets is a more complex task and, hence, it is more difficult to reach agreement on which summarizers performed better from different ranking approaches.

We assumed that some summarizers could generate quality summaries for both groups of document sets while others could have a fair performance for the group of similar document sets but not for the group of diverse document sets. Thus, the ranking orders for the summarizers would be different when evaluated based on different types of document sets, especially by human

**Table 3.3.** Correlation between the rankings obtained based on three different approaches ($*p < 0.05, **p < 0.01$).

| | | $DG$ | $S$ | $H$ |
|---|---|---|---|---|
| Group of similar documents | $DG$ | – | | |
| | $S$ | 0.21 | – | |
| | $H$ | 0.45 | 0.92** | – |
| Group of diverse documents | $DG$ | – | | |
| | $S$ | 0.54 | – | |
| | $H$ | 0.48 | 0.44 | – |

judgment. The experiments showed that the *DG* approach indicated that there is a significant difference in summarizer performance when working on the group of similar document sets versus the group of diverse document sets. The Spearman rank correlation coefficient of the ranking orders obtained when working with both groups is only 0.45, and the correlation is not statistically significant (Table 3.4). In other words, at least some summarizers performed differently with two different groups of document sets. However, the sentence matching approach could not identify this difference as clearly as the *DG* approach; the correlation coefficient for the two rankings being 0.72, and the *p-value* indicted the correlation is statistically significant. Surprisingly, the human also failed to realize the difference ($r = 0.75$ and $p = 0.006$) (Table 3.4).

Although we can not completely rule out the possibility that all 10 automatic summarizers performed similarly well with the group of similar document sets but not the group of diverse document sets, the low correlation between the two ranking orders for different types of document sets based on *DG* still supported our hypothesis.

Previous studies showed that human judgment differences could be one of the variations that affect the performance scores (Harman and Over 2004). In our experiments, five participants generally agreed with each other on the performance of the automatic summarizers, where in pair-wised comparisons, 7 out of 10 pairs showed the correlation between two human judges was statistically significant (Table 3.5). Also, in this experiment, not only were the extracts provided by the 10 summarizers included, but also included the model summaries. In fact, the model summaries generated by humans were always ranked as the best ones, except in one case where it ranked as the third best;

**Table 3.4.** Correlation between the rankings obtained by the same method on different groups of document sets ($**p < 0.01$).

| *DG* | *S* | *H* |
|---|---|---|
| 0.45 | 0.72** | 0.75** |

**Table 3.5.** Pair-wised correlation analysis on ranking orders obtained based on the scores assigned form different judges (the experiments were performed on all data, including both groups of document sets) ($*p < 0.05, **p < 0.01$).

| Judge | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|
| 1 | – | | | | |
| 2 | 0.37 | – | | | |
| 3 | 0.80** | 0.39 | – | | |
| 4 | 0.60* | 0.62* | 0.64* | – | |
| 5 | 0.92** | 0.45 | 0.67* | 0.57* | – |

which indicated that human participants were doing a good job on evaluating the summary quality.

In our user evaluation experiments, in trying to eliminate possible bias, we told the participants that a quality summary should cover important content, but deliberately avoided directing what kind of content should be considered as important (is it high level analysis? or is it detailed information? or where the focus should be? etc.). Unfortunately, this also resulted in the fact that no specific task was assigned for the summarization process. In this situation, human judges might assign the same scores to the summaries that covered some general topics and detailed information, although their coverage and focuses could be quite different. When generating model summaries for DUC 2002 data collection, judges were likely facing the same problem.

Furthermore, we calculated pair-wised correlations among five human judges separately on the group of similar document set and the group of diverse document sets. When working with the group of diverse document sets, there is only 1 pair in a total of 10 that correlated with each other at a level of statistically significant, with average $r = 0.19$. For the group of similar document sets, there are 4 pairs rankings showing statistically significant correlation, and the average $r$ is increased to 0.47. It suggested that, as individuals, the human judges had more disagreements with each other when working with diverse document sets.

## Discussion and conclusion

Information gathering is the first step in intelligence analysis, it is critical and a challenge for the analysts to be able to collect the right information under often severe time pressure. Multi-document summarization is a very useful technique to assist intelligence analysts in their daily work to find relevant information from the massive amount of available data, in particular, when the analysts are most likely to receive information from different sources. The process of automatic summarization can be decomposed into three steps: analyzing the input text, extracting important information, and synthesizing an appropriate output. As we discussed above, a summarization should be user- and task-oriented. Therefore, understanding humans and their needs are crucial for a good text summarization system.

Various technologies have been attempted for generating summaries, such as term frequency (Luhn 1958), predefined templates (McKeown et al. 2002b), and latent semantic indexing (Gong and Liu 2001). However, much less has been done in identifying how individual differences affect the perception of a good/quality summarization. There has been a closely related effort that used the utility of query biased summaries to help users identify relevant documents (Sanderson 1998), in which Local Context Analysis (LCA) has been used to expand topics contained in the baseline summaries with additional words and phrases. LCA is a technique for automatic query expansion using

pseudo feedback. It examines the context surrounding the topic terms in the top ranked documents for query expansion. It has been found that users could judge the relevance of documents based on their summaries, almost as accurately as if they accessed the full texts. Sakai and Masuyama (2004) proposed an interactive approach for multi-document summarization realizing a user's summarization need. Their system extracts keywords from a document set and shows k best keywords with scores to a user on the screen. The user then has the opportunity to select those that reflect his/her information needs. The approach helped improve system performance.

Our *DG* approach can automatically identify if a summary is created from a broad and diverse document set (as opposed to a highly focused set). It then can serve as an alert to the user when there is a high risk of missing information the user may be interested in. In our experiments, human evaluators, as a group, were not very sensitive to this difference; however, as individuals, they had more disagreement with each other when working with document sets covering diverse content. This result, again, suggested that each individual has his or her own information needs. We also believe that the difference in individual information needs would be much more obvious when they are working on certain tasks or have specific goals.

More importantly, the difference between these two groups of document sets and the human response to it actually reflects the requirements for a good summary. People would generally agree with each other on the most common content needed to be covered in a summary; however, they do have their own individual interests. Therefore, a quality summary should cover the most general topics and also various related pieces of information from different aspects or details that are relevant to the user's specific needs. Thus, for *user-centered summarization*, an important task is to identify what kinds of the details in information needs to be retrieved and kept in the summary to satisfy specific user interests according to the user's knowledge and the specific tasks he or she has.

McKeown et al. (2005) found that for multi-event input document sets, the difficulty in generating a quality summary comes from the breadth and diversity of the documents in the sets. Sometimes, even humans seem to have a hard time determining how to produce a good summary. In this situation, their summaries were often quite different from each other (as showed by the example above). The method proposed to handle this situation is to apply different strategies/techniques on different categories of documents in the input sets. For DUC 2002 data, they used four different strategies, one for single events, one for multiple related events, one for biographies, and one for discussion of an issue with related events (McKeown et al. 2002a). The problem associated with this approach is that, the categories are classified based on the content of the documents. There should be more categories than the four being used in DUC 2002 collection, and a method is also needed to automatically classify documents. McKeown et al. (2002b) suggest creating document sets and define a set of criteria by automatically filtering and clustering large

online data, while it is a manual process in DUC 2002. In addition, the user's needs have not been considered.

*DG* approach has the potential to overcome this problem. *DG* generation is a process of information extraction and representation. As a result, the important concepts, entities and relations in the text are captured. We can then generate a summary from each *DG* of each document. First, a core of a summary (also in *DG* format) that contains the most general information is needed. For document sets, the core can be constructed by majority vote; while for a single document, the core can be created based on the weights of the relations. We then expand the core *DG* by inserting relevant relations based on the underlying graph structure. The most relevant relations would be decided with the help from a user model module that captures a user's knowledge and foci/interests, which answers to the challenge that a good summary should target who reads it. Finally, we generate a summary based on this *DG*, which should be biased towards a user's individual interests, and better meet his or her information needs. This approach naturally fits into our previous efforts on using a user model to provide proactive assistance in information searching process. We are currently pursuing this effort and are focused on formally defining the appropriate graph theoretic measures for expanding *DG*s from multiple documents.

## Acknowledgments

This work was supported in part by the Advanced Research and Development Activity (ARDA) U.S. Government. Any opinions, findings and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the U.S. Government. A preliminary version of this work was presented at the HICSS 39 Workshop (Zhao et al. 2006).

## References

Cheng, J., J. Dang, R. Emami, H. Goradia, J. Huang, M. Huhns, L. Kerschberg, H. Nguyen, E. Jr. Santos, M. Valtorta, H. Wang, S. Xi, and Q. Zhao. 2005. A cognitive framework for user modeling, reuse of prior and tacit knowledge, and collaborative knowledge services. *Proceedings of the 38th Annual Hawaii International Conference on Systems Sciences (HICSS-38)* IEEE Press 293c. Big Island, HI.

Elhadad, N. 2004. User-sensitive text summarization. *AAAI Doctoral Consortium* 987-988. San Jose, CA.

Goldstein, H. 2006. Modeling terrorists, new simulation could help intelligence analysts think like the enemy. *IEEE Spectrum* 26–33.

Gong, Y., and X. Liu. 2001. Generic text summarization using relevance measure and latent semantic analysis. *Proceedings of the 24th Annual International ACM*

*SIGIR Conference on Research and Development in Information Retrieval* 19-25. New Orleans, LA.

Harman, D., and P. Over. 2004. The effects of human variation in duc summarization evaluation. *Proceedings of ACL 2004, Workshop on Text Summarization Branches Out* 10–17. Barcelona, Spain.

Luhn, H. P. 1958. The automatic creation of literature abstracts. *IBM Journal* 159–165.

Mani, I., and M. T. Maybury, (Eds.) 1999. *Advances in Automatic Text Summarization*. Cambridge, MA: The MIT Press.

McKeown, K., R. Barzilay, and S. Blair-Goldensohn. 2002a. The Columbia multi-document summarizer for DUC 2002. *Document Understanding Conference*.

McKeown, K., R. Barzilay, D. Evans, V. Hatzivassiloglou, J. L. Klavans, A. Nenkova, C. Sable, B. Schiffman, and S. Sigelman. 2002b. Tracking and summarizing news on daily basis with Columbia's Newsblaster. *Proceedings of 2002 Human Language Technology Conference (HLT)*. San Diego, CA.

McKeown, K., R. J. Passonneau, and D. K. Elson. 2005. Do summaries help? A task-based evaluation of multi-document summarization. SIGIR'05. Salvador, Brazil.

Montes-y-Gómez, M., A. Gelbukh, and A. Lópes-López. 2000. Comparison of conceptual graphs. *Proceeding of MICAI-2000 - 1st Mexican International Conference on Artificial Intelligence*. Acapulco, Mexico.

Myers, J. L., and A. D. Well. 1995. *Research Design and Statistical Analysis* 488–490. New Jersey: Lawrence Erlbaum Associates.

Nguyen, H., E. Santos Jr., Q. Zhao, and C. Lee. 2004a. Evaluation of effects on retrieval performance for an adaptive user model. *Adaptive Hypermedia 2004: Workshop Proceedings - Part I* 193–202. Eindhoven, the Netherlands.

Nguyen, H., E. Santos Jr., Q. Zhao, and H. Wang. 2004b. Capturing user intent for information retrieval. *Proceedings of the 48th Annual Meeting for the Human Factors and Ergonomics Society (HFES-04)* 371–375. New Orleans, LA.

Over, P., and W. Liggett. 2002. Introduction to DUC-2002: An intrinsic evaluation of generic news text summarization systems. *Document Understanding Conference*. `http://duc.nist.gov/`.

Sakai, H., and S. Masuyama. 2004. A multiple-document summarization system introducing user interaction for reflecting user's summarization need. *Working Notes of NTCIR-4*. Tokyo.

Sanderson, M. 1998. Accurate user directed summarization from existing tools. *Proceedings of the 7th International Conference on Information and Knowledge Management* 45–51. Bethesda, MD.

Santos, E. Jr., G. Johnson, H. Nguyen, P. Thompson, and Q. Zhao. 2005. A cognitive framework for information gathering with deception detection for intelligence analysis. *Proceedings of 2005 International Conference on Intelligence Analysis*. McLean, VA.

Santos, E. Jr., A. A. Mohamed, and Q. Zhao. 2004. Automatic evaluation of summaries using document graphs. *Proceedings of ACL 2004, Workshop on Text Summarization Branches Out* 66–73. Barcelona, Spain.

Santos, E. Jr., H. Nguyen, and S. M. Brown. 2001. Kavanah: An active user interface. *Information Retrieval Agent Technology* 412–423. Maebashi, Japan.

Santos, E. Jr., H. Nguyen, Q. Zhao, and E. Pukinskis. 2003a. Empirical evaluation of adaptive user modeling in a medical information retrieval application.

P. Brusilovsky, A. Corbett, and F. de Rosis (Ed.). *Lecture Notes in Artificial Intelligence 2702: User Modeling 2003* 292–296. Springer, New York.

Santos, E. Jr., H. Nguyen, Q. Zhao, and H. Wang. 2003b. User modelling for intent prediction in information analysis. *Proceedings of the 47th Annual Meeting for the Human Factors and Ergonomics Society* 1034–1038.

Santos, E. Jr., H. Nguyen, Q. Zhao, and H. Wang. 2005. Impacts of user modeling on personalization of information retrieval: An evaluation with human intelligence analysts. *4th Workshop on the Evaluation of Adaptive Systems, in conjunction with UM'05* 27–36.

Sleator, D. D., and D. Temperley. 1993. Parsing English with a link grammar. *Proceedings of the 3rd International Workshop on Parsing Technologies* 277–292.

Van Rijsbergen. 1979. *Information Retrieval*, 2*nd* ed. Butterworths, London.

Zhao, Q., E. Santos Jr., H. Nguyen, and A. A. Mohamed. 2006. What is needed for a good summary? – Two different types of document sets yet seemingly indistinguishable to human users. Presented at *the Hawaii International Conference on System Sciences* (HICSS 39). Koloa Hawai, HI.

# 4

# A Prototype Search Toolkit

Margaret M. Knepper, Kevin L. Fox, and Ophir Frieder

**Summary.** Information overload is now a reality. We no longer worry about obtaining a sufficient volume of data; we now are concerned with sifting and understanding the massive volumes of data available to us. To do so, we developed an integrated information processing toolkit that provides the user with a variety of ways to view their information. The views include keyword search results, a domain specific ranking system that allows for adaptively capturing topic vocabularies to customize and focus the search results, navigation pages for browsing, and a geospatial and temporal component to visualize results in time and space, and provide "what if" scenario playing. Integrating the information from different tools and sources gives the user additional information and another way to analyze the data. An example of the integration is illustrated on reports of the avian influenza (bird flu).

## Introduction

While many today experience the reality of the information explosion of recent years, the School for Information Management and Systems (SIMS) of the University of California at Berkeley has attempted to measure it. In studies conducted in 2000 (Lyman and Varian 2000) and 2003 (Lyman and Varian 2003), SIMS researchers attempted to measure how much information was produced in the world each year. The result—SIMS estimated that new stored information grew about 30% a year between 1999 and 2002—more than doubling in that three year time period.

Thus, there has been a vast increase in the total volume of information produced and stored digitally in the world in recent years. However, total data consumption has not kept pace with the growth in new information (Brown 2003). As the growth in the volume of new information continues, consumers of information are, therefore, faced with a growing need to filter and select relevant information. To do this, consumers need to identify relevant information more easily before selecting what they want to consume.

Information processing necessitates the composition of data collection, search, and understanding. Our efforts focus on developing an integrated

S. Argamon, N. Howard (eds.), *Computational Methods for Counterterrorism*, DOI 10.1007/978-3-642-01141-2_4,

toolkit that supports all levels of users' information needs—knowledgeable, familiar, and novice. This requires, at a minimum, the integration of efficient search, domain customization, and geospatial visualization for scenario playing.

We caution the reader that the effort described herein is, by no means, complete. Furthermore, we note that we are not attempting to provide a definitive statement as to which tools comprise a minimal set for information discovery. Our goal is simply to illustrate one toolkit and demonstrate the type of information discovery made possible by such a system.

The described "analyst toolkit" was developed as part of Harris Corporation's Internal Research and Development effort in information management. In this toolkit, an assortment of tools is provided to assist users in their information discovery endeavors. Specifically:

- Keyword search is augmented with domain customization using Query Improvement Elevation Technique (QUIET), to incorporate topic vocabularies into the search and re-ranking of results, biasing by domain knowledge. A detailed description of QUIET is available on-line in Knepper et al. (2005).
- Navigational pages enable users to browse the document collection, allowing each user to view documents based on source, date, document theme, or other metadata properties.
- Geospatial and temporal visualization allow for the integration of large data sets and for the user to see changes over space and time. Additionally, this facilitates "what if scenario game playing" to analyze the data.

The navigation and QUIET tools provide a mechanism to capture domain expertise and share this expertise among users. Search tools return an inordinately sized result set without providing means to easily determine how to produce a better query, hence potentially higher overall search accuracy (Grossman and Frieder 2004). Providing the user with the ability to easily review the words used within the retrieved documents, compare terms from previous queries, and quickly identify new terms to incorporate into the query aides the user, consequently improving accuracy. Likewise, visualizing the response and providing an ability to impose various constraints, e.g., only consider certain locations or certain demographics, further enhances user understanding.

## System components

The information discovery process differs depending on the sophistication of the user. Clearly, a specialist looking for a highly in depth treatise of an issue demands a different exploration tool then a novice looking for a broad but technically shallow overview of the subject. That said, at times, even the expert may wish a broader overview, and the novice may desire a detailed

exposition for a specific topic (Marchionini 2006). We provide a search enhancement tool for detailed searching called QUIET, a navigational capability for general browsing, and a geospatial and temporal visualization option. An advantage of our toolkit is the integration of these tools. We demonstrate the practicality of this integration and its benefits via our avian influenza (bird flu) scenario. In this section, we describe our key components.

## QUIET

Search tools return an inordinately sized result set without providing means to easily determine how to produce a better query, and hence, potentially higher overall search accuracy. Providing the user with the ability to easily review the words used within the retrieved documents, compare terms from previous queries, and quickly identify new terms to incorporate into the query aides the user, consequently improving accuracy.

The *Query Improvement Elevation Technique (QUIET)* improves the results sets returned by information retrieval engines through automatic algorithms or interaction with the user. Domain vocabularies are created, maintained, and altered as needed to capture the knowledge within a given area of interest. By sharing these vocabularies among the various users, the domain understanding of one user can be capitalized on by other users.

QUIET improves the accuracy of current systems that use keyword searches and is intended for queries that are repeated over time. Vocabularies are developed for specific user domains, e.g., airport security, terrorism, nuclear weapons proliferation, etc., and define the relevant/irrelevant words for each of these domains. A vocabulary represents expert knowledge for the domain that can be shared with other users to improve query results.

Figure 4.1 illustrates the concept—selection of an appropriate domain vocabulary that can be used to distinguish the type of "Piracy" of interest to a particular user. As new ideas are introduced to the domain, QUIET information can continue to be updated by the analyst.

A vocabulary can be built several ways. Namely, it can be built on the fly by selecting the words from the top $N$ documents from the search tool retrieved list or using relevant documents of any type, e.g., documents on "newly found" disks, monitored e-mail traffic, or simply available documents on a specified domain. The vocabulary can continue to grow by adding new relevant documents to the vocabulary as they are identified.

Topics contain information about the relevancy of words and documents. A topic can point to any vocabulary. Any number of topics can point to a vocabulary. Since topics can cover multiple vocabularies, pointing the topic to a different vocabulary can potentially help with relevant document identification.

The vocabulary is used to calculate word statistics based on word counts in the documents and the number of documents containing the words. This approach markedly improved search precision of the top 30 documents retrieved

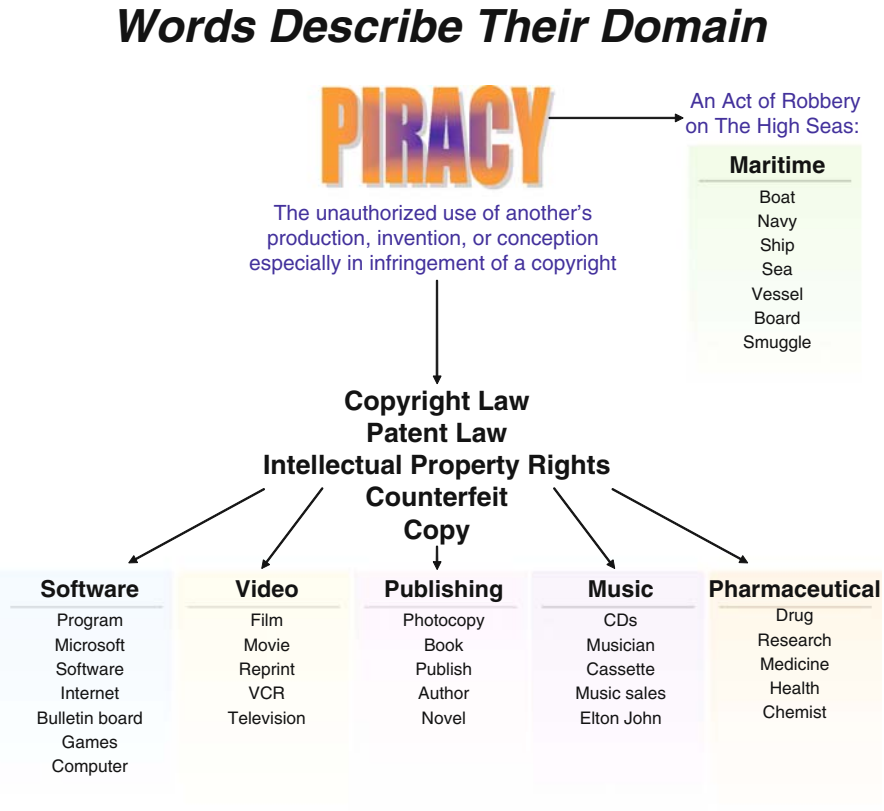



**Fig. 4.1.** The type of "Piracy" of interest to a user can be distinguished from other forms of piracy by the selection of an appropriate vocabulary.

for both an academic and a commercial search engine we tested without the need to modify either engine. We focused on the top 30 documents, as it was perceived as a more realistic reflection of an analyst's attention span/patience. Generally, the best results were obtained by combining the results of QUIET and the search engine.

Multiple search engines and languages may be used throughout the environment. This combined approach allows for best of breed tools to be used for specific data sets or queries. QUIET can fuse/correlate/consolidate results from multiple search engines, queries or users. QUIET provides a centralized interface to review those results. The user can view retrieved documents and words in different ways, gaining new perspectives on the data. The QUIET structure provides several benefits to the end user. The user is able to:

- **Share domain models**. QUIET allows the analyst to articulate and communicate expertise within their domain. Custom vocabularies and topics

improve query results and allow the analyst to share their expertise with other analysts; thus, potentially allowing analysts to fill new roles as they analyze new domains with confidence.

- **Identify new ideas**. The vocabulary makes it easy to identify new terms in the documents. These terms may represent novel ideas in the topic. They may also rapidly identify irrelevant documents. The analyst can easily update the vocabulary with these new terms and start to use them in his or her queries.
- **Domain insight**. The analyst can interact with the system to identify relevant/irrelevant words and documents. This interaction improves the understanding of the domain and further improves query results.

### Navigation

The navigation technique is applied when the user is unsure of which keywords to use to identify relevant documents, allowing the user to navigate (browse) through the data to identify different areas of interest. In Fig. 4.2, we show an example of a navigation screen, implemented with Siderean Software's Seamark Navigator. This screen allows the user to navigate through the documents based on the document's themes, i.e., quarantine, human catching avian influenza, avian influenza vaccine, etc. An approach called facet-based navigation (Hearst 2006) can be used to narrow down the universe of potential documents through selection of values for facets of the information being sought. A facet is a fundamental category (or set) that can be used to describe a document or other digital asset, such as subject, date created, domain, language, etc. A user can browse the collection by selecting values for facets (or categories or sets) and examining the number and content of documents that continue to match the facet values. Through progressive navigation, a user can gain an overall view of the information space and discover relevant documents. The facet-based navigation technique can be used in conjunction with more traditional search techniques to broaden or refine the search terms being used and to find relevant documents that do not contain any of the expected keywords. Different levels of metadata tagging can be used for this technique.

Digital objects or assets are associated with categories (represented as members of a set) based on associated metadata. Navigation enables filtering based on iterative refinement of the categories of interest, that is, the iterative refinement of sets (membership in subsets) and the intersection of sets. For example, we mapped a set of 1,384 news reports relating to terrorist incidents into a number of categories, such as Form of Attack, Target of Attack, Location of Attack, Terrorist—Name, Terrorist—Location, Year of Attack, and Number of Fatalities. Suppose a user is interested in bombing as the form or method of attack. By selecting that category or set, the number of relevant documents is reduced to 616. That number can be further reduced. Suppose further that the user is interested in attacks that took place in the Middle East (361 documents), in the year 2002 (207) and targeted at civilians (380).

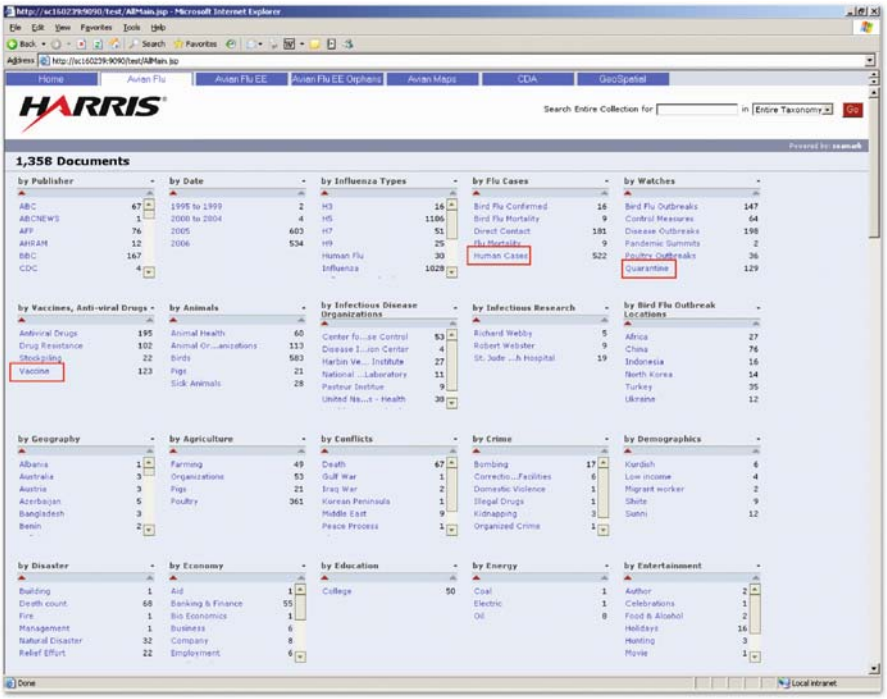



**Fig. 4.2.** A custom navigation screen for an avian flu analyst. Categories include: Different influenza, headlines—bird flu confirmed, quarantine, antiviral drugs, vaccine research, organizations and people to watch, and outbreak areas.

Only 16 terrorist incident reports fit all four of those categories, that is, are members in each of those sets, and thus fall in the intersection of those sets. So the user has quickly reduced the number of reports to be examined from 1384 to 16, a much more manageable number, see Fig. 4.3.

The methodology behind this search technique is that different users both think about and know different aspects about the information they are searching for. This gives these users multiple ways to navigate the same data. The navigation tool does not require users to have existing knowledge about the subject. Exploration of the data is also simplified using categories, allowing them to easily enter and back out of different facets. They are free to explore any aspect of the problem they want.

The difficult aspect of meaningful navigation is creating the metadata, or aspects, about the resources to be browsed (navigated through) and later searched. Since the navigation system solely uses the metadata to navigate to the resources, it is closely tied to the meta-data creation process. We currently use a variety of entity extractors and taggers to accomplish most of the

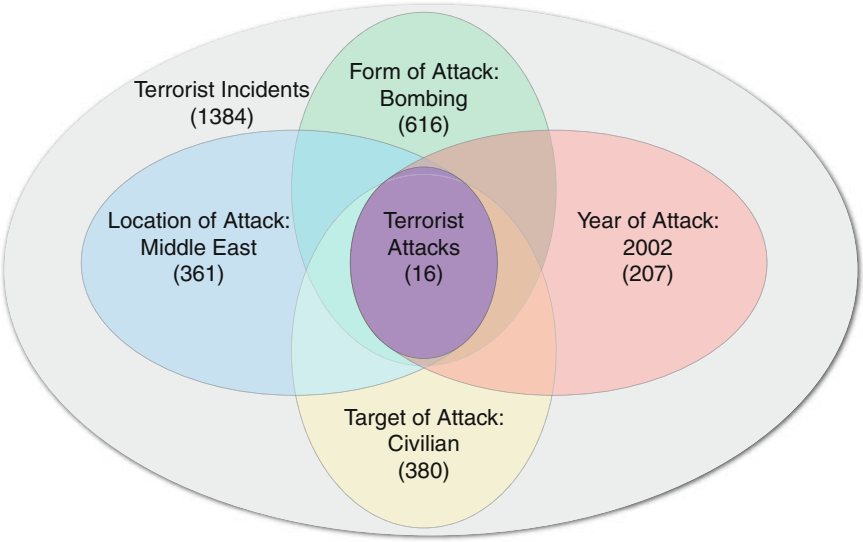



**Fig. 4.3.** Navigation of terrorist incident reports is accomplished by iteratively selecting categories of interest (set membership).

metadata creation. Once the user has navigated to their known aspects of the data, then keyword searches can be done to search the reduced data set.

### Geospatial and temporal visualization for analysis and exploitation

Geospatial and temporal visualization is accomplished through use of ESRI ArcIMS capabilities (ESRI's ArcIMS March 25, 2006). We use ArcIMS as it provides standard map navigation functions—zoom in, zoom out, zoom to full extent, pan, etc. Custom layers are created on top of the map containing dynamically updated geospatial information including graphics and text. Different map layers can be toggled on and off by the user.

### Document building process

The processing and displaying of a document's themes and data through the system is shown in Fig. 4.4. The document is ingested into the system. Documents are indexed, so the user can perform keyword searches. Metadata comes from a variety of sources; the integrated result is shown in the navigation pages. The top phrases are identified in the document as a form of high-level theme detection. If a document has a theme of interest (i.e., nuclear, avian flu, terrorism), it is run through the entity extraction tool to obtain more detailed metadata information.

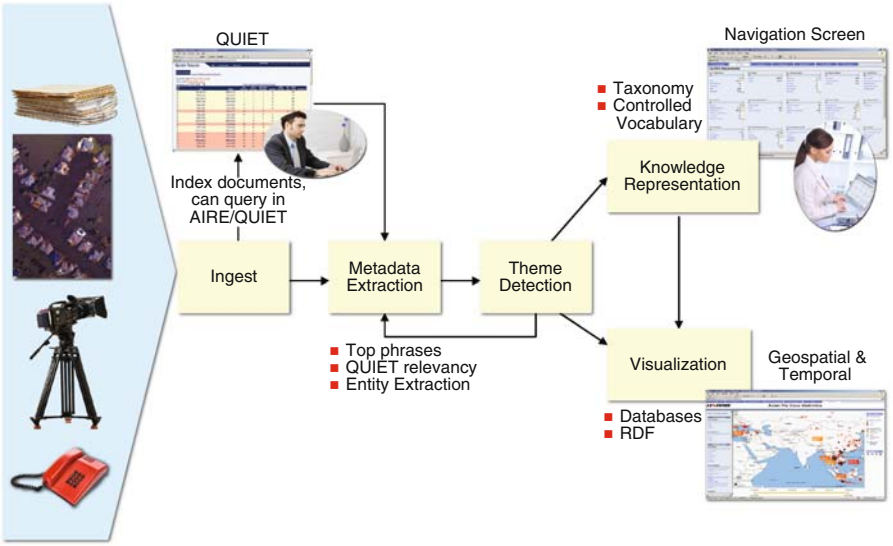



**Fig. 4.4.** Document/content processing extracts metadata and themes to facilitate user navigation and the issuance of queries.

The navigation screens represent the document themes. A variety of taxonomies are built with controlled vocabulary terms pointing to the taxonomy. The document's metadata are compared to the controlled vocabulary to enable the taxonomy navigation. Additionally, information can be pulled from the documents and stored in a database to be displayed geospatially.

## Integration of the tools

We illustrate the use of the above technologies to provide multiple ways to search for information regarding the avian influenza, H5N1-type (alias the bird flu). Data related to avian influenza were collected from the Web, and in particular from RSS news feeds.

### Navigation screens

In the case of bird flu, there are many different areas to watch. The initial navigation screen was developed to reflect those wide ranges of interest. In Fig. 4.2, we presented an overview of the categories for all the data collected and associated with the avian influenza. The categories of interest include: Flu Cases, Watches (Quarantines, Outbreaks), Animals, Vaccines and Infectious Disease Organization. Documents can be associated with multiple categories. Any digital asset with metadata tags can be associated with any category.

Documents identified in categories of interest were run through an Entity Extraction (EE) tool to extract additional information—names, organizations, etc. A user can use the point-and-click navigation to drill down to the desired level of detail.

The navigation screens are intended to help people understand the available data set. Novice users can use the navigation screens to start understanding the domain. More advanced users can use the navigation screen to get a feel for the type of data that are available.

## QUIET

Navigation provides a quick overview of the data. However, it does not always detect and provide all the theme/meaning of the document. Eventually, the user begins to comprehend the domain and may want to perform specific searches for the data of interest. Since our focus, in this example, is on the avian influenza specifically, the user needs to be informed of the different types of influenza that may be occurring at the same time. What is happening? Is it the avian influenza or is something else that is happening? An area to watch may be the "Unknown Disease"—is there something happening that needs to be confirmed? A variety of techniques can be utilized to identify the documents to the end user. Sometimes usage of a specific tool might help find documents of interest. For example, a bird flu analyst may be interested in documents discussing "unknown diseases." The document may be found through theme detection. Key phrases like "mystery illness" or "unknown illness" are easy to identify.

Keyword searches can also identify documents and their themes. For example, QUIET identified documents corresponding to the different areas in avian flu. A vocabulary was built for avian flu and then specific topics were built for unknown illness, avian flu vaccine, quarantines, and drug resistance to bird flu. Each of these topics customizes the avian flu vocabulary. In Fig. 4.5, we show the bird flu vocabulary for unknown illness and avian flu vaccine. The word relevancy is set differently for each topic to reflect the importance of the words. In the case of unknown diseases, the word vaccine is not relevant since the disease is unidentified and a vaccine cannot be applied.

In Fig. 4.6, we illustrate a display of the documents as ranked by QUIET. In this case, a document is ranked fourth by QUIET and ninety-third by the search tool. Color-coding in the display is as follows: Yellow represents relevant documents, and orange represents irrelevant documents as rated by the analyst. In this document, the boy's death is described as:

- "Was not killed by bird flu."
- "It is not clear what killed."
- "The disease is not caused by bird flu or any other flu virus."

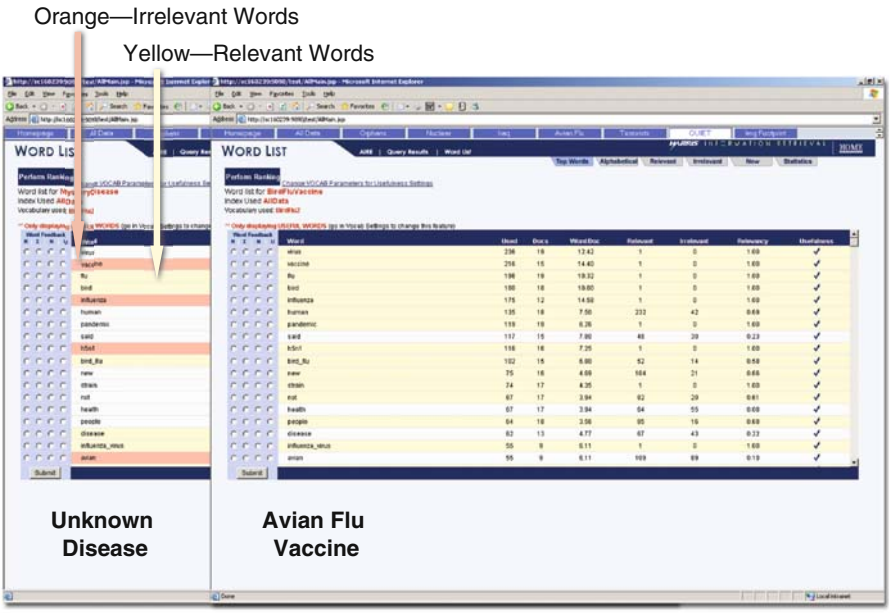



**Fig. 4.5.** Avian flu—QUIET—The bird flu vocabulary is set differently for the topics—Mystery disease and bird flu Vaccine. Words are given different relevancies based on the topic. *Yellow* indicates relevant words, *orange* irrelevant.

There are no keyword phrases like "unknown illness" or "mysterious illness" used to describe this unknown death. By building the QUIET vocabulary, this document appeared within the top 5 documents in the QUIET list, and most likely would have been reviewed by the analyst. Documents ranked by QUIET as relevant or unknown relevancy appear as a document theme in the navigation pages. This document was not identified as "unknown illness" theme by looking at the top phrases for the document. However, it was associated with unknown illness through the QUIET ranking. Results from QUIET are displayed in the navigation window as shown in Fig. 4.7 as a document theme.

### Geospatial-temporal visualization of bird flu outbreak locations

In Fig. 4.8, we present a sequence of geospatial and temporal visualization views. The user can observe changes over time by changing the date (using the slider bar on the bottom of the display). As shown, the number of outbreaks increases as time progresses, and the disease spreads north and west.

Clearly, the presented screen shots do not serve as proof of concept. However, they do demonstrate the potential value of our approach. Currently, we are developing additional domain vocabularies so as to provide additional canned environments for our toolkit.

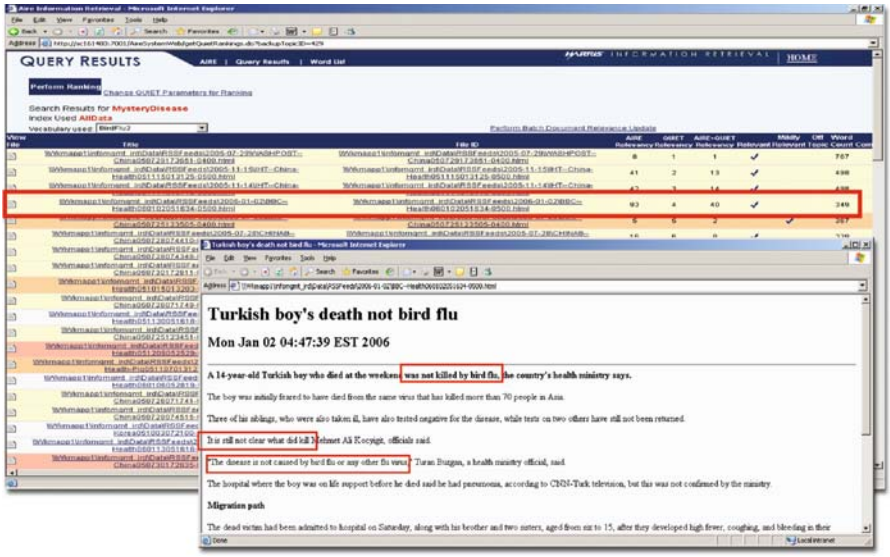



**Fig. 4.6.** Avian flu—in QUIET, a "Mysterious Illness" topic was performed and a vocabulary was built. The document describing a 14 year old boy's death as "not killed by bird flu" and "it is not known what did kill" was ranked #4 in mysterious illness.

## Potential future enhancements

A variety of techniques are required to help people find their information needs. Different tools present these techniques to the user. Implementation of these systems requires behind the scenes work—capture expertise, represent the knowledge, create processes to move the data through the steps to get the final products. On the back end, it requires a focus on the customer needs, a willingness to solicit user feedback to get the correct representations for their needs. New data and constant changing requirements require the system to be maintained and flexible enough to allow the new data representations.

Briefly stated, the ideas for key additions to QUIET include:

- **Multilingual QUIET**. QUIET needs to be tested in a multilingual environment. QUIET builds vocabularies, and the technique used to do so is language independent. In the future, vocabularies could be built in different languages and then merged to help the analyst find relevant documents in multiple languages.
- **Domain identification**. The vocabularies represent domain expertise. Mathematical models can be used to compare the document to the topic vocabulary to determine the document domain. This would provide additional information for the document theme on the navigation pages.

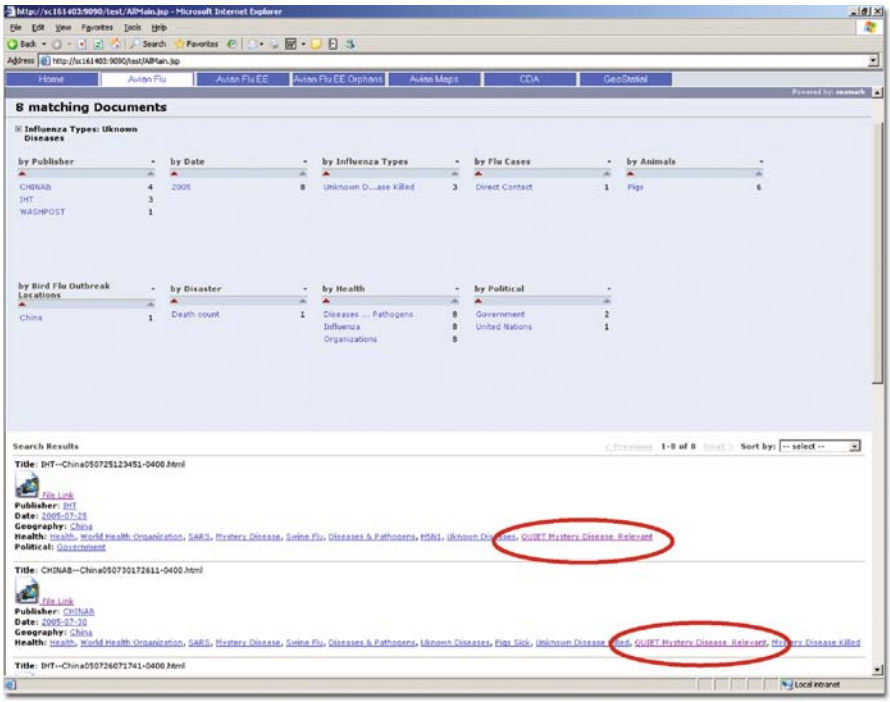



**Fig. 4.7.** Avian flu—results from the search engine can be included in the navigation.

## Conclusions

We developed an integrated information processing toolkit that provides all levels of users with a variety of ways to view their information. New information can be added to the system. Current views include:

- Keyword search based algorithms to improve results
- Domain specific ranking system that allows for adaptively capturing topic vocabularies to customize and focus the search results
- Navigation pages for browsing
- Knowledge representation using taxonomy and ontology
- Visualization of ontology relationships
- Geospatial and temporal component to visualize results in time and space and provide "what if" scenario playing.
- Entity extraction
- Operates in a SOA environment

Integrating the information from different tools and sources gives the user additional information and another way to analyze the data. The goal of our

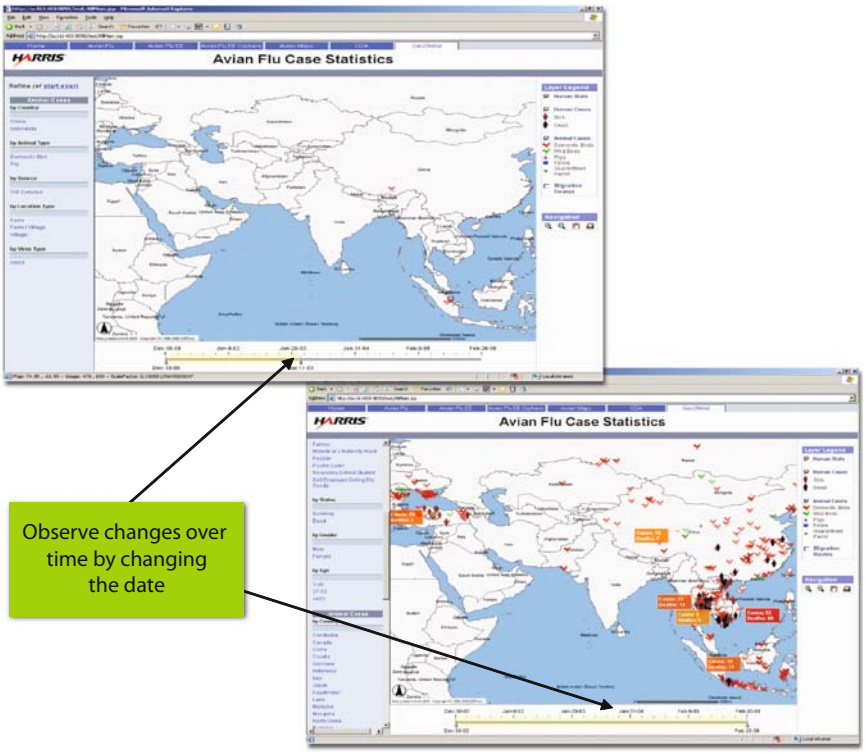



**Fig. 4.8.** Geospatial and temporal visualization—observe changes over time by changing the date.

system is to assist the user in finding the right data at the right time, enabling the user to:

Share domain models. Analysts can articulate and communicate expertise within their domain. Custom vocabularies, relationships, and topics improve query results and allow analysts to share their expertise with other analysts; thus, potentially allowing analysts to fill new roles as they analyze new domains with confidence.

Identify new ideas. The different views allow the user different ways to look at the data. These multiple views of vocabulary, relationships, geospatial, and time may generate new ideas on the topic. These views may also rapidly identify irrelevant information. Analysts can easily update information and start to use it in their domain.

Obtain domain insight. Analysts can interact with the system to identify relevant/irrelevant words and documents. This interaction improves the understanding of the domain and further improves results.

## References

Brown, P. 2003. *Information Architecture with XML*, 5. John Wiley & Sons, Ltd, New York.

ESRI's ArcIMS. March 25, 2006. http://www.esri.com/arcims

Grossman, D. A., and O. Frieder. 2004. *Information Retrieval: Algorithms and Heuristics*, 2*nd* ed. Springer Science and Business Media, New York.

Hearst, M. A. 2006. Clustering versus faceted categories for information exploration. *Communications of the ACM* 49(4).

Knepper, M. M., K. L. Fox, and O. Frieder. 2005. Query Improvement Elevation Technique (QUIET). *International Conference on Intelligence Analysis (IA05)*. McLean, VA. https://analysis.mitre.org/proceedings/Final_Papers_Files/188_Camera_Ready_Paper.pdf

Lyman, P., and H. R. Varian. 2000. How much information. School for Information Management and Systems (SIMS) of the University of California at Berkeley. http://www2.sims.berkeley.edu/research/projects/ how-much-info/

Lyman, P., and H. R. Varian. 2003. How much information. School for Information Management and Systems (SIMS) of the University of California at Berkeley. http://www.sims.berkeley.edu/how-much-info-2003

Marchionini, G. 2006. Exploratory search: From finding to understanding. *Communications of the ACM* 49(4).

Siderean Software's Seamark Navigator. August 29, 2006. http://siderean.com/products_suite.aspx

# Part II

# Text Analysis

# 5

# Unapparent Information Revelation: Text Mining for Counterterrorism

Rohini K. Srihari

**Summary.** Unapparent information revelation (UIR) is a special case of text mining that focuses on detecting possible links between concepts across multiple text documents by generating an evidence trail explaining the connection. A traditional search involving, for example, two or more person names will attempt to find documents mentioning both these individuals. This research focuses on a different interpretation of such a query: what is the best evidence trail across documents that explains a connection between these individuals? For example, all may be good golfers. A generalization of this task involves query terms representing general concepts (e.g. indictment, foreign policy). Previous approaches to this problem have focused on graph mining involving hyperlinked documents, and link analysis exploiting named entities. A new robust framework is presented, based on (i) generating concept chain graphs, a hybrid content representation, (ii) performing graph matching to select candidate subgraphs, and (iii) subsequently using graphical models to validate hypotheses using ranked evidence trails. We adapt the DUC data set for cross-document summarization to evaluate evidence trails generated by this approach

## Introduction

Open source document collections reflect diverse sources and authors; they often reveal interesting information other than what is explicitly stated. The goal of information analysts is to sift through these extensive document collections and find interesting links that connect facts, assertions or hypotheses that may be otherwise missed. What is required is a set of automated tools that will expose such links, or at least generate plausible patterns. We refer (Srihari et al. 2007) to this special case of text mining as unapparent information revelation (UIR).

Currently, analysts perform this task with assistance from social network analysis (SNA) and graph or pattern matching tools. Such techniques while powerful are fragile: (i) they rely on an information extraction system to accurately tag key entities and relationships, (ii) they do not take into account

S. Argamon, N. Howard (eds.), *Computational Methods for Counterterrorism*, DOI 10.1007/978-3-642-01141-2_5,

more general concepts such as 'trucking industry', but are limited to named entities, and (iii) they require users to anticipate and predefine specific scenarios of interest; this in turns involves building complex *domain models*. This process is a chokepoint. In order for domain models to be effectively used in pattern matching, they should reflect the *data model* which is derived from processing a corpus. Analysts cannot anticipate all types of event patterns leading to a specific scenario: the system should discover these patterns! On the other hand, analysts are able to specify the set of concepts typically involved in the pattern being sought. The goal of this research is to take as input such broad queries, referred to as *concept graph queries*, and generate the corpus-specific hypothesis (pattern) that corresponds to it. Each hypothesis will typically involve more concepts than the original query, and will be backed up by a textual *evidence trail* that shows the logical connection between the original input concepts. Multiple hypotheses, along with multiple evidence trails may be generated. Such a technique is designed to reduce the burden on analysts to do cumbersome modeling.

Figure 5.1 illustrates a typical concept graph query. The analyst is looking for patterns involving a religious leader participating in some activity involving both a US and foreign city. The query is simply a graph representing these three key concepts. Part (ii) of the figure shows the specific matching subgraph in the corpus based on the evidence trail shown in part (iii). The sentences are prefixed by the documents from which they are extracted. In this case, the subgraph in part (ii) is the corpus-specific hypothesis that is generated. An analyst could then modify the subgraph to generalize the pattern, or make

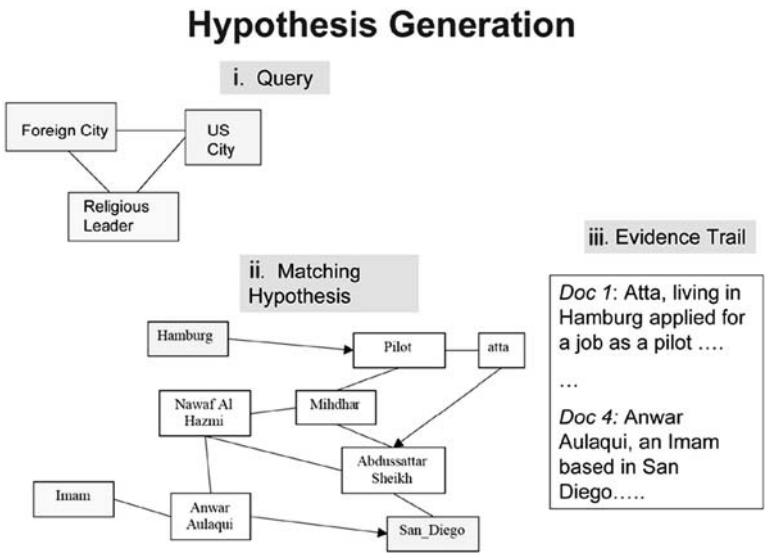



**Fig. 5.1.** Hypothesis generation: (i) query graph, (ii) matching subgraph, (iii) corresponding evidence trail.

it more specific. This could then be matched against the graph representing the data model. The arrows in the latter graph are only used to highlight the connections in the resulting evidence trail; the graph itself is undirected.

Previous work in UIR (Srihari et al. 2007) focused on concept chain queries, a special case of concept graph queries involving only two concepts. Various models were used to derive the best concept chains. The goal was to generate the best cross-document chain connecting two concepts with document sets (rather than sentences) as evidence. The work described here extends the previous work by focusing on (i) evidence trail generation, and (ii) concept graph queries. The UIR solution framework consists of several steps ranging from information extraction to graph mining to graphical models for evidence trail generation. Experimentation on the Document Understanding Conference (DUC) data set validates the approach through quantitative evaluation. The DUC data set is designed to evaluate cross-document summarization; evidence trails are a special case of this. The next section describes some differences between the two tasks.

## Criteria for ranking evidence trails

There are various criteria for ranking concept chains, including: (i) recency, (ii) most interesting, (iii) most plausible, and (iv) going through certain specified concepts. In this work, we are focused on finding chains that are *coherent* and *informative*.

Coherence is the main criteria that needs to be satisfied when generating chains. It is not sufficient for individual links between concepts to make sense: the chain as a whole must make sense also. Chains of length 5 or greater often have problems since at some point, the thread gets lost. In formulating ranking algorithms for concept chains, coherence is the primary factor to be considered. Later sections describe how techniques used to evaluate student essays for coherence are adapted to rank evidence trails. In some cases there is not a single thread connecting all concepts; the resulting evidence trails reflect this and represents disjoint sets of connections spanning all concepts.

We also require evidence trails to be informative. For example, if person A eats breakfast and person B also eats breakfast, then although eating breakfast is a valid connection, it is not of interest. On the other hand, if both of them have a liking for exotic spicy food from Southeast Asia, the connection starts to get more interesting. Developing quantitative measures for judging the informativeness of a chain is part of this effort. Techniques such as distributional similarity (Lee 1999) are useful in at least determining the concepts which are important to the domain, based on the difference between general and corpus-specific usage.

Finally, it should be noted that while the evidence trails serve as summaries of how concepts are connected across documents, this is not the same as the task of cross document summarization (Radev 2000). In later sections we use

data sets and queries designed for evaluating cross document summarization; however there is a difference in the intended use of evidence trails. They are intended to weave a thread between query concepts and place heavy emphasis on the order in which the concepts appear in the evidence trail. The resulting explanation may focus on less important themes and topics than what a traditional summary would aspire to cover. Nevertheless there is enough overlap between the goals of the two tasks to motivate us to use the same data sets.

## Related work

There are several research efforts that are related to the work on concept chain queries described here. The DARPA EELD program has resulted in text mining efforts that use more sophisticated information extraction (IE) output such as named entities, relationships and events (Weiss et al. 2004). Such systems typically use IE tools to extract salient entities and relationships; these are then input to either visualization or link analysis tools. Since IE systems have yet to achieve a level of recall for relationship and event detection that is sufficient for this type of analysis, such techniques may fail to capture significant links/paths. Wang et al. (2005) describes text mining on a corpus of govt. documents with the goal of discovering interesting patterns involving groups of entities and topics. While not focused on specific concept chain queries, it does go beyond named entities and establishes correlations across documents between entities and general concepts.

There has been work on discovering connections between concepts across documents using social network graphs, where nodes represent documents, and links represent connections (typically URL links) between documents. However much of the work on social network analysis has focused on different types of problems, such as detecting communities (Gibson et al. 1998). Faloutsos et al. (2004) is the work which is closest to the research presented here, at least in its goals. The authors model the problem of detecting associations between people as finding a connection subgraph and present a solution based on electricity analogues. The most notable difference is their reliance on URL links to establish connections between documents; our approach extracts associations based on content (textual) analysis. Second, the connection subgraph approach presents all paths together, while our approach presents the paths individually. This allows greater user input in determining the *best* paths, including recency, novelty, semantic coherence, etc. Third, the approach presented here attempts to generate an explanation of the chains, whereas the connection subgraph approach does not. Finally, the connection subgraph solution only addresses named entities whereas this approach extends to general concepts.

IR-based approaches have also been employed to solve similar problems. Srinivasan (2004) discusses a technique based on constructing semantic profiles of concepts to derive connections between concepts in biomedical

documents. This reflects an attempt to replicate the results of Swanson's (1988) pioneering efforts in text mining through purely automated techniques; the results are impressive and hence this is used as a baseline model for our own experiments described later.

## UIR framework for text mining

This section describes the UIR framework and an overview of the solution to the hypothesis generation problem. A key part of the solution is the representation framework. What is required is something that supports traditional IR models (such as the vector space model), graph mining and probabilistic graphical models. We have formulated a representation referred to as concept chain graphs (CCG).

Formally a CCG is a hypergraph $G(E, V)$ with $E$ edges and $V$ nodes representing a set of documents $D$ with the following properties:

- each node $v$ represents a term, a concept or a document
- each edge $e$ represents an association between two concepts or a membership link (e.g. link between a document and a concept, *is-a* links, or links between a concept and its contituent terms)

Nodes can have attributes such as the weight of a concept; edge weights indicate the strength of an association. The weighted graph representation enables traditional graph mining algorithms to be applied to the CCG. The CCG can also be viewed as a specialized index consisting of four layers:

(i) Document Layer. This layer contains documents and links (e.g. hyperlinks) between them.
(ii) Concepts and Associations Layer. Consists of concepts and associations coming from corpus or ontology mapping.
(iii) Instances Layer. Tracks instances of concepts and associations detected in the corpus back to documents. Also maintains instance specific information last offsets and type.
(iv) Index Terms Layer. Consists of index terms and hits.

Figure 5.2 illustrates a schematic representation of a small portion of the CCG that has been constructed from the 9/11 document collection. The figure illustrates the explicit representation of connections between concepts; the concept *bin_ladin* in document D121 is linked to concepts occurring in document D236. Edges labeled Ax represent associations between concepts. The number of hits for a concept in a document is also recorded. A typical forward index view can be obtained by traversing concepts associated with a document. A reverse index view is obtained by examining nodes and links emanating from a given concept (and what documents they are connected to). Ontological links are also illustrated; *white_house* is a type of *organization*.

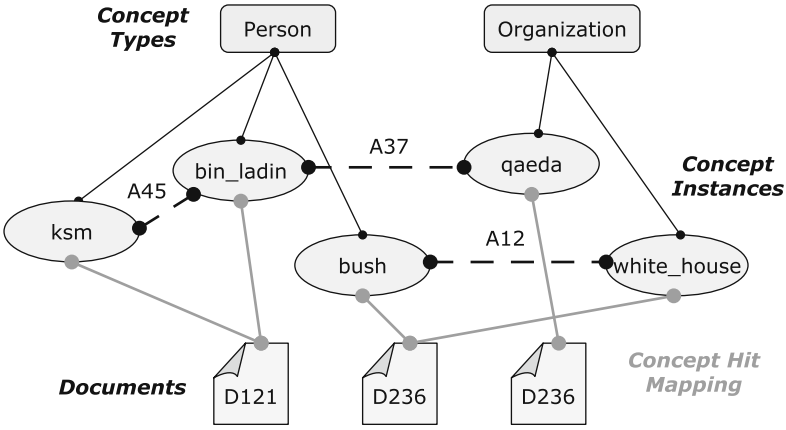



**Fig. 5.2.** Portion of the CCG.

The solution to the hypothesis generation problem can then be described as the following steps:

CCG construction: Process the corpus through an information extraction system that can extract concepts and associations; generate the CCG.

Graph Matching: Using the query concepts, generate hypothesis graph candidates. This step is based solely on weighted graph matching algorithms.

Evidence Trail Generation: Generate evidence trails corresponding to each candidate using a graphical model based on the CCG; rank evidence trails for coherence and informativeness, and return the best hypothesis.

There is an observation that should be made with respect to the above process. We have adopted a strategy whereby graph matching techniques are first used to generate candidate hypothesis matches; the candidates are then ranked based on the quality of the evidence trails that can be associate with them. Ideally, the hypothesis generation process should be driven by the evidence trail generation phase, not the other way around. It is possible to formulate a hierarchical graphical model solution to this problem based on selecting the best sequence of sentences from the corpus that connect all the concepts; this in turn would generate the best matching hypothesis (subgraph). Since the initial focus is on evidence trail generation, we have chosen an approach where that subtask can be independently evaluated and refined.

## Generating the CCG

The process of generating the CCG includes the following steps: (i) domain customization, (ii) concept extraction and filtering, (iii) mapping concepts to target ontology, and (iv) construction of the UIR index. Domain customization

consists of (i) developing a suitable ontology, and (ii) tailoring the information extraction engine to the domain in question. This includes customization of the named entity tagger. The Semantex[1] engine (Srihari et al. 2006) was used for this effort. It permits customization of the lexicons, as well as defining new entity types which were useful for this effort. We adapted Teknowledge's Terrorism Ontology[2] but modified it to suit our task. It contains 21 top-level concepts, a total of 180 concepts, and extends to a maximum depth of 5 although the typical depth is 3. Other top-level ontology nodes can be seen in Table 5.1 which is discussed further in the next section. In addition to the top-level classes shown, it also includes the standard named entity categories, *person, organization, location etc.* to handle instances of those types. To facilitate mapping of concepts into the ontology, each node in the ontology was associated with WordNet synsets.

**Table 5.1.** Portion of terrorism ontology.

| Root Concept | Examples |
|---|---|
| Feeling | *gravity, desire* |
| Document | *fbi_report, conspirator_statement* |
| Building | *national_counterterrorism_center* |
| Region | *staging_area, american_soil* |
| Person(role) | *visa_applicant,survivor* |
| Human Action | *mistake, moderation, nomination patriot_act* |

### Concept extraction and selection

Concept extraction involves running an information extraction engine, Semantex on the corpus. Semantex tags named entities, common relationships associated with person and organization, as well as providing subject-verb-object (SVO) relationships. We extract as concepts all named entities, as well as any noun phrases participating in SVO relationships. All named entity instances are retained as instances of their respective named entity concept category. Concepts that are *not* named entities undergo filtering and mapping phases. The extracted set of concepts are filtered on the basis of their global significance in the document corpus. Singleton concepts are removed. Some concept merging is performed in an attempt to consolidate aliases across documents, e.g., spelling variations for names. Table 5.1 illustrates some of the concepts that are extracted by the system; all these have survived the filtering processing and have automatically been mapped into the ontology; mapping

[1] Courtesy of Janya Inc.

[2] `http://ontology.teknowledge.com`

is discussed in the next section. Concepts in the table such as *fbi_ report, american_ soil* indicate the data-driven manner in which these concepts are extracted.

### Relationship extraction

Semantex generates a modest number of semantic relationships involving person, organization and location entities in particular such as *employed_ by, whence* etc. All these are used in our content representation. However, we recognize that this does not even begin to capture all the various relationships that are important. To increase the coverage of significant associations, we also exploit key syntactic relationships that Semantex generates. In particular, if two concepts are linked through a set of subject-verb and verb-object links, we generate an (unlabeled) association between them. We consider this to be an improvement over simply using sentence-level co-occurence to generate associations.

WordNet is used to map concepts to ontology nodes using a hierarchical traversal technique. As Table 5.1 illustrates, the technique used above sometimes produces erroneous results. For example, *patriot_ act* is mapped into the ontology node *human action* rather that into *document* which includes acts of legislation. Concepts that cannot be mapped into categories are placed in a miscellaneous category and can still participate in concept chains. Finally, the CCG is constructed which also records weights on associations, number of hits, etc. The E4Graph package is employed which enables us to persist the graph.

## Graph mining: Hypothesis candidate generation

This section discusses the graph matching techniques used to generate candidates for the hypothesis subgraph. The objective is to take as input a set of query concepts, and find the best matching subgraph that connects all of these. In cases where no direct edges are present between certain concepts, the system attempts to find the best chain of concepts connecting them. We first describe this simpler case involving only two query concepts: we refer to this as a concept chain. This is followed by a generalization to concept graph queries involving three or more concepts.

### Concept chains

Generating the best concept chain for a given concept chain query is performed using a Markov chain model. A concept can make a transition to another concept through one of the "recognized relations". If a concept $X$ is related to another concept $Y$ which has a similar context as that of $X$, then such

a relation can be coherent and meaningful. Each link can be seen as some *drift* away from the original concept. Keeping this in mind we calculate the transition probabilities of the concepts based on their contexts.

A *concept* can be formed by one or more terms. These terms define a *term vector* for that concept in the $n$ dimensional Vector Space Model (Baeza-Yates and Ribeiro-Neto 1999). A *context vector* of a concept is given by the union of term vectors of the concept and the term vectors of its related concepts; these related concepts are extracted from a domain-specific ontology. For example, concepts related to concept "wmd" would include "chemical_weapon, biological_weapon, nuclear_weapon", etc.

In the Context Model, for any concept $C_i$ and $C_j$, the transition probability is given as

$$P\left(C_i, C_j\right) = \frac{sim\left(\overline{C_i}, \overline{C_j}\right)}{\sum sim\left(\overline{C_i}, \overline{C_k}\right)} \quad \forall C_k \in \{neighbors\left(C_i\right)\} \tag{5.1}$$

where $sim\left(\overline{C_i}, \overline{C_j}\right)$ is the cosine distance between the context vectors of concept $C_i$ and concept $C_j$. It is important to note that even though the similarity measures are symmetric in that $sim(\overline{C_i}, \overline{C_j}) = sim(\overline{C_j}, \overline{C_i})$, the transition probabilities are not symmetric i.e. $P(C_i, C_j) \neq P(C_j, C_i)$. This asymmetricity arises from the fact that each concept has a different neighborhood. It is in a way interesting to have the forward probabilities differ from the backward probabilities in that it gives a possibility to get a different best Markov Sequence from $C_i$ to $C_j$ than from $C_j$ to $C_i$.

### Concept graph

As an extension to the Concept Chain model described previously, we pursued techniques for generating a graph of concepts rather than a linear chain in response to the user query. The primary motivation behind this approach was to ease the restriction of just two input query concepts; the concept graph model can accept any number of inputs and produces a connected graph of concept nodes and associations as output. Unlike the Markov Chain Model used for the Concept Chain generation model, we approached this problem as a pure graph matching problem, using the entire CCG concept neighborhood as the parent graph and the query concepts as key vertices within the graph. Although superficially this problem resembles the Minimum Spanning Tree (MST) problem, closer analysis reveals that it is in fact more similar to the Steiner Tree problem.

Formally, given a weighted graph *G(V,E,w)* and a vertices subset $S$; we call a Steiner Tree an acyclic sub graph of $G$ that connects all vertices in $S$. A Minimum Steiner Tree is a minimum edge-weight instance of a Steiner Tree for a given graph. The vertex subset $S$ are also called Steiner Points of the sub-graph.

We use Mehlhorn's Algorithm (Mehlhorn 1998) to build the graph:

(i) Compute the entire distance network $N_d$ between all the vertices in $S$ [Dijkstra's Algorithm]
(ii) Build a Minimum Spanning Tree $T_d$ in $N_d$ [Kruskal's Algorithm]
(iii) Transform $T_d$ into a subnetwork $N$ of $T_d$ by replacing every edge of $T_d$ with its corresponding shortest path
(iv) Build a Minimum Spanning Tree $T$ for the subnetwork $N$ of $T_d$ [Kruskal's Algorithm]
(v) Transform $T$ into a Steiner Tree $S_k$ by successively deleting non-terminal leaves

The bottleneck in this algorithm is the first step of computing the entire distance network, especially in the case of larger graphs with larger queries. This is done by calling Dijkstra's Algorithm once for every terminal. However, once the complete distance network is known, the magnitude of the problem is significantly reduced.

Two small modifications were made to this algorithm to account for the sparseness of the data set we used in our experiments: (i) While computing the distance network, to enable the algorithm to favor smaller length *chains*, we introduced a dampening factor of 0.2 for every edge encountered along the path. This ensured that shorter chains were favored over longer ones, but at the same time a long chain with a sufficiently high score would still be chosen over *weaker* short chains. (ii) We put a cap on the maximum depth of the shortest path algorithm to search for nodes at a maximum distance of 5 away from the source node.

Additionally, we experimented with improving candidate subgraph generation by introducing the notion of *concept importance* to the algorithm. The motivation for using concept importance includes (i) eliminating (or reducing) subgraphs containing several trivial concepts such as "part", and (ii) the desire to dampen the effect of *sink* nodes which ultimately led to a poor summary being generated. Sink nodes (e.g. "USA") are characterized by a high number of associations, many of which do not carry much semantic importance. Faloutsos et al. (2004) refer to this as the pizza delivery man phenomenon: since he delivers pizza to many people in an area, there is a trivial connection between many people through him. As an initial step, we adopted purely graph-based techniques to assign the concept weights; future plans call for measuring the distributional similarity of a concept in a domain specific language model as compared to a generic language model.

Figure 5.3(a) shows a small subset of the entire CCG graph; Fig. 5.3(b) shows a sample Concept Graph for query concepts *atta*, *ksm* and *hazmi*.

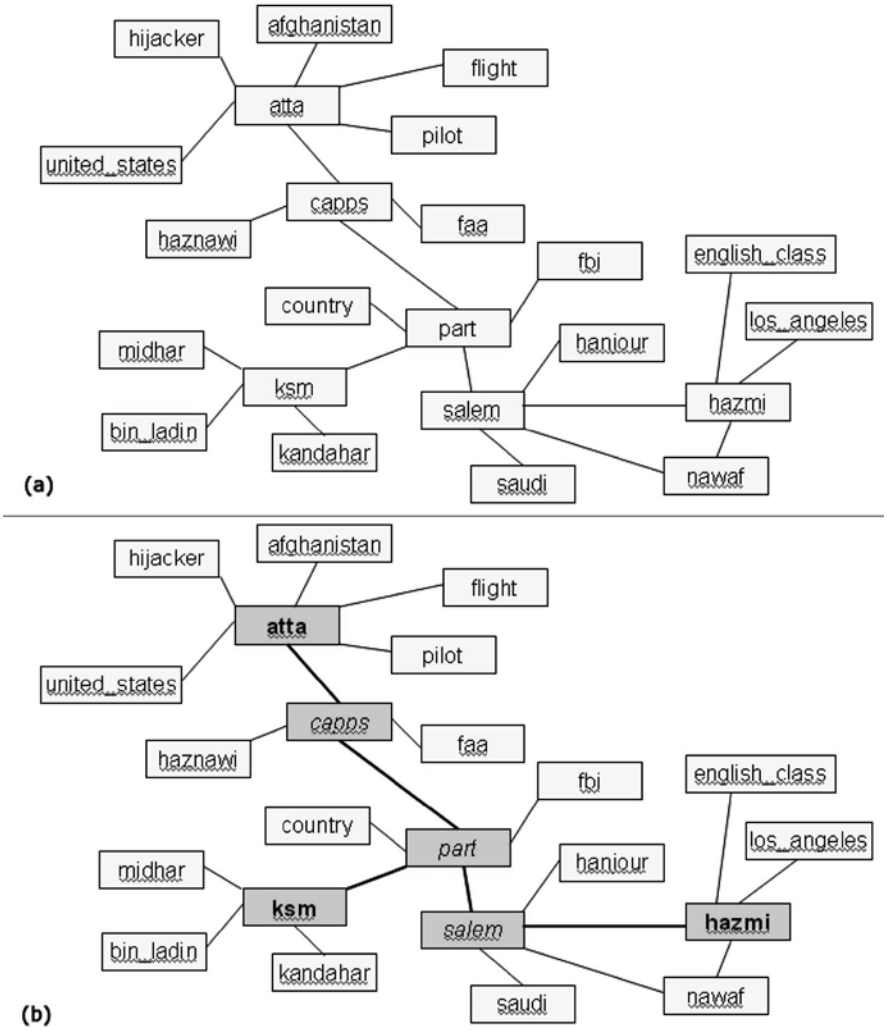



**Fig. 5.3.** Sample concept graph produced from CCG.

## Evidence trail generation

This section describes a Hidden Markov model to generate the best evidence trail given a set of concept chains. We suppose all the documents within the corpus belong to one topic. We follow Harabagiu and Lacatusu (2005) to represent a topic as a structure of themes. A theme is defined as a cluster of sentences which convey the same semantic information. Themes represent events or facts that are repeated throughout the document collection. A content model (Barzilay and Lee 2004) is built to capture the relationships among

these themes. Evidence trails are generated by Viterbi decoding on the content model. Desirable evidence trails need to be both informative and coherent. Informativeness is incorporated in concept chains induced from the Concept Chain Graph. Coherence is obtained by the drift of topic themes captured in the Content Model.

**Content model construction**

A Content model is essentially a Hidden Markov Model in which states correspond to themes and state transitions indicate the probabilities changing from one theme to another by examining the theme ordering in the topic. It is built on the assumption that all texts describing a given topic are generated by a single content model. We adapt the Content Model described above to our problem.

Given a set of sentence clusters $c_1, c_2, ...c_m$, we construct a Content Model whose states $s_1, s_2, ...s_m$ corresponds to these clusters. Here onwards, we do not differentiate between cluster $c_i$ and state $s_i$. Each state is associated with a state-specific language model. For state $s_i$, a state-specific bigram language model is built as:

$$P_{lm_i}(w_2|w_1) = \frac{f_{c_i}(w_1 w_2) + \delta_1}{f_{c_i}(w_1) + \delta_1 |V|} \tag{5.2}$$

where $f_{c_i}(x)$ is the frequency with which word sequence $x$ occurs in cluster $c_i$; $V$ is the vocabulary; $\delta_1$ is a smoothing parameter.

The sentence emission probabilities are determined by the state-specific language model and association overlap.

$$P(x|c_i) = \lambda_1 * P_{lm_i}(x) + (1 - \lambda_1) * \frac{\sum_{a_k \in x} count(a_k, c_i)}{\sum_k count(a_k, c_i)} \tag{5.3}$$

where $count(a_k, c_i)$ is the number of times that association $a_k$ is observed in cluster $c_i$; $P_{lm_i}(x)$ is the probability of sentence x being generated by the language model associated with cluster $c_i$.

The transition probability measures the likelihood of state $s_i$ preceding state $s_j$, which is captured by considering sentence ordering in the original documents, and the overlap of concepts between them. The transition probability is computed as follows:

$$P(c_j|c_i) = \lambda_2 * \frac{D(c_i, c_j) + \delta_2}{D(c_i) + \delta_2 m} + (1 - \lambda_2) * \frac{\sum_{cpt_k \in c_j} count(cpt_k, c_i)}{\sum_k count(cpt_k, c_i)} \tag{5.4}$$

where $m$ is the number of states; $D(c_i, c_j)$ is the number of documents in which a sentence from $c_i$ immediately precedes a sentence from $c_j$; $count(cpt_k, c_i)$ is the number of times that association $cpt_k$ is observed in cluster $c_i$; $\delta_2$ is a smoothing parameter.

The sentence clusters are initialized by complete-link analysis. We follow Barzilay and Lee (2004) who use an EM-like Viterbi re-estimation procedure to build the model: re-cluster sentences by placing them in the state which is most likely to have generated it after the Viterbi decoding. The new clusters are then used as input to estimate HMM parameters. Repeat this cluster/estimation procedure until the clusters stabilize.

**Evidence trail generation**

The generation of evidence trails is accomplished by (1) traversing through the hypothesis graph returned in the previous section to enumerate all possible chains between the input concepts, (2) treating these chains as input to the content model to decode sequences of themes, (3) organizing sentences from decoded themes to form evidence trails.

This process takes as input a concept chain which can be also viewed as a sequence of associations. Observe that here the emission of the content model changes from sentences into associations: the content model needs to be modified accordingly. The state transition probabilities stay the same. The emission probability is now defined as:

$$P(a_j|c_i) = \frac{count(a_j, c_i)}{\sum_k count(a_k, c_i)} \tag{5.5}$$

where $count(a_j, c_i)$ is the number of times that association $a_j$ is oberserved in cluster $c_i$. This emission probability is not smoothed because we are not interested in cluster $c_i$ if it doesn't include the current association $a_j$.

Viterbi decoding of the input chain results in a sequence of sentence clusters which has the highest possibility of generating the observed concept chain. Since one association can appear in several sentences within a cluster, there could be several candidate sentences for each association in the chain. Since each cluster is a topic theme conveying similar semantic meaning, it does not impact the informativeness of the evidence trail if different sentences are chosen. It is readability that is of our concern now, which is taken care of by the ranking metric later. So we treat all sentences which contain the corresponding association in the decoded sentence cluster as candidates, and generate all possible evidence trails out of them. The evidence trails for concept chains enumerated from one hypothesis graph are combined together with duplicate associations and their corresponding sentences removed to form the final result.

**Ranking evidence trails**

Evidence trails are ranked based on two main criteria, coherence and informativeness. A key consideration is the lack of a gold standard with which to compare the evidence trail: a standalone technique for ranking is required. There

have been previous efforts to model coherence (Barzilay and Lapata 2005). For the current research we are motivated by efforts to measure coherence in essays written by grade school students. The Coh-Metrix tool (Graesser et al. 2004) in particular is relevant to the current work. Coh-Metrix is a computational tool that produces indices of the linguistic and discourse representations of a text. It takes into account many other dimensions of good writing such as readability, concreteness, sophistication of writing etc. Since we are not judging the quality of sentences, only the selection and ordering, many of these factors are not relevant to our work. Currently, the evidence trail ranking module uses *latent semantic analysis* (Landauer and Dumais 1997) to compute a score. LSA is an algebraic technique for analyzing relationships between a set of documents and the terms they contain resulting in a reduced concept space that best characterizes the set of documents and terms. Sentences are represented as vectors in this space; similarity between sentences can be computed using cosine distance. We use a sentence to sentence comparison technique whereby a summary of $n$ sentences results in $n$-1 cosine comparisons between the sentences. The mean of the individual comparisons is used to judge the goodness of an evidence trail. The semantic space is based on college level general English.[3] We are in the process of retraining the LSA model on a comprehensive set of DUC documents and using this reduced space matrix for the summary ranking system. While this does not capture all the criteria we desire, it is a good starting point. Experiments have shown that permuting the sentences in a manually judged good evidence trail significantly alters the LSA score.

## Experiments

This section describes the experiments that were conducted. The Document Understanding Conference (DUC) 2005 data set was used. The data set is organized into a set of topics where each topic has a set of 25–50 documents relevant to it. The task: given a user profile, a DUC topic, and a cluster of documents relevant to the DUC topic, create from the documents a brief, well-organized, fluent summary which answers the need for information expressed in the topic, at the level of granularity specified in the user profile. The summary should include (in some form or other) all the information in the documents that contributes to meeting the information need. Some generalization may be required to fit everything in. The summary can be no longer than 250 words.[4]

The DUC data set was selected since it is one of the few data sets involving a cross document analysis task. It has been explained earlier that the

[3] The tools available at `http://lsa.colorado.edu/` were used in this module.

[4] More information can be found at `http://www-nlpir.nist.gov/projects/duc/duc2005/tasks.html`

**Table 5.2.** DUC query set.

| | Title |
|---|---|
| 1 | Argentine British relations post Falkland War |
| 2 | Amazon Rainforest Problems |
| 3 | New Successful Applications of Robot Technology |
| 4 | Tourism in Great Britain |
| 5 | Saving Tourists and Tourism |
| 6 | Welsh devolution and British Parliament |

evidence trail, though a form of a cross-document summary is not designed to maximize the criteria being evaluated by the DUC task. However it comes tantalizingly close, and hence we use this data set for formal, quantitative evaluation. Table 5.2 illustrates the six queries that were chosen for this experiment; only the titles, not the full narrative description of each query is shown. These queries were selected since they call for question-answering techniques, and thus lend themselves well to concept graph queries where some chain of thought is required.

### Query processing

The following steps are used to process each query.

- Each query narrative is processed by the Semantex engine, and the significant concepts are extracted.
- For a second configuration, query expansion was used to select similar concepts. For each query concept, the most *similar* concepts were determined based on semantic similarity. Lin (1997) discusses semantic similarity based on grammatical dependency relationships as well as proximity. This was used[5] to derive semantically similar words for concepts that are not named entities. For named entities, string heuristics along with ontologies were used to find "similar" concepts; for example, the query concept *Argentina* resulted in additional concepts such as *Argentinian, Buenos Aires* being added to the query. No more than 25 concepts were considered in a query.
- The set of concepts was used as a concept graph query by the hypothesis generation module described previously. At most 5 matching hypotheses candidates were selected for evaluation.
- Evidence trails were generated for each of the hypothesis candidates. Due to the current limitation in the evidence chain technique, each candidate hypothesis graph was traversed and the set of concept chains enumerated.

[5] `http://www.cs.ualberta.ca/~lindek/demos.htm`

This resulted in multiple chains per graph; the evidence trail module selectively merges the individual trails to produce a single summary.

- The set of evidence trails (corresponding to each candidate hypothesis graph) were ranked using the LSA-based ranking module resulting in the best evidence trail overall, and hence, the best hypothesis graph. This top-ranked evidence trail was used in the DUC summary evaluation described below.

**Evaluation**

The highest ranking evidence trails produced by the system is evaluated against human generated summaries provided by DUC. Several such summaries are provided; we use the average of these scores. A scorer, which calculates a distance metric based on a sliding window of $n$-grams is used to judge the system produced summary against the gold standard. The lower the score, the more similar to the desired summary.

**Baseline system**

To facilitate quantitative evaluation of our technique, we devised a naiive baseline algorithm to produce a cross document summary. The algorithm works in two distinct steps: (i) Using a greedy search technique, we determine a maximally weighted set of associations such that every query term is connected by at least one association. If no direct association exists between any two query concepts, we try to find the best chain of maximum length 2 between the two; if this does not exist, we ignore the concept pair for our summary. (ii) For the best set of associations extracted in the previous step, we select the 1st sentence in which that association occurs in and add it to the summary (after eliminating duplicate sentences). This is a simple technique for generating a cross-document summary based on sentence selection.

Tables 5.3 and 5.4 illustrate a complete example, including the original query and the top-ranked evidence trail (using query expansion) from the algorithm described above. This was judged the best of 5 hypothesis graph candidates. The three concept chains (enumerated from the hypothesis graph candidate) are also shown. Each sentence is prefixed with the document number that it emanated from: 6 unique documents are used to generate this trail. It is also interesting to see the quality of evidence trails produced by the different methods on this query. The table shows the output of the Baseline algorithm on this query as well as one of the Reference (gold standard) summaries. The baseline summary contains additional sentences at the end that were omitted for space reasons. Although both summaries cover similar topics, the one based on the Hypothesis Generation and Evidence Trail ranking algorithm appears to be more succinct and meanders less. For example, in the baseline summary, sentence D26 is out of place. Finally, both system-generated evidence trails seem to be covering the same themes that

**Table 5.3.** Sample query and resulting evidence trail.

| Number | d324e |
|---|---|
| Title | Argentine British relations post Falkland War |
| Narrative | How have relations between Argentina and Great Britain developed since the 1982 war over the Falkland Islands? Have diplomatic, economic, and military relations been restored? Do differences remain over the status of the Falkland Islands? |
| Chains | (i) C343:argentine::C270:carlos-menem::C269:foreign-secretary-douglas-hurd::C182:argentina::C233:falkland<br>(ii) C320:relation::C188:britain::C182:argentina::C233:falkland<br>(iii) C277:economy::C381:cavallo::C202:douglas-hurd::C233:falkland |
| HypothGen | **D13**: ARGENTINE President Carlos Menem has ordered an 'immediate' investigation into war crimes allegedly committed by British troops during the 1982 Falklands War. **D6**: UK foreign secretary Douglas Hurd will meet President Carlos Menem in Argentina next week, the first senior British official to visit since the Falklands war 10 years ago, Reuter reports. **D9**: Provided the Argentine decision has a commercial, rather than predatory, motivation, the British government is virtually powerless to influence the move. **D10**: British Gas said last year it had discussed joint offshore exploration close to the disputed waters with YPF, Argentina's state - owned oil company. **D17**:Relations between the two countries were soured again last May when Britain decided to extend territorial waters to 200 miles around South Georgia and the South Sandwich Islands, neighbours of the Falklands, and to introduce a fishing licence system there to halt overfishing by third countries. **D14**:Last month, Argentina suffered another diplomatic slap in the face when Britain announced it would extend territorial waters to 200 miles around the remote and uninhabited South Georgia and South Sandwich island groups, which Argentina also claims. **D16**:As well as meeting UK government officials, Mr Cavallo will give a lecture at the London School of Economics and a speech at the Confederation of British Industry. **D16**: Britain has shown growing interest in Latin America as economic reforms across the continent create new trade and investment opportunities. |

are mentioned in the human generated summary. This includes mention of diplomatic visits, the oil and gas industry, tensions related to fishing, as well as the specific mention of President Carlos Menem. As expected, our summaries are richer in specifics, such as people and organization names since the sentences are selected from the corpus.

**Table 5.4.** Sample query and resulting evidence trail.

| | |
|---|---|
| Baseline | **D21**: A year ago, Mr Douglas Hurd, foreign secretary, became the first UK cabinet minister to visit Argentina since the 1982 Falkland islands conflict. **D16**: Mr Cavallo, the first Argentine cabinet minister to make an official visit to Britain since the 1982 Falklands conflict, will meet Mr Kenneth Clarke, the chancellor of the exchequer, Mr Douglas Hurd, the foreign secretary, and Mr Eddie George, governor of the Bank of England. **D16**: Mr Cavallo became economy minister in 1991 and rapidly transformed Argentina into one of the world's fastest-growing economies. **D9**: WHEN Foreign Secretary Douglas Hurd arrives in Buenos Aires this morning, the first UK cabinet minister to visit Argentina since the 1982 Falkland Islands conflict, he will fly into a barrage of inflated expectations. **D20**: BRITAIN AND Argentina have reached a new agreement on sharing fish resources in the South Atlantic and waters surrounding the Falkland islands after two days of talks in Buenos Aires. **D26**: London and Buenos Aires broke diplomatic relations during the conflict. **D14**: Eventually, London will have sufficient confidence in Argentina to consider transferring sovereignty. **D6**: UK foreign secretary Douglas Hurd will meet President Carlos Menem in Argentina next week, the first senior British official to visit since the Falklands war 10 years ago, Reuter reports. |
| Reference | Argentina and Great Britain engaged in a war in 1982 over the Falkland Islands, islands in the South Atlantic claimed by Argentina but occupied by Britain. After the war, Britain imposed an arms embargo and Argentina imposed restrictions on imports from Britain. Relations between Argentina and Britain eventually improved: in 1989, Argentina lifted restrictions on imports from Britain, and they cooperated during 1991 Gulf War. Full diplomatic relations between the two countries resumed in February 1990. Argentina made discreet attempts to resume military ties, but was rebuffed and the arms embargo continued. Carlos Memen assumed the office of President of Argentina in 1991 and adopted free-market and pro-western policies which made trade with Argentina and investments there more attractive. Trade between Argentina and Britain then began to recover. Argentina sought help from Britain on its privatization program and encouraged British investment. Britain reduced the cost of insurance protecting against possible losses resulting from investments in Argentina. In 1993, for the first time since the war, high-level ministers from both countries exchanged visits. A British consortium bought Argentina's largest gas distribution company in a privatization auction. Both countries held talks regarding joint oil and gas projects in Falkland waters. Argentina also hoped to attract British investment in their privatized nuclear industry. Two continuing sources of tension between Argentina and Britain have been Argentina's Falkland sovereignty claims and Britain's efforts to protect Falkland fishing revenues from illex squid, also competed for by Argentine and Asian fisherman and thought to be over-exploited. |

## Results

Table 5.5 illustrates the results obtained for six of the DUC queries. The score is based on the *n*-gram distance metric discussed in the previous section. Four systems are compared: the baseline system with and without query expansion, and the Hypothesis Generation and Evidence Trail (HypoGen) solution described above, also with and without query expansion. The score reflects the mean distance between the system-generated summary and each of the human generated summaries. Since the value of concept weighting in hypothesis graph generation is still inconclusive, it was not used in generating these results.

In all cases, the HypoGen technique scored better than the baseline. Furthermore, HypoGen with query expansion consistently produced better scores. While the results between systems are comparable in most queries, in a few cases, such as queries 3, 4, and 6, the baseline with query expansion fared much worse. Query expansion often introduces noise and a simple technique such as the baseline can easily be led astray. It is interesting that the HypoGen technique with query expansion still performed the best in this query, and was not adversely affected by query expansion.

It should be noted that the evidence trail generation technique was designed for a slightly different purpose, namely finding the best connections between concepts. If the goal were to produce a better cross-document summary, a post-processing module could be developed which would produce higher scoring summaries. Currently, evidence trails are being cut off to keep the length manageable; these could be extended. Temporal issues are not being taken into account; it is possible to order sentences based on this as well as coherence. High-scoring DUC summaries tend to choose sentences from the initial portions of documents; we are not taking this into account. Finally, the queries themselves could be processed in a more intelligent manner, with the goal of better summaries. Nevertheless, the technique used for hypothesis generation has resulted in a viable method for generating query-driven cross-document summaries.

**Table 5.5.** Evaluation results based on *N*-gram distance metric.

| | Q1 | Q2 | Q3 | Q4 | Q5 | Q6 |
|---|---|---|---|---|---|---|
| Base | 0.0493 | 0.0754 | 0.0918 | 0.1039 | 0.0899 | 0.1142 |
| Base-Q | 0.0483 | 0.0825 | 0.2525 | 0.2557 | 0.0845 | 0.2702 |
| HypG | 0.0397 | 0.0535 | 0.0517 | 0.0639 | 0.0590 | 0.0439 |
| HypG-Q | 0.0392 | 0.0427 | 0.0453 | 0.0535 | 0.0413 | 0.0419 |

## Summary

A new framework for generating corpus-specific hypotheses graphs has been described. This approach has the promise of reducing the effort on the part of analysts in constructing domain models that can be matched against data collections to look for scenarios of interest. The highlight is the generation of evidence trails, cross-document summaries that explain how the query concepts are connected. Results from experimentation show that this can also be viewed as a general technique for cross-document summarization. Ongoing work includes: (i) fine-tuning graph matching algorithms to account for the importance of concepts, (ii) generating evidence trails directly from the hypothesis graph candidates, and (iii) improved techniques for ranking evidence trails.

## Acknowledgments

This work is sponsored by NSF grant IIS-0325404 and FAA grant 032-G-009.

## References

Baeza-Yates, R., and B. Ribeiro-Neto. 1999. *Modern Information Retrieval*. Addison Wesley, New York.

Barzilay, R., and M. Lapata. 2005. Modeling local coherence: An entity-based approach. In *Proceedings of the 43rd Annual Meeting of the ACL(ACL'05)*, pp. 141–148, Ann Arbor, Michigan, June 2005. Association for Computational Linguistics.

Barzilay, R., and L. Lee. 2004. Catching the drift: Probabilistic content models, with applications to generation and summarization. In D. M. Susan Dumais and S. Roukos, (Eds), *HLT-NAACL 2004: Main Proceedings*, pp. 113–120, Boston, Massachusetts. Association for Computational Linguistics.

Faloutsos, C., K. S. McCurley, and A. Tomkins. 2004. Fast discovery of connection subgraphs. In *Proceedings of the Tenth ACM SIGKDD International Conference KDD-2004*, pp. 118–127, Seattle, Washington 2004. ACM Press, New York.

Gibson, D., J. Kleinberg, and P. Raghavan. 1998. Inferring web communities from link topology. In *Proceedings of the Ninth ACM Conference on Hypertext and Hypermedia*, pp. 225–234.

Graesser, A. C., D. S. McNamara, M. M. Louwerse, and Z. Cai. 2004. Coh-metrix: Analysis of text on cohesion and language. *Behavior Research Methods, Instruments and Computers*, 36:193–202.

Harabagiu, S. and F. Lacatusu. 2005. Topic themes for multi-document summarization. In *SIGIR '05: Proceedings of the 28th Annual International ACM SIGIR Conference*, pp. 202–209, New York. ACM Press.

Landauer, T. K. and S. T. Dumais. 1997. A solution to plato's problem: The latent semantic analysis theory of the acquisition, induction, and representation of knowledge. *Psychological Review*, 104:211–240.

Lee, L. 1999. Measures of distributional similarity. In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics*, pp. 25–32.

Lin, D. 1997. Using syntactic dependency as local context to resolve word sense ambiguity. In *Proceedings of the 35th Annual ACL Conference*, pp. 64–71, Morristown, NJ. Association for Computational Linguistics.

Mehlhorn, K. 1988. A faster approximation algorithm for the steiner problem in graphs. *Information Processing Letters*, 27(2):125–128.

Radev, D., H. Jing, and M. Budzikowska. 2000. Centroid-based summarization of multiple documents: Sentence extraction, utility-based evaluation, and user studies. In *Proceedings of the ANLP/NAACL 2000 Workshop on Automatic Summarization*.

Srihari, R. K., W. Li, T. Cornell, and C. Niu. 2006. Infoxtract: A customizable intermediate level information extraction engine. *Natural Language Engineering*, 12(4):1–37.

Srihari, R. K., L. Xu, and T. Saxena. 2007. Use of ranked cross document evidence trails for hypothesis generation. In *Proceedings of the 13th International Conference on Knowledge Discovery and Data Mining (KDD 2007)*, San Jose, CA, August 2007, pp. 677–686.

Srinivasan, P. 2004. Text mining: Generating hypotheses from medline. *JASIST*, 55:396–413.

Swanson, D. R. 1988. Migraine and magnesium: Eleven neglected connections. *Perspectives in Biology and Medicine*, 31(4):552–557.

Wang, X., N. Mohanty, and A. McCallum. 2005. Group and topic discovery from relations and text. In *Proceedings of Neural Information Processing Systems (NIPS)*.

Weiss, S., N. Indurkhya, and D. T. Zhang. 2004. *Text Mining: Predictive Methods for Analyzing Unstructured Information*. Springer, New York.

# 6

# Identification of *Sensitive Unclassified* Information

Kazem Taghva

**Summary.** *Sensitive Unclassified* information is defined as any unclassified information that may cause adverse consequences against the government facilities. In this chapter, we explore the use of categorization techniques and information extraction to discover this kind of information in scanned documents.

We show here that the combined use of a K-Dependence Bayesian categorization engine and a semi-automated review application reduce by nearly 95% the number of man hours required to redact sensitive unclassified information. We also discuss and provide statistics on how OCR errors can affect the information extraction tasks.

## Introduction

The Freedom of Information Act (US Government 2004) is based on the premise that government records belong to the public and should be disclosed unless they fall within one of the excludable categories. For example, Exemption 6 covers private information. After the events of September 11th, 2001, federal agencies ordered a re-examination of existing policies for the dissemination of information normally provided to the public. *Sensitive Unclassified* (SU) information has been defined as any unclassified information not otherwise protected from disclosure by law or regulation that may cause an adverse consequence against government facilities or infrastructure. This creates a dilemma when records held by government agencies are required by law to be released in the public domain, but these same records may contain SU.

With the explosion of electronic media, in particular the Internet, the problem of inadvertent release of SU is a major security concern. But with limited budgets, few resources, and a lack of necessary technology, the problem of identifying and if necessary, redacting SU becomes unwieldy. This is especially true when paper records are ordered for release. The United States Department of Energy's (DOE) Licensing Support Network (LSN) (US DOE 2001) is an example of a collection that require redaction of SU. The LSN is a large document repository of over forty million pages that will provide

S. Argamon, N. Howard (eds.), *Computational Methods for Counterterrorism*, DOI 10.1007/978-3-642-01141-2_6,

information to the proceedings for licensing Yucca Mountain's nuclear waste repository. The information contained in the LSN is obviously of public interest but could easily contain SU information since it consists of a variety of documents about nuclear waste. To review and redact such a collection would have taken thousands of man hours.

The technology and sciences of Information Retrieval (IR), Text Categorization (TC) and Information Extraction (IE) can play an important role in identification of relevant information for any information related task. The same technology can also play an important role in understanding data, relating concepts, and inferring new facts from the data. Although these information access technologies are not completely mature, they can be used to solve some of the problems associated with information dissemination. In what follows, we will show how we use such information access methods to partially overcome the problem of identifying SU information.

## Problem description

Immediately after 9/11, the intelligence community started to examine written information and data via what is known as the "three-legged stool" methodology. Any stool, in order to be balanced, requires at least three legs of equal length to keep the seat horizontal, and this premise can be applied to the analysis of documents for SU. For example, suppose a document analyst (known as a "derivative classifier" in the intelligence community) is looking for clues that could guide an adversary to plan an act of terrorism. In order to plan such an act, a target must be identified (e.g. a dam), its location must be found (e.g. via a map), and its accessibility must be revealed (e.g. bypassing physical security barriers). In the three-legged stool analogy, the target, location, and accessibility are the three legs. Information about these three is exactly what a human classifier and an adversary look for. A document thus becomes SU if it contains enough information in the appropriate proportions to keep the seat of the three-legged stool horizontal. Of course, evaluating if the stool is balanced requires years of experience in intelligence background and education.

After much experience in reviewing SU materials, the intelligence community has come to a better understanding of the security principles involved. In particular, the three-legged stool approach has been generalized to define hundreds of scenarios that an adversary could possibly come up with. For example, in the basic three-legged stool approach, each document is judged solely based on its content without taking into account key relationships with other information that may be available in other forms or other documents.

As mentioned above, the LSN with over forty million documents was to become a public database. The first approach to identification of SU material was to use an IR approach. In this setting, classifiers will search for sensitive

words, phrases, or patterns to find documents of interest. Although a labor-intensive approach, it was successful in identifying a set of SU documents. This approach thus led to building of an initial collection which could be used to devise an automated process.

The basic idea behind automating the process was to treat the SU problem as a binary text categorization problem. A successful solution would sort documents into SU and NOT-SU bins. Based on sampling of the data, it was clear that less than one-tenth of one percent of documents were SU. This means that if a categorizer put all the documents into the NOT-SU bin, its accuracy rate would be 99.9%! This fact, combined with a high penalty associated with the release of an SU information, made us rethink this binary categorization approach.

In the three-legged stool approach, a human classifier focuses on sections of document that are about target, location, or accessability. He then analyzes the found sections with respect to the rest of the document. If he sees enough information to build a harmful scenario, then he redacts part of the document. In our approach, we seek to assist the human classifier by providing him the appropriate sections of the documents that may be about target, location, or accessibility (termed *POSSIBLY-SU*). In this way, the human analyst is still responsible for identifying harmful scenarios, but needs to examine substantially fewer documents.

This solution was implemented by the use of a binary classifier to identify POSSIBLY-SU materials or, equivalently, identify the NOT-SU bin with a high degree of accuracy. In this approach, we can use probabilities to calculate odds for a document being NOT-SU. This, combined with a detailed review process turned out to be a practical solution to the problem. The next two sections will provide a detailed look at our solution.

## System architecture

As mentioned above, since some LSN documents may contain sensitive information, SU review is required prior to LSN submission. Because of the large number of documents DOE expects to submit, manual SU review is prohibitive. The Information Science Research Institute (ISRI) has developed a methodology and a set of software tools to generate a specialized classification system. The system is designed to separate NOT-SU information from SU information with a very high level of accuracy. The fundamental rationale is to identify NOT-SU information with 100% precision; the remaining "possibly sensitive" document pages are marked for subsequent review. Figure 6.1 shows the overall flow of how an incoming document gets classified.

### Control flow

A document in the `Document Directory` consists of page images and metadata (e.g. title, author). The system uses an Optical Character Recognition

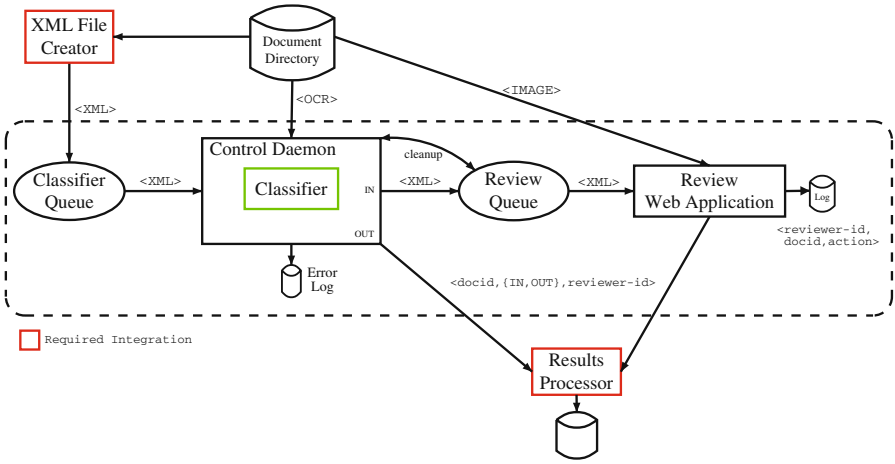



**Fig. 6.1.** Architecture of our semi-automated SU classification system.

(OCR) device to produce an ASCII file that contains the text of the documents in addition to font and geometry information for each word. The `XML File Creator` builds an XML file describing the metadata for the document which is then passed to the `Clasifier Queue`.

Features are extracted from the incoming document (both the text and the XML description) and processed by the `classifier`. The `classifier` assigns probabilities to this document's features and computes its overall probability of being POSSIBLY-SU. If indeed a document is found to be POSSIBLY-SU, the offending pages are tagged and the document is placed in the `review queue`; NOT-POSSIBLY-SU documents are also appropriately tagged as "out" or not requiring review and forwarded to the `results processor`.

The `review web application` is the interactive interface that aids the manual reviewer. Recall that the classifier has tagged certain pages in a document as POSSIBLY-SU. The interactive interface allows reviewers to quickly "zero in" on sensitive passages within these pages. Although the complete document is available for inspection, this tool simplifies SU identification by highlighting the most sensitive passages. The highlighting is done on the document images using the geometry information provided by OCR.

### Text classification

The classifier can be viewed as the "brains" of the system. This subsystem is a *K-Dependence Bayesian system* that was built based on features extracted from training materials, as will be described in detail below. Features are various objects determined useful as discriminants among the pre-defined categories. For textual data like documents, words, phrases, and textual patterns are almost always used as discriminating "features." Other features, though,

can also play a role for a particular classification task. For example, logical rules can be applied as "features."

There are two main processes during the training phase: *feature selection* and *network generation*. Features are selected through expert manual analysis and automatic evaluation of pre-selected training data. Network generation is an autmoated learning process that discovers dependencies among these features and calculates corresponding probabilities.

One point worth mentioning is that each page was considered a document for the purpose of categorization. In this subsystem, the notion of identifying "document pages" is important because our analysis indicates that there could be just a single passage in a very large document that contains SU information.

**Document representation**

Following McCallum and Nigam (1998), assume we have a feature set or vocabulary $V = (X_1, X_2, \ldots, X_{|V|})$ for our collection, then a document $d_i$ can be represented by a vector:

$$d_i = (N_{i1}, N_{i2}, \ldots, N_{i|V|}) \tag{6.1}$$

where $N_{ij}$ is the number of occurrences of the feature $X_j$ in the document $d_i$.

Many groups of features were used to represent each document. Here are three prominent groups:

(i) **Words**. A set of words is chosen according to their information content. Training documents are first indexed, then the number of index terms reduced by using dimensionality reduction techniques. Lastly the *Mutual Information* ($MI$) of each remaining term $x$ is computed to identify important terms that discriminate between categories.

(ii) **Manually selected phrases**. These are phrases that need to be selected by subject matter experts. In the case of junk email, these could be phrases such as "Free Money" or "Free Pics" and perhaps can be chosen by anyone who has received junk email. In a formal setting such as government agencies, this may require completely different criteria. Examples of these phrases are `alarm device for high radiation area`, `drip shield`, and `fuel cladding`.

(iii) **Selected rules**. The intention of rule selection is to include a representation for meta-data in the classifiers. Most of the commercially available classifiers are word based and ignore rules governing meta-data. For example, a list of authors who are known to write SU documents.

We also used other types of manually-defined features such as `sensitive patterns`. These represent features for specific materials and the like.[1] In

[1] Due to security concerns, we cannot release full details about these features at this time.

general, these patterns were manually selected and formulated as regular expressions. The simplest form of a pattern is the co-occurrence of certain words and numbers within a specific distance.

## K-dependence categorization

Most extant statistical text categorization techniques are based on the probabilistic approach introduced by Maron and others (Maron and Kuhns 1960; Maron 1967; Lewis 1991). Following McCallum and Nigam (1998), we assume we have a set of $C = \{C_1, C_2, \ldots, C_{|C|}\}$ classes that we want to assign to our document collection. One basic assumption is that each document falls into exactly one category (i.e. exhaustive and incompatible).

In this framework, we are interested in finding $P(C_j|d_i)$, or the conditional probability that a document belongs to category $C_j$. Using Bayes' theorem, we can calculate this probability by:

$$P(C_j|d_i) = P(C_j)\frac{P(d_i|C_j)}{P(d_i)} \tag{6.2}$$

Bayes' theorem provides a method to compute $P(C_j|d_i)$ by estimating the conditional probability of seeing particular documents of class $C_j$ and the unconditional probability of seeing a document of each class. If we make the *word independence* assumption which states that the probability of each word occurring in a document is independent of the occurrences of other words in the document, then this probability can be estimated by:

$$P(d_i|C_j) = P(|d_i|)|d_i|!\prod_{t=1}^{|V|}\frac{P(X_t|C_j)^{N_{it}}}{N_{it}!} \tag{6.3}$$

In this formula, the $P(X_t|C_j)$ probabilities are estimated from the word frequencies in training documents. This is the well-known Naive Bayes text classifier.

Often, various dimensionality reduction techniques are used to decrease the size of the vocabulary, increase the efficiency of the computation, and avoid "over-fitting." Furthermore, the notion of word dependence may be incorporated in the model to achieve a better categorization accuracy.

For the rest of this section, we will describe the notions of mutual information and K-Dependence which we apply for dimensionality reduction and accounting for word interdependences.

Recall that a document is represented as a vector of variable (feature) values $X_1, X_2, \ldots, X_{|V|}$. In general, these variables are not independent of each other. For example, suppose the two variables $X_i$ and $X_j$ represent "sunshine" and "rain" respectively, then it is reasonable to expect some dependency between these two variables. The question then is how to represent (and learn) just those dependencies between variables which are significant.

**Table 6.1.** Example document/feature/category matrix.

| Doc | Cat | $X_1$ | $X_2$ | $X_3$ | $X_4$ | $X_5$ | $X_6$ |
|---|---|---|---|---|---|---|---|
| 1 | $C_1$ | 1 | 0 | 1 | 1 | 0 | 0 |
| 2 | $C_1$ | 1 | 0 | 1 | 0 | 0 | 0 |
| 3 | $C_1$ | 1 | 0 | 0 | 0 | 0 | 0 |
| 4 | $C_2$ | 1 | 1 | 1 | 1 | 0 | 0 |
| 5 | $C_2$ | 0 | 1 | 0 | 1 | 1 | 1 |
| 6 | $C_3$ | 0 | 0 | 1 | 0 | 1 | 0 |
| 7 | $C_3$ | 0 | 0 | 1 | 0 | 1 | 1 |
| 8 | $C_3$ | 0 | 0 | 0 | 0 | 1 | 1 |
| 9 | $C_3$ | 0 | 0 | 1 | 0 | 1 | 1 |
| 10 | $C_3$ | 0 | 0 | 1 | 1 | 1 | 1 |

In what follows we will use the example shown in Table 6.1, which represents ten documents, containing six different features (words, say), classified in three different categories. Each row represents a document, its label, and presence (1) or absence (0) of the feature in the document.

### Bayesian networks

A *Bayesian network* is a directed acyclic graph, where each node represents a variable, and each edge represents a probability dependency. In general, the joint probability of a set of variables may be computed based on knowing the conditional probability of each variable given the others, as for example for six variables:

$$\begin{aligned} P(X_1, X_2, X_3, X_4, X_5, X_6) = P(X_1) \cdot P(X_2|X_1) \cdot P(X_3|X_2, X_1) \\ \cdot P(X_4|X_3, X_2, X_1) \cdot P(X_5|X_4, X_3, X_2, X_1) \\ \cdot P(X_6|X_5, X_4, X_3, X_2, X_1) \end{aligned} \tag{6.4}$$

This is difficult to calculate, though, since it requires knowing potentially very large conditional probability tables.

However, if we can assume that $X_1, X_2, X_3, X_4, X_5, X_6$ are independent, then:

$$P(X_1, X_2, X_3, X_4, X_5, X_6) = P(X_1) \cdot P(X_2) \cdot P(X_3) \cdot P(X_4) \cdot P(X_5) \cdot P(X_6) \tag{6.5}$$

which is easy to calculate, though the independence assumption is quite strong and may therefore lead to poor classification.

As we will see, however, we can uncover the key dependencies for conditional probabilities of $P(X_1, X_2, X_3, X_4, X_5, X_6)$, balancing computational ease with effectiveness. Suppose we assume the variables are assumed to be

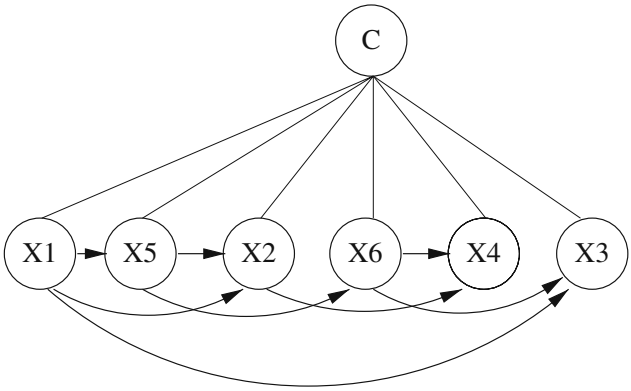



**Fig. 6.2.** Bayesian net.

related in a Bayesian network as in Fig. 6.2. Then we compute the joint probability over the variables by:

$$P(X_1, X_2, X_3, X_4, X_5, X_6) = P(X_1) \cdot P(X_2) \cdot P(X_3|X_1) \cdot P(X_4|X_1, X_2) \cdot P(X_5|X_2, X_3) \cdot P(X_6|X_2, X_5) \tag{6.6}$$

To estimate probabilities from a set of training documents, we consider how many documents contain each feature or are assigned to each class. For example, the joint probability $P(C_1, X_1 = 1)$ may be estimated from the training data as the fraction of documents both in category $C_1$ and containing $X_1$. The simple maximum-likelihood estimate thus gives $P(C_1, X_1 = 1) = \frac{3}{10}$.

Considering the estimates for all classes and features in our example, we would estimate:

$$\begin{aligned} P(C_1, X_1 = 1) &= 3/10 \\ P(C_1, X_1 = 0) &= 0/10^* \\ P(C_2, X_1 = 1) &= 1/10 \\ P(C_2, X_1 = 0) &= 1/10 \\ P(C_3, X_1 = 1) &= 0/10^* \\ P(C_3, X_1 = 0) &= 5/10 \end{aligned} \tag{6.7}$$

For the zero cases marked with a *, we use a smoothing method due to Kohavi et al. (1997), in which we replace the zero count with a factor that is inversely proportional to the number of instances $n$:

$$\frac{N + 1/n}{n + 2/n} \tag{6.8}$$

where $n$ is the total number of documents and $N$ is the number of times the feature and category co-occur. The only case we care about here is when

$N = 0$. In other words, instead of adding 1 to the numerator and 2 to the denominator as in Laplace smoothing, add $1/n$ to the numerator and $2/n$ to the denominator, but *only in cases where the numerator is zero.*

So we would estimate:

$$\begin{aligned} P(C_1, X_1 = 0) &= \frac{0 + 1/10}{10 + 2/10} = \frac{1/10}{102/10} = \frac{1}{102} \\ P(C_3, X_1 = 1) &= \frac{0 + 1/10}{10 + 2/10} = \frac{1/10}{102/10} = \frac{1}{102} \end{aligned} \tag{6.9}$$

This smoothing is done to overcome the situation where a feature does not occur together with a category. In theory, the non-zero cases should be normalized so the sum of probabilities becomes 1.0, however in practice this adjustment is ignored in the implementation.

Now, to calculate the conditional probabilities:

$$\begin{aligned} P(C_1|X_1 = 1) &= P(C_1, X_1 = 1)/P(X_1) \\ &= \frac{3/10}{4/10} = \frac{3}{4} \\ P(C_1|X_1 = 0) &= P(C_1, X_1 = 0)/P(X_1) \\ &= \frac{1/102}{4/10} = \frac{10}{408} \end{aligned} \tag{6.10}$$

We associate a table of conditional probabilities to each node:

| $X_1$ | $P(X_1)$ $C_1$ | $P(X_1)$ $C_2$ | $P(X_1)$ $C_3$ |
|---|---|---|---|
| 1 | $\frac{3}{3}$ | $\frac{1}{2}$ | $\frac{1/102}{5/10} = \frac{10}{510}$ |
| 0 | $\frac{1/102}{3/10} = \frac{10}{306}$ | $\frac{1}{2}$ | $\frac{5}{5}$ |

To deal with the zero cases, Kohavi's trick is used. Consider the case where $X_1 = 0$ for category $C_1$. What we are really calculating is $P(X_1 = 1|C_1)$ which is:

$$P(X_1, C_1)/P(C_1) = \frac{0/10}{3/10} \Rightarrow \frac{\frac{0+1/10}{10+2/10}}{3/10} = \frac{1/102}{3/10} = \frac{10}{306} \tag{6.11}$$

An example for two features:

| $X_3$ | $X_1$ | $P(X_3\|X_1)$ $C_1$ | $P(X_3\|X_1)$ $C_2$ | $P(X_3\|X_1)$ $C_3$ |
|---|---|---|---|---|
| 1 | 1 | $\frac{2}{3}$ | $\frac{1}{2}$ | $\frac{1}{102}$ * |
| 1 | 0 | $\frac{1}{102}$ * | $\frac{1/102}{1/10} = \frac{10}{102}$ | $\frac{4}{5}$ |
| 0 | 1 | $\frac{1}{3}$ | $\frac{1/102}{1/10} = \frac{10}{102}$ | $\frac{1}{102}$ * |
| 0 | 0 | $\frac{1}{102}$ * | $\frac{1}{2}$ | $\frac{1}{5}$ |

Here there is a problem with a $\frac{0}{0}$. For the case $X_3 = 1$ and $X_1 = 0$ what we are really computing is the probability that $X_3$ is 1 given that $X_1 = 0$ and we only care about category $C_1$:

$$P(X_3 = 1|X_1 = 0, C_1) = \frac{P(X_3 = 1, X_1 = 0, C_1)}{P(X_1 = 0, C_1)} = \frac{0/10}{0/10} \tag{6.12}$$

You can't apply Kohavi's trick to both the numerator and denominator, because then the probability would be 1! Nor can you add $1/10$ to the numerator and $2/10$ to the denominator since then the probability is $1/2$. So, in case of $\frac{0}{0}$, we replace the whole thing with Kohavi's trick applied to the numerator: $\frac{0}{0} \Rightarrow \frac{0+1/10}{10+2/10} = \frac{1}{102}$.

The probability of $X_5$ given $X_2$, $X_3$, and we are in category $C_1$:

| $X_5$ | $X_2$ | $X_3$ | $C_1$ |
|---|---|---|---|
| 1 | 1 | 1 | $\frac{0/10}{2/10} \Rightarrow 10/204$ |
| 1 | 1 | 0 | $\frac{0/10}{0/10} \Rightarrow 1/102$ |
| 1 | 0 | 1 | $\frac{0/10}{2/10} \Rightarrow 10/204$ |
| 1 | 0 | 0 | $\frac{0/10}{0/10} \Rightarrow 1/102$ |
| 0 | 1 | 1 | $\frac{0/10}{0/10} \Rightarrow 1/102$ |
| 0 | 1 | 0 | $\frac{0/10}{0/10} \Rightarrow 1/102$ |
| 0 | 0 | 1 | $\frac{2/10}{2/10} = 1$ |
| 0 | 0 | 0 | $\frac{1/10}{1/10} = 1$ |

Recall:

$$P(X_5|X_2, X_3, C_1) = \frac{P(X_5, X_2, X_3, C_1)}{P(X_2, X_3, C_1)} \tag{6.13}$$

**Learning Bayesian networks**

A Bayesian network may be learned from the training data sets in two steps. In the first step, the structure of the graph is found by figuring out the dependencies between variables (i.e. which node is connected to the other). The structure of the graph defines which variables are conditionally independent of which other variables. In the second step, the probability distributions are calculated by using the empirical conditional frequencies from the training data sets. We attach the distribution tables to the graph nodes.

In the step one of network learning, we first assume an ordering of variables, say $X_1, X_2, \ldots, X_{|V|}$. Now for category node $C$, we create a network with $V+1$ nodes labeled $X_1, X_2, \ldots, X_{|V|}$, $C$. The node $C$ is connected to each node $X_i$ via an edge from $C$ to $X_i$. Since it is computationally expensive to discover all the dependencies in the network, we put certain constraints on this structure. For example, in the Naive Bayes network, it is assumed that there are no

more edges. According to Sahami (1996), a *K-Dependence Bayesian* classifier is a Bayesian network in which each feature $X_i$ may have a maximum of K features as parents. For example, the network in Fig. 6.2 is a 2-Dependence Bayesian network.

Sahami (1996) uses the notion of *Mutual Information (MI)* to find dependencies in the network. Intuitively, MI is a data-driven criterion for ranking features according to their usefulness in distinguishing categories.

For a feature $X_i$, define its mutual information (MI) with class $C$ as:

$$MI(X_i, C) = H(C) - H(C|X_i) \tag{6.14}$$

where $H(C)$ is the entropy of category $C$ and $H(C|X_i)$ is the conditional entropy. So that:

$$MI(X_i, C) = -\sum_{C} P(C) \log P(C) + \sum_{C, X_i} P(C, X_i) \log P(C|X_i) \tag{6.15}$$

We may also define the conditional mutual information for two features $X_i$, $X_j$, and category $C$ as:

$$MI(X_i, X_j|C) = \sum_{X_i, X_j, C} P(X_i, X_j, C) \log \frac{P(X_i, X_j|C)}{P(X_i|C)P(X_j|C)} \tag{6.16}$$

We used the following algorithm from (Sahami 1996) for creating a K-Dependence Bayesian net:

(i) $\forall X_i$ compute $\alpha \leftarrow MI(X_i, C)$
(ii) Sort and renumber features $X_1 \ldots X_n$ in descending order by $\alpha$
(iii) $\forall i \neq j$ compute $\gamma \leftarrow MI(X_i, X_j|C)$
(iv) for i=1 $\ldots n$ do

  (a) $r \leftarrow \min(i - 1, k)$
  (b) $X_{di} \leftarrow r$ features $X_{J_1} \ldots X_{J_r}$ with largest $\gamma_{iJ_l}$ where $J_l < i$
  *this step is usually modified by picking $\gamma_{iJ_l} > \theta$ for some threshold $\theta$. This way some dependencies with be eliminated. In experiments $\theta \cong 0.03$.*
  (c) parents$(X_i) \leftarrow \{C\} \cup X_{di}$

(v) compute the conditional probability tables using the network structure and training data set.

**K-dependence learning example**

We now show a detailed example from Sahami (1996) that illustrates the operation of the above algorithm. Recall that the MI of a feature $X_i$ is defined as:

$$MI(X_i, C) = -\sum_{C} P(C)\log P(C) + \sum_{C,X_i} P(C, X_i)\log P(C|X_i) \tag{6.17}$$

where $X_i \in \{1,0\}$, $C \in \{C_1, C_2, C_3\}$ and log is shorthand for $\log_2$. Then we may compute the class entropy:

$$\begin{aligned} -\sum_{C} P(C)\log P(C) &= -(P(C_1)\log P(C_1) + P(C_2)\log P(C_2) + P(C_3)\log P(C_3)) \\ &= -(3/10\log(3/10) + 2/10\log(2/10) + 5/10\log(5/10)) \\ &= -(0.521 - 0.464 - 0.5) \\ &= 1.485 \end{aligned} \tag{6.18}$$

And now, to calculate the second term, we must compute:

$$\begin{aligned} \sum_{C,X_1\in\{0,1\}} P(C, X_1)\log P(C|X_i) &= P(C_1, X_1 = 1)\log P(C_1|X_1 = 1) \\ &= P(C_1, X_1 = 0)\log P(C_1|X_1 = 0) \\ &= P(C_2, X_1 = 1)\log P(C_2|X_1 = 1) \\ &= P(C_2, X_1 = 0)\log P(C_2|X_1 = 0) \\ &= P(C_3, X_1 = 1)\log P(C_3|X_1 = 1) \\ &= P(C_3, X_1 = 0)\log P(C_3|X_1 = 0) \end{aligned} \tag{6.19}$$

So, to complete the MI calculation, we evaluate:

$$\begin{aligned} \sum_{C,X_1\in\{0,1\}} P(C, X_1)\log P(C|X_1) = &\frac{3}{10}\log\frac{3/10}{4/10} \\ &+ \frac{1}{102}\log\frac{1/102}{6/10} \\ &+ \frac{1}{10}\log\frac{1/10}{4/10} \\ &+ \frac{1}{10}\log\frac{1/10}{6/10} \\ &+ \frac{1}{102}\log\frac{1/102}{4/10} \\ &+ \frac{5}{10}\log\frac{5/10}{6/10} \\ = &-.124 - .058 - .2 \\ &-.258 - .052 - .132 \\ = &-.825 \end{aligned} \tag{6.20}$$

And then we have:

$$MI(X_1, C) = 1.485 - .825 = .66 \tag{6.21}$$

**Conditional mutual information for two features given $C$**

Recall that the conditional mutual information for two features $X_i$, $X_j$, and category $C$ was defined as:

$$MI(X_i, X_j|C) = \sum_{X_i, X_j, C} P(X_i, X_j, C) \log \frac{P(X_i, X_j|C)}{P(X_i|C)P(X_j|C)} \tag{6.22}$$

Thus, for example, to compute $MI(X_3, X_1|C)$, we perform the following calculations (note the use of Kohavi's smoothing to deal with zero counts):

$$\begin{aligned} P(X_3, X_1, C) => & P(X_3 = 1, X_1 = 1, C_1) = 2/10 \\ & P(X_3 = 1, X_1 = 0, C_1) = 1/102 \\ & P(X_3 = 0, X_1 = 1, C_1) = 1/10 \\ & P(X_3 = 0, X_1 = 0, C_1) = 1/102 \\ & P(X_3 = 1, X_1 = 1, C_2) = 1/10 \\ & P(X_3 = 1, X_1 = 0, C_2) = 1/102 \\ & P(X_3 = 0, X_1 = 1, C_2) = 1/102 \\ & P(X_3 = 0, X_1 = 0, C_2) = 1/10 \\ & P(X_3 = 1, X_1 = 1, C_3) = 1/102 \\ & P(X_3 = 1, X_1 = 0, C_3) = 4/10 \\ & P(X_3 = 0, X_1 = 1, C_3) = 1/102 \\ & P(X_3 = 0, X_1 = 0, C_3) = 1/10 \end{aligned} \tag{6.23}$$

$$\begin{aligned} P(X_3, X_1|C) => & P(X_3 = 1, X_1 = 1|C_1) = 2/3 \\ & P(X_3 = 1, X_1 = 0|C_1) = \frac{1/102}{3/10} = 10/306 \\ & P(X_3 = 0, X_1 = 1|C_1) = 1/3 \\ & P(X_3 = 0, X_1 = 0|C_1) = \frac{1/102}{3/10} = 10/306 \\ & P(X_3 = 1, X_1 = 1|C_2) = 1/2 \\ & P(X_3 = 1, X_1 = 0|C_2) = \frac{1/102}{2/10} = 10/204 \\ & P(X_3 = 0, X_1 = 1|C_2) = \frac{1/102}{2/10} = 10/204 \\ & P(X_3 = 0, X_1 = 0|C_2) = 1/2 \\ & P(X_3 = 1, X_1 = 1|C_3) = \frac{1/102}{5/10} = 10/510 \\ & P(X_3 = 1, X_1 = 0|C_3) = 4/5 \\ & P(X_3 = 0, X_1 = 1|C_3) = \frac{1/102}{5/10} = 10/510 \\ & P(X_3 = 0, X_1 = 0|C_3) = 1/5 \end{aligned} \tag{6.24}$$

$$\begin{aligned} P(X_3|C) => & P(X_3 = 1|C_1) = 2/3 \\ & P(X_3 = 0|C_1) = 1/3 \\ & P(X_3 = 1|C_2) = 1/2 \\ & P(X_3 = 0|C_2) = 1/2 \\ & P(X_3 = 1|C_3) = 4/5 \\ & P(X_3 = 0|C_3) = 1/5 \end{aligned} \tag{6.25}$$

$$\begin{aligned} P(X_1|C) => & P(X_1 = 1|C_1) = 3/3 \\ & P(X_1 = 0|C_1) = \frac{1/102}{3/10} = 10/306 \\ & P(X_1 = 1|C_2) = 1/2 \\ & P(X_1 = 0|C_2) = 1/2 \\ & P(X_1 = 1|C_3) = \frac{1/102}{5/10} \\ & P(X_1 = 0|C_3) = 5/5 \end{aligned} \tag{6.26}$$

And thus we may finally calculate:

$$\begin{aligned} MI(X_3, X_1|C) = & (2/10) \log \frac{2/3}{(2/3)(3/3)} = 0.0000 \\ & + (1/102) \log \frac{10/306}{(2/3)(10/306)} = .0057 \\ & + (1/10) \log \frac{1/3}{(1/3)(3/3)} = 0.0000 \\ & + (1/102) \log \frac{10/306}{(1/3)(10/306)} = .01553 \\ & + (1/10) \log \frac{1/2}{(1/2)(1/2)} = 0.100 \\ & + (1/102) \log \frac{10/204}{(1/2)(1/2)} = -0.023 \\ & + (1/102) \log \frac{10/204}{(1/2)(1/2)} = -0.023 \\ & + (1/10) \log \frac{1/2}{(1/2)(1/2)} = 0.100 \\ & + (1/102) \log \frac{10/510}{(4/5)(10/510)} = 0.0032 \\ & + (4/10) \log \frac{4/5}{(4/5)(5/5)} = 0.000 \\ & + (1/102) \log \frac{10/510}{(1/5)(10/510)} = 0.0228 \\ & + (1/10) \log \frac{1/5}{(1/5)(5/5)} = 0.0000 \\ & = 0.2009 \end{aligned} \tag{6.27}$$

### How to use the Bayesian net

Suppose, now that we have trained our Bayesian network, we want to classify a new document, #11.

| Doc | Cat | $X_1$ | $X_2$ | $X_3$ | $X_4$ | $X_5$ | $X_6$ |
|---|---|---|---|---|---|---|---|
| 11 | ? | 1 | 0 | 1 | 1 | 1 | 1 |

Each node in the Bayesian net will have a table of probabilities for each term or feature $X_1$ through $X_6$ as depicted in Fig. 6.3. For each category, the document will calculate the probability:

$$P(X_1 = 1, X_2 = 0, X_3 = 1, X_4 = 1, X_5 = 1, X_6 = 1) \\ = P(X_1 = 1)P(X_5 = 1|X_1 = 1)P(X_2 = 0|X_1 = 1, X_5 = 1)P(X_6 = 1|X_5 = 1) \\ \times P(X_4 = 1|X_2 = 0, X_6 = 1)P(X_3 = 1|X_6 = 1, X_1 = 1) \tag{6.28}$$

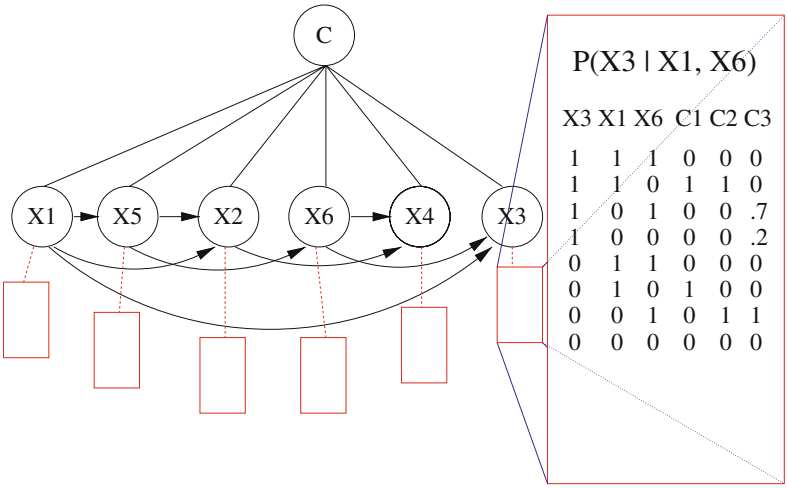



**Fig. 6.3.** Bayesian net with tables

As Fig. 6.3 shows, the value for $P(X_3 = 1|X_1 = 1, X_6 = 1)$ is 0 for $C_1$, 0 for $C_2$ and 0 for $C_3$.

The category giving the highest result from Eq. 6.28 is the predicted category of document #11.

### K-dependence algorithm example

The algorithm from (Sahami 1996) for creating a Bayesian net was discussed above. Figure 6.2 is the correct network for the original example assuming $\theta$ is 0. The sorted MI's for the terms are:

```
x1 0.6603039572
x5 0.6603039572
x2 0.5743644504
x6 0.3834239394
x4 0.276307015
x3 0.04483810157
```

They have not been renumbered simply to avoid confusion. The conditional MI's are:

```
C = {C1, C2, C3}
mi(x1,x1|C) = 0.2579106921
mi(x1,x2|C) = 0.1236067166
mi(x1,x3|C) = 0.2011057519
mi(x1,x4|C) = 0.06680177638
mi(x1,x5|C) = 0.2579106921
mi(x1,x6|C) = 0.2282188377

mi(x2,x1|C) = 0.1236067166
mi(x2,x2|C) = 0.1466508072
mi(x2,x3|C) = 0.06680177638
mi(x2,x4|C) = 0.08984586704
mi(x2,x5|C) = 0.1236067166
mi(x2,x6|C) = 0.09391486213

mi(x3,x1|C) = 0.2011057519
mi(x3,x2|C) = 0.06680177638
mi(x3,x3|C) = 0.676754736
mi(x3,x4|C) = 0.1041562155
mi(x3,x5|C) = 0.2011057519
mi(x3,x6|C) = 0.2015543796

mi(x4,x1|C) = 0.06680177638
mi(x4,x2|C) = 0.08984586704
mi(x4,x3|C) = 0.1041562155
mi(x4,x4|C) = 0.5654948511
mi(x4,x5|C) = 0.06680177638
mi(x4,x6|C) = 0.06725040403

mi(x5,x1|C) = 0.2579106921
mi(x5,x2|C) = 0.1236067166
mi(x5,x3|C) = 0.2011057519
mi(x5,x4|C) = 0.06680177638
mi(x5,x5|C) = 0.2579106921
mi(x5,x6|C) = 0.2282188377
```

```
mi(x6,x1|C) = 0.2282188377
mi(x6,x2|C) = 0.09391486213
mi(x6,x3|C) = 0.2015543796
mi(x6,x4|C) = 0.06725040403
mi(x6,x5|C) = 0.2282188377
mi(x6,x6|C) = 0.5038790175
```

For $k = 2$ we get the following for each loop at step 4 of the algorithm:

| i | r | Term | Possible Parents | Parents Selected |
|---|---|---|---|---|
| 1 | 0 | $X_1$ | none | none |
| 2 | 1 | $X_5$ | $X_1$ | $X_1$ |
| 3 | 2 | $X_2$ | $X_1$ $X_5$ | $X_1$ $X_5$ |
| 4 | 2 | $X_6$ | $X_1$ $X_5$ $X_2$ | $X_1$ $X_5$ |
| 5 | 2 | $X_4$ | $X_1$ $X_5$ $X_2$ $X_6$ | $X_2$ $X_6$ |
| 6 | 2 | $X_3$ | $X_1$ $X_5$ $X_2$ $X_6$ $X_4$ | $X_1$ $X_6$ |

## Evaluation results

Many tests have been perfomed to evaluate the effectiveness of the classification system. Based on experimental results, we set the threshold to 0.9999. In other words, if the odds of a document being NON-SU outweighs the odds of it being SU by more than 0.9999, then the document is marked as NON-SU. The likelihood of a document being marked as NON-SU when it is actually SU is 1 in 10,000.

In a sample of documents with known SU documents which were selected by DOE, testing showed that 100% of SU documents were marked as "potentially SU" and 70% of the NON-SU documents were marked correctly. This left only 30% of the documents requiring review. The review application reduces this manual review task even further. First, only 20% of the pages are marked as POSSIBLY- SU and these pages are highlighted with the most sensitive passages. Using the image-based review application only 10% of these pages required a manual review before the document was determined to be either SU or NON-SU.

The combined use of the classifier and the interactive tool reduced manual efforts by 95%. Based on ISRI's page-by-page SU analysis review, on average, manual inspection with no automated processing takes approximately 2 min 45 s per page, so for every million pages processed, the amount of work is reduced by nearly 95% from approximately 40,277 to 2400 h.

In actual operational enviornment, the classification system has only marked 15% of the documents as POSSIBLY-SU based on processing of a few million documents.

## OCR error analysis

The majority of LSN documents are only available in hard copy. So before processing for SU information, they were converted to electronic form via an OCR system.

ISRI has extensively studied the effectiveness of IR on OCR versus clean text (Taghva et al. 1996). As mentioned previously, we fed many groups of features such as sensitive patterns to the K-Dependence algorithms. Typically these features are extracted from documents using IE technologies. An interesting question is whether OCR errors affect IE.

In recent studies it has become apparent that the *information extraction* task differs from text retrieval in that performance is affected for the worse by OCR errors. Miller et al. (2000) noted degradation in the accuracy of their Hidden Markov Model (HMM) information extraction system IdentiFinder. They printed copies of Wall Street Journal articles and then ran experiments on progressively degraded images of these with progressively higher word error rates. The system suffered a 0.6 point loss in F1-measure for each percentage point increase in word error. We had similar difficulties applying address extracting HMM and date-of-birth extracting HMM to noisy texts (Taghva et al. 2006).

In our applications, the extractors look for relations among features in the documents. To detect a particular type of feature, it is useful to identify other types of entities which often appear near that particular type. In some cases, to be considered a feature, there must be a relationship between data of different types. To make this point more clear, let us consider the extraction of date-of-birth for the purpose of privacy identification and redaction:

A date in isolation is not private information. Nor does it become private if it is identified as a birthdate with an identifier such as *date of birth*. A date becomes private information only when it is correctly associated with a person. The statement *John Doe's date of birth is 5/17/55* is private, but the phrase *someone's birthdate is 5/17/55* is not, unless a referent for *someone* is implied contextually. The date-of-birth identification is an example of `relational extraction` of the form $E_1RE_2$, where $E_i$ and $R$ denote an `entity` and its `relation`, respectively. In the above text segment, "John Doe" is the value of $E_1$, "5/17/55" is the value of $E_2$, and the relationship $R$ is birthdate which is deduced from the phrase "date of birth."

When recognizing a document via OCR, two types of errors can occur. The first type of error is *classification*. For example, OCR the word "born" could potentially be recognized as "bom" by an OCR device. Visually, the characters "rn" are similar to the character "m." Similarly, characters such as "1,", "l,", "i," and "j" cause confusion for OCR. These confusions can cause misspellings in both entities and indicators. In general, for any misspelling of the form $B_0B_1 \cdots B_n$, our goal is to find its correct spelling. This can be achieved by maximizing the conditional probability $P(A_0A_1 \cdots A_n|B_0B_1 \cdots B_n)$ for any correct word $A_0A_1 \cdots A_n$ in our lexicon. In other words, the probability that

given the OCR has produced $B_0B_1 \cdots B_n$, the word $A_0A_1 \cdots A_n$ is the correct word. Assuming independence, we can write $P(A_0A_1 \cdots A_n|B_0B_1 \cdots B_n)$ as $P(A_0|B_0) \star P(A_1|B_1) \cdots \star P(A_n|B_n)$, where $P(A_i|B_i)$ is the probability that given the OCR produced $B_i$, it is actually $A_i$. These marginal probabilities can be estimated using Bayes' formula. In our setting, we use Bayesian techniques to estimate probabilities such as $p(born|bom)$. This technique is not only applied to indicators but to entities as well. Figure 6.4 shows some of the OCR errors we encountered while collecting training and performing testing.

| *Error* | *Correct* |
|---|---|
| 1S40 | 1940 |
| 12/12[53 | 12/12/53 |
| ,', /e.2 / | 11/02/44 |
| Z | 08/28/30 |
| IL | 08/28/30 |
| 9//s /ro | 9/16/64 |
| 03111/1964 | 03/11/1964 |
| ^9so | 04/02/1960 |
| duty of birth | date of birth |
| 5/13/55 | 5.413,L4;^ |
| 12/1/24 | 12 124 |
| DATE :.::D | DATE AND |
| PLACE CF | PLACE OF |
| ;,9RTH:- | BIRTH |

**Fig. 6.4.** OCR errors affecting extraction.

The second type of errors that can affect extraction is *zoning* error. An example of a zoning errors is when the OCR confuses a two column page as one column. In our experiments, we calculated that 37.5% of the errors were classification errors and 63.5% due to zoning.

Because of the classification and zoning errors in extraction, we decided to only process a document for SU information if its word accuracy was over a certain threshold. Obviously, this adds to the manual task!

## Conclusion and future work

The SU classifier has been used to sieve through millions of documents looking for homeland security sensitive materials. The system has eliminated over 95% of the manual efforts. The solution and implementation of the SU classifier took over two years. During these two years, many issues and problems have been posed. We believe our solutions to many of these problems can help other security related document processing projects.

There are many open problems such as minimum accuracy requirements for retrieval, categorization, and information extraction. The information display is also a challenge. Many non-technical users do not distinguish between image and text. Most of our image-based interfaces are built with this assumption. Unfortunately, the text being manipulated behind the image does not always match what the user sees so some operations cannot be easily implemented. For example, many of the reviewers would like to view images and have "cut and paste" capabilities. Having an interface as clean as the original image but as flexible as ASCII is not yet available. Our research though has made some progress in this direction.

## References

Kohavi, R., B. Becker, and D. Sommerfield. 1997. Improving simple bayes. In *Proceedings of ECML-97*. `http://robotics.stanford.edu/users/ronnyk/ronnyk-bib.html`

Lewis, D.D. 1991. Evaluating text categorization. In *Proceedings of the Speech and Language Workshop*. `http://robotics.stanford.edu/users/ronnyk/ronnyk-bib.html`

Maron, M.E. 1967. Automatic indexing: An experimental inquiry. *Journal of the ACM*, 8:404–417.

Maron, M.E. and J.L. Kuhns. 1960. On relevance, probabilistic indexing and information retrieval. *Journal of the ACM*, 7(3):216–240.

McCallum, A. and K. Nigam. 1998. A comparison of event models for naive bayes text classification. In *Proceedings of AAAI-98 Workshop on Learning for Text Categorization*. URL `citeseer.nj.nec.com/mccallum98comparison.html`

Miller, D., S. Boisen, R. Schwartz, R. Stone, and R. Weischedel. 2000. Named entity extraction from noisy input: Speech and OCR. In *Proceedings of the Sixth Conference on Applied Natural Languae Processing*, pp. 316–324.

Sahami, M. 1996. Learning limited dependence Bayesian classifiers. In *Second International Conference on Knowledge Discovery in Databases*. `http://robotics.stanford.edu/users/sahami/papers.html`

Taghva, K., J. Borsack, and A. Condit. 1996. Evaluation of model-based retrieval effectiveness with OCR text. *ACM Transaction on Information Systems*, pp. 64–93.

Taghva, K., R. Beckley, and J. Coombs. 2006. The effects of OCR error on the extraction of private information. In *Proceedings of 7th IAPR Workshop on Document Analysis Systems (DAS 2006)*, pp. 348–357.

U.S. Government. 2004. The Freedom of Information Act (FOIA), 5 USC Section 552(b)(6). `http://www.usdoj.gov/oip/exemption6.html`

U.S. Department of Energy. 2001. Licensing support network baselined design requirements. `http://www.lsnnet.gov/`

# 7

# Rich Language Analysis for Counterterrorism

Mathieu Guidère, Newton Howard, and Shlomo Argamon

**Summary.** Accurate and relevant intelligence is critical for effective counterterrorism. Too much irrelevant information is as bad or worse than not enough information. Modern computational tools promise to provide better search and summarization capabilities to help analysts filter and select relevant and key information. However, to do this task effectively, such tools must have access to levels of meaning beyond the literal. Terrorists operating in context-rich cultures like fundamentalist Islam use messages with multiple levels of interpretation, which are easily misunderstood by non-insiders. This chapter discusses several kinds of such "encryption" used by terrorists and insurgents in the Arabic language, and how knowledge of such methods can be used to enhance computational text analysis techniques for use in counterterrorism.

## Introduction

The world today is shaped by an information society where massive amounts of data circulate in multiple languages on the same critical subjects. These data often contain strategically useful information and reflect perceptions relevant on both the local and global levels. However, the vast amount of information transferred hinders real time evaluation of the flow of these data and the evolution of the perceptions they contain.

For the benefit of decision makers, trivial and irrelevant information must be distinguished from information essential to decision-making. Indeed, an overabundance of data can disturb perception. Furthermore, individuals require a hierarchy of information that depends on their immediate relevance, as most of the available data is inconsequential. Only relevant elements merit particular attention. This complex situation calls for comprehension of how the human mind understands both explicit and implicit meanings from text, so that we may search out truly relevant information.

The meaning of a text cannot be considered in isolation, but rather with reference to its culture and social context. The relationship between the social frame of reference and individual aspirations is quite significant, since an

S. Argamon, N. Howard (eds.), *Computational Methods for Counterterrorism*, DOI 10.1007/978-3-642-01141-2_7,

individual's experience can generate meanings that differ in terms of certain dominant social norms.

Mastering different levels of linguistic analysis is indeed essential for a sound comprehension of salient information on a particular phenomenon or situation. The meaning of a given word or phrase depends crucially on its context, both for disambiguation (consider *head* the body part vs. *head* the organizational role) and to understand deeper cultural meanings. These meanings may arise from stereotyped interactions, such as the use of "How are you?" as a form of greeting rather than an inquiry in modern American English. They may also arise by reference to a shared cultural history, as in how the phrase "to be or not to be" may evoke tragic indecision, referring to Shakespeare's play *Hamlet*. Thus, meaning should be broadly construed to include not just semantics, but also pragmatic and cultural aspects of language use. This is particularly important to keep in mind when dealing with a radically foreign culture, as for westerners dealing with Islamic fundamentalists.

Avoiding errors of judgment based on misperception of such meanings that are very frequent in daily social, political or military situations thus requires awareness of language as a network of complex systems.

The naming of Baghdad's "Green Zone" is a case in point. For the western coalition forces, the expression evokes the relative security of the area, a "secure space." But for jihadists or Islamic insurgents, the "Green Zone" is a privileged target for "sacrifice" because of its very name. In Islam, green symbolizes what is sacred; not only is it the Prophet Mohammad's color, but it also refers to paradise for all Muslims. Fighters therefore desperately seek to sacrifice themselves as martyrs against this "green" zone, whose very name makes it an ideal target for suicide attack. This case represents a typical misconception (by coalition forces), which leads to misperception (by Islamic fighters). They interact and generate a clash of perceptions with concrete implications in the physical world (increased suicide attacks against a zone that is supposed to be safe).

Similarly, invalid understanding of and reaction to the cultural meaning of the Iranian President's political statements on the State of Israel or on nuclear development, may lead to a dangerously incorrect evaluation of the situation. On January 1, 2006, Ahmadinejad declared, "The Europeans have established a Jewish camp in the heart of the Islamic nations... The Zionist regime is a part of Europe that has been detached... It [the regime] is naturally anti-Islamic." Are these statements a serious threat or mere exaggeration? What is their sociocultural context? What should be the response to these words?

The potential answers to questions like this depend on assessing meaning at several levels of linguistic perception. The purpose of this chapter is to outline a framework for considering such questions in terms of three levels of linguistic perception, considering explicit semantic meaning, cultural associations, and goal-oriented intentions. The fundamental question is how to process massive quantities of data in a foreign language (Arabic), using these insights, to extract salient and strategic information about individuals

or groups, particularly radical groups and organizations. We sketch in this chapter several areas of automated language analysis in which our framework can be applied to address such text analysis problems.

More advanced computational text processing techniques promise to enable important applications in counterterrorism:

- To authenticate the attribution of a document, i.e. if an individual/organization $x$ has produced it or not, especially when the source is uncertain;
- To search and identify other documents that match a given document's characteristics, thus allowing us to find other documents written by the same individual or a group;
- To characterize a rhetorical/ideological profile of individuals or groups based on their writing and discourse;
- To characterize the identity (religious, ideological, cultural) of different terrorist groups;
- To make an inventory of the images and the representations of their cultural behavior.

In many of these cases, the task is essentially one of *text classification*, the assignment of input documents to a particular category based on their content. Our goal here is to examine how a deeper analysis of language meaning structures may help with these tasks. While the work reported here work is still preliminary, it points the way towards more effective text analysis for counterterrorism applications.

**Scope**

The main corpus used for the work reported here includes texts relevant to the following countries:

- Iraq: all the armed groups fighting against the government and the coalition forces
- Afghanistan: Taliban groups, especially Al-Qaeda
- Saudi Arabia: Salafi groups and Al-Qaeda related groups
- Algeria: Salafi and Jihadi groups merged in Al-Qaeda

The following Iraqi insurgency groups were studied:

- Al-Qaeda in the Land of the Two Rivers [Iraq]
- Ansar Al-Sunna Army
- Islamic Army in Iraq
- Islamic Front in Iraq
- Mujahidin Army
- Rashidin Army
- Fatihin Army
- Other minor groups

## Levels of meaning

As we discussed, the meaning of any "rich" text must be understood with reference to its cultural and intentional context, not just its denotational meaning. We formulate this notion by considering three levels of analysis of language, as follows:

**Signification:** Language as a system for designating objective meanings directly (denotation). For example, the Arabic word *khadim*[1] literally means "servant";
**Cultural:** Language as a knowledge system embedded in a cultural context. The word *khadim*, for example, is used as the official title of the King of Saudi Arabia (*khadim al-haramayn*, Servant of the Holy Lands), so that the term evokes the title;
**Intentional:** Language as a communication system between individuals sharing a cultural/historical context and value system (expressed in part through a set of canonical proverbs and texts). *Khadim*, for example, is used in the Arabic proverb "*khadimu al-qawmi sayyiduhum*" [the servant of the people is their lord], which reverses the literal meaning of the term since "servant" on the signified level becomes "lord" on the intentional level.

From a cognitive point of view, these three levels of linguistic perception fit together to produce a specific meaning for a given message at a given time from a given sender to a particular recipient in a particular situation. Due to the complexity of the full system (requiring deep knowledge of the culture and history as well as the particulars of the communication situation), an individual may only perceive one or another aspect of a message's meaning, and may misunderstand others.

## Cultural encryption

Understanding these levels of meaning allows us to investigate the use of deep culture and shared goals for message encryption. We call this *cultural encryption*. For example, Islamic terrorists and insurgents often use Arabic poetry in order to encrypt their messages and communications. The poetry itself is a public "key" accessible to a wide scope of people (i.e. all those who understand Arabic will understand the literal meaning of the words). But there are layers of meaning with more private keys. One of these keys resides at the cultural level and others reside at the intentional level.

To understand this, consider the following example. Before his well-known video where he appeared in the desert with his main commanders, Al-Zarqawi disseminated an audio message, where he quoted Arabic poetry. At that time, few people understood the reason behind using Arabic poetry in the context

[1] We use throughout the Arabic transcription of the *Encyclopaedia of Islam*, Netherlands, Brill, 2002.

of guerrilla warfare. The poetry verse he used was a very well known verse by the medieval poet Al-Mutanabbi.

الخيل والّيل والبيدَاء تعرفني
والسيف والرمح والقرطَاس والقلم

*Al-khaylu wal-laylu wal-bayda'u ta'arifuni*
*Was-sayfu war-rumhu wal-qirtasu wal-qalamu*

The horse and the night and the wilderness know me
And the sword and the spear and the paper and the quill

When we compared the content of this verse with Al-Zarqawi's subsequent actions, we discovered the culturally encrypted message behind the poetic communication, as decoded in the table below:

| | |
|---|---|
| Al-khaylu | means the way: "Convoy" |
| Al-laylu | indicates Time: "by night" |
| Al-bayda | indicates Space: "in the desert" |
| Ta'arifuni | literally: "I will be there, you will see me" |
| As-sayfu | indicates the weapons: light weaponry |
| Ar-rumhu | symbolizes the Instrument: Rocket |
| Al-qirtasu | indicates the Method: Video taped |
| Al-qalamu | indicates the Objective: Propaganda |

These correlations have been verified as more general by analyzing other texts where these terms appear. Cultural encryption has been demonstrated for a number of words when insurgents use Arabic poetry, such as:

"Horses" = Convoy
"Spear" = Rocket
"Arch" = Sniper
"Sword" = Kalashnikov

For the general audience (Arabic-speaking Muslims), it would be just a verse in poetry that is normally used for motivation. For the specific audience (Mujahiddin), it is an announcement of a big upcoming operation with media coverage. Finally, for the targeted audience, his own group and commanders, who share his intentions, the literary text could be decrypted as follows:

> We will be in convoy (*Al-kahylu*), we will meet by night (*Al-laylu*), in the desert (*Al-bayda*), I will be there and you will see me (*Taarifuni*), we will use light weaponry (*Assayfu*), rockets (*Arrumhu*) and I want it to be covered (*Al-qirtasu*), in order to achieve propaganda objectives (*Al-qalamu*).

All these details were encrypted in one Arabic verse—invisible to any not aware of the cultural and intentional context. Such cultural encryption appeals

to collective cultural memory and thus serves a secondary (though important) purpose of reinforcing group identity and loyalty. In order to uncover such messages and counter them, we need to understand how terrorists think and decipher the way they express their thoughts and intentions.

More generally, cultural encryption or encoding can be categorized into three basic types:

**Historical encryption:** Some historical events such as medieval battles (Badr, Hittin, Yarmuk, etc.) are used to express a type of operation to be planed.
**Poetic encryption:** Some verses and poets such as Islamic panegyrists (Hassan Ibn Thabit, Ibn Rawaha, etc.) are used to convey a specific message based on their poetry (including poetry inspired songs).
**Religious encryption:** Some reference to medieval religious authorities such as Ibn Taymiyya are used to indicate the specific way to act or treat the enemy.

### Encryption in names

One interesting type of encryption is that of names of fighters (Guidère 2006a). Every fighter will choose the name of a companion of the Prophet Mohammed and the chosen name expresses at the same time the personal history of the fighter and his projected image in the future (how he wants to die and for what). So the fighters adopt and the roles and the characters of historically known heroes of Islamic civilization. It is a form of impersonation (Guidère

**Table 7.1.** Different levels of meaning for Arabic names of individuals and groups involved in the Iraq conflict.

| Name | Meaning Level | Knowledge Level | Communication Level |
|---|---|---|---|
| Abu Moussab al-Zarqawi | He is from Zarqa *origin, country* | Abu + Moussab = (father)+(son) *affiliation, bloodline* | *Man of difficulties* sa'ab =*difficult* |
| Abu Maysara al-'Iraqi | He is from Iraq *origin, country* | Abu + Maysara = (father)+(son) *affiliation, bloodline* | *Man of ease* Yassir =*easy* |
| Katibat al-Firdaws | Paradise Brigades *attitudinal* | Martyrs Brigades *cultural* | Personal invitation *spiritual, personal* |
| Katibat al-Haq | Truth Brigades *goal-oriented* | Brigades of God *religious/cultural* (haq is one of the names given to Allah) | Divine justice *military, legal* |

2006b). To illustrate this phenomenon, a list of Arabic names is given in Table 7.1 with their interpretations at the three levels of meaning.

These examples highlight the complexity of the Arabic system of meaning and how perception of different levels of meaning depends on the reader/hearer's level of sociocultural and linguistic knowledge. Opinions and decisions also depend on the type of perception applied at a given time. Establishing a communication goal appropriate for local populations requires conscious attention to these distinctions in a broader goal to follow social and political evolutions behind these visions.

## Noun phrase structure

A central task for automated text analysis systems is finding important key terms that can serve as meaningful indices for the texts. Once such important terms are extracted, they can be classified according to known categories of meanings, using methods as described above. While a number of statistical methods have been developed for finding and classifying important "named entities" in general texts (Nadeau and Sekine 2007), we consider here some linguistic considerations that we have found useful for analysis of terrorist and insurgent documents.

We analyzed a corpus of documents, magazines, and files comprising writings from a variety of Islamist terrorists, including Iraqi insurgents as well as Saudi and Afghan Al-Qaeda cells. Our main task was to extract the names of weapons used by them in battle. The writings included in the corpus represented the doctrine for their warfare since they do not follow a conventional training or doctrine. One example document from the corpus is the "Mawsuat Al-Jihad Al-Afghani" (The Encyclopedia of Afghan Jihad), which consists of more than 1000 pages describing diverse weapons in Arabic. Other important sources included *Sawt Al-Jihad* (*The Voice of Jihad*) and *Muaskar Al-Battar* (*Camp of Al-Battar*), both of which were magazines issued by Al-Qaeda in the Holy Lands (i.e. the Kingdom of Saudi Arabia). Nearly 30 issues of *Sawt Al-Jihad* were produced between 2003 and 2004, comprising more than 500 pages of explanations of weapons and military training in Arabic.

When attempting to extract the weapon names used, we observed that the names used in these documents were different from the standard names of these weapons; indeed, straightforward English translation would not yield the normal English terms for the weapons considered.

Our hand-analysis showed that in most cases, the names of weapons and weapon types fit into a relatively small number sequential syntactic patterns. In fact, many weapon names fit a very simple adjective-noun pattern as shown in Table 7.2. Some more complex patterns were also identified. One of the more common is that shown in the noun phrase "Aslihat Al-Damar Al-Shamil" (weapons of mass destruction) is an instance of the pattern $[N_1 + Art + N_2 + Art + Adj]$ (where $N$ represents a noun, $Art$ an article, and $Adj$ an adjective).

**Table 7.2.** Simple noun phrases and patterns denoting weapon names in the corpus.

| Weapon's Name in Arabic | Weapon's Name in Transcription | Weapon's Name in English |
|---|---|---|
| | $[N + Adj]$ | $[Adj + N]$ |
| سلَح ستَرَتيجي | "Silah Strategi" | Strategic Weapon |
| سلَح مطلق | "Silah Mutlak" | Absolute weapon |
| سلَح بيولوجي | "Silah Biulugi" | Biological weapon |
| سلَح شضوية | "Silah Shadhwi" | Fragmentation weapon |
| اَسلحة ذرّية | "Asliha Dharriyah" | Atomic weapons |
| اَسلحة نوَويّة | "Asliha Nawawiyyah" | Nuclear weapons |

For comparison, the English equivalent "weapons of mass destruction" would have the pattern $[N_1 + \text{of} + Adj + N_2]$. Other examples of this pattern include weapons-related phrases such as "prohibition of atomic weapons" (Hazer Al-Asliha Al-Dharriyyah) and "proliferation of nuclear weapons" (Intishar Al-Asliha Al-Nawawiyyah).

We note that these patterns exemplify a typical noun-phrase pattern in Arabic, i.e., an appositive sequence of nouns and adjectives, with intervening articles. This characteristic implies, and examination of the corpus shows, that in Arabic certain key terms can be found in the intersection of two common patterns. For example, the term "Al-Aslihah" (the weapons) occurs at the intersection of "Intishar Al-Aslihah" ("weapons proliferation," pattern $[N_1 + Art + N_2]$) and "Al-Asliha Al-Nawawiyyah" ("the nuclear weapons," pattern $[Art + N + Art + Adj]$). The combination of the two patterns, $[N_1 + [Art + N_2] + Art + Adj]$, gives the common phrase "Intishar Al-Asliha Al-Dharriyyah" (proliferation of nuclear weapons). Hence, we can use the intersection of common patterns to find significant smaller terms (such as "Al-Asliha" here) that can be used as useful indexing terms in their own right, and also as cues to find other meaningful phrases in the texts.

## Text classification

To analyze holistic properties of texts, such as authorship characteristics, we use the vector space model of information retrieval (Grossman and Frieder 2004) and text categorization algorithms (Sebastiani 2002). First, each document in the corpus is broken down into its most essential linguistic characteristics, token-by-token. Depending on our purpose, the output may include the original words, word roots, parts-of-speech, semantic or cultural categories, or combinations of all of these. Next, the numbers of occurrences of each type

of token are counted up, and their frequencies used to compute a numeric vector which represents the overall contents of the document (relative to the chosen feature set). In some cases, the elements of these vectors will be the simple relative frequencies of the features considered, while for other applications more complex formulas such as the tf-idf formula of information retrieval (Grossman and Frieder 2004) may be applied.

Whatever features are used in a particular experiment, we represent the document vector as $X = (x_1, ..., x_i, ..., x_n)$, where n is the number of features and $x_i$ is the relative frequency of feature $i$ in the document.

In text classification, we are given a corpus of training documents, each labeled according to some category for a particular profiling dimension. For example, documents may be labeled as "martyr's biography" or "other document". Each training document is processed into a vector, labeled by its category for the task. The labeled training vectors are then fed as input to a machine learning algorithm which learns a classifier that can assign new (unseen) documents to known categories. Such a method computes a classifier that, to the exten possible, classifies the training examples correctly. To test the effectiveness of this method for a task, the predictive power of the classifier is tested on out-of-training data.

Most effective multi-class (i.e., more than two classes) classifiers for text classification used today share the same structure: The algorithm computes a weight vector $W_j = (w_{1j}, ..., w_{ij}, ..., w_{nj})$ for each category $c_j$ and then assigns each input document $X$ to the class for which the inner product $W_j * X$ is maximal. There are a number of effective algorithms for learning the weight vectors; the one used here is the SMO support vector machine method (Platt 1998).

Several types of features can be used for construction of document vectors as above. One type, based on previous research on authorship attribution, are function words, typically members of closed word classes such as pronouns, prepositions, and the like (e.g., "you", "of", "also", and "was" in English, or "Al-" and "huwa" in Arabic). In Arabic some of these are not separate words, and so some morphological analysis of text tokens must be used. A second type of features are significant words and phrases, determined by use of statistical methods and linguistic analyses as above. Different sorts of documents will use these terms with different characteristic frequencies. A related feature set are the syntactic patterns of such characteristic phrases. Finally, a novel feature set consists of occurrences of culturally-determined codes, such as the encodings found in the use of Arabic verses used by Al-Zarqawi. These encodings, being specific to a given culture and intention group, can give strong clues as to the nature of the sender and intended recipient of a message.

In one experiment to validate this methodology, we used the vector-space text classification methods described above, using function words and pattern terms meanings to select biographies of "Al-Qaeda martyrs" (suicide bombers) from Arabic-language web pages. The classification approach proved highly

effective, distinguishing biographies of "Al-Qaeda martyrs" from even highly similar biographies of members of the "Al-Aqsa Martyrs Brigade" in Palestine.

**Table 7.3.** Composition of the Arabic blogs corpus for dialect classification, by country of blogger origin. Given are the numbers of blogs from each country and the number classified correctly (under 10-fold cross-validation).

| Country/Region | Num. Blogs | Num. Correct |
|---|---|---|
| Algeria | 4 | 1 |
| Egypt | 19 | 16 |
| Iraq | 10 | 6 |
| Jordan | 25 | 18 |
| Lebanon | 9 | 8 |
| Morocco | 10 | 8 |
| Palestine | 22 | 20 |
| Saudi Arabia | 15 | 10 |
| **Total** | 114 | 87 |

In another experiment, we considered a small corpus of 114 blogs from the *Maktoob* Arabic-language blog site (`http://www.maktoobblog.com`), to see if different Arabic dialects could be distinguished based on language style. Blogs were selected for inclusion as those (i) listing the country of the blogger, where that country was an Arab country or region (i.e., excluding countries such as the U.S., where the true dialect of the blogger could not be assumed). The task was then to classify each blog by its national origin, based only on the text of its blog posts. The full text of the posts in each blog was first processed by Buckwalter's morphological analyzer (Buckwalter 2002), and the relative frequencies of different Arabic parts-of-speech were computed (in case of ambiguity, all alternatives were counted separately—note that this method could only degrade performance and would not produce spuriously high accuracies). The document vectors thus produced were fed into a linear support vector machine for classification; overall 10-fold cross-validation accuracy was 76.3% for eight-way classification (see Table 7.3). While this study was quite small, the results are quite promising for the eventual classification of Arabic dialects in informal text, which can give important clues as to the identity of the text's author.

## Conclusion

The study of language offers an strategic opportunity for improved counterterrorist intelligence, in that it enables the possibility understanding of the Other's perceptions and motivations, be he friend or foe. As we have seen, linguistic expressions have levels of meaning beyond the literal, which it is

critical to address. This is true especially when dealing with texts from a high-context traditionalist culture such as those of Islamic terrorists and insurgents.

Computational and statistical methods offer the promise of reliable automated methods for analysis and authentification of open-source documents. But such methods must also incorporate deeper levels of semantic processing to produce effective intelligence. This chapter has sketched some of the ways in which this may be done, in the context of processing Arabic language documents. With further study, these ideas may also be useful for counter-intelligence and counter-propaganda efforts. A deeper understanding of how texts encrypt cultural assumptions and intentions may enable new types of psychological operations, by allowing us to produce texts that play on deeper aspects of messages currently circulating in the target audience. This approach, applied systemically on a large scope using popular media outlets, can disturb the audience confidence in terrorist's propaganda and create doubt in the minds of the general public in relation to those groups. Thus the groups would lose undeclared public support or even sympathy.

## References

Barnbrook, G. 1996. Language and computers. Edinburgh: Edinburgh University Press.

Biber, D., S. Conrad, and R. 1998. Reppen corpus linguistics, investigating language structure and use. Cambridge: Cambridge UP.

Buckwalter T. 2002. Buckwalter Arabic Morphological Analyzer Version 1.0. Linguistic Data Consortium, catalog number LDC2002L49 and ISBN 1-58563-257-0. http://www.ldc.upenn.edu/myl/morph/buckwalter.html

Charteris-Black, J. 2004. Corpus approaches to critical metaphor analysis. Basingstoke: Palgrave Macmillan.

Chilton, P. 1996. Security metaphors: Cold War discourse from containment to common house : Conflict and consciousness; v.2. New York: Peter Lang.

Croft, W., and D. A. Cruse. 2004. Cognitive linguistics: Cambridge textbooks in linguistics. Cambridge: Cambridge University Press.

Cuyckens, H., R. Dirven, and J. R. Taylor (eds.) 2003. Cognitive approaches to lexical semantics. Cognitive linguistics research; 23. Berlin; New York: Mouton de Gruyter.

Dirven, R., and M. Verspoor (eds.) 2004. Cognitive exploration of language and linguistics. Amsterdam: John Benjamins.

Dirven, R., R. Frank, and C. Ilie (eds.) 2001. Language and ideology. Amsterdam; Philadelphia: John Benjamins.

Dirven, R., R. M. Frank, and M. Putz eds. 2003. Cognitive models in language and thought: ideology, metaphors and meanings. Cognitive linguistics research; 24 . Berlin: Mouton de Gruyter.

Fauconnier, G., and M. Turner. 2002. The way we think: conceptual blending and the mind's hidden complexities. New York: Basic Books.

Fauconnier, G. 1997. Mappings in thought and language. Cambridge, U.K.; New York,: Cambridge University Press.

Goldberg, A. ed. 1996. Conceptual structure, discourse and language. Stanford: CLSI.

Grossman, D., and O. Frieder. 2004. Information retrieval: algorithms and heuristics, 2nd ed. Springer Publishers.

Guidère M. 2006a. Al-Qaeda's Noms de Guerre: How should we decode terrorists' names?. *Defense Concepts*, 1(3):6–16.

Guidère, M. 2006b. *The Al-Qaeda "Martyrs"*, Paris : Editions du Temps, p. 240.

Hudson, R. 2006. Language networks. The new word grammar. New York: Oxford University Press.

Langacker, R. W. 1987. Foundations of cognitive grammar. Stanford, CA.: Stanford University Press.

Langacker, R. W. 2001. Discourse in cognitive grammar. *Cognitive Linguistics* 12:143–88

Lebart, L., A. Salem, and L. Berry. 1997. Exploring textual data. Kluwer Academic Publishers.

Lee, D. 2001. Cognitive linguistics: an introduction. Melbourne; Oxford: Oxford University Press.

Madigan, D., Genkin, A., Lewis, D., Argamon, S., Fradkin, D., and L. Ye. 2005. Author identification on the large scale. Joint Annual Meeting of the Interface and the Classification Society of North America (CSNA).

Mehler, A. 2007. Large text networks as an object of corpus linguistic studies. In: Corpus linguistics. L|deling, A. and Kytv, M. (eds.). An international handbook of the science of language and society, Berlin/New York: de Gruyter.

Nadeau, D. and S. Sekine. 2007. A survey of named entity recognition and classification. *Linguisticae Investigationes*, 30(1):3–26.

Palmer, G. 1996. Toward a theory of cultural linguistics. Austin: University of Texas Press.

Partington, A. 1998. Patterns and meanings: Using corpora for english language research and teaching. Philadephia, PA: John Benjamins.

Semino, E., and J. Culpeper (eds.) 2002. Cognitive stylistics: language and cognition in text analysis . Linguistic approaches to literature; v. 1 . Amsterdam: John Benjamins.

Stefanowitsch, A., and S. Thomas Gries. 2006. Corpora in cognitive linguistics: Corpus-based approaches to syntax and lexis, New York: De Gruyter. Young, S. (ed.). 1997. Corpus-based methods in language and speech processing. Kluver Academic publishers.

Talmy, L. 2000. Toward a cognitive semantics. Cambridge, MA; London: MIT Press.

Tomasello, M. 1998. The new psychology of language: Cognitive and functional approaches to language structure. Mahwah, NJ: Laurence Erlbaum.

Ungerer, F., and S., Hans-Jorg. 1996. An Introduction to Cognitive Linguistics. Longman.

# Part III

# Graphical Models

# 8

# Dicliques: Finding Needles in Haystacks

Robert M. Haralick

**Summary.** A key task in counterterrorism is finding useful records and combinations of records in very large heterogeneous databases. The bits and pieces of information come from many sources and the pieces do not all tightly connect together. Some (possibly disconnected) pieces tightly connect to some other (possibly disconnected) pieces. The databases are very large and the number of records that pertain to terrorist activity is very small. The problem is to find a needle in a haystack, that involves connections between different types of relevant data. This chapter describes the mathematical properties of the diclique structure, which encodes a key concept of what such meaningful 'needles' may look like. We show that the set of dicliques forms a lattice and in two different ways forms commutative idempotent monoids. By investigating groupoid homomorphisms on the monoid structure, we show how to coalesce overlapping dicliques and how coalesing a pair of dicliques may require coalescing other dicliques. These results point towards useful new data mining tools.

## Introduction

An important process in counterterrorism is the information processing of large databases that have been assembled from a variety of sources such as internet exchanges, e-mail, bank records, telephone records, credit card records, travel records, and observations of many sorts. The bits and pieces of information come from many sources and the pieces do not all tightly connect together. Some (possibly disconnected) pieces tightly connect to some other (possibly disconected) pieces. This is the nature of terrorist cell operations. Those in one cell do not know those in another cell. The person who routes money to the cells does not know or have any dealings with the one who handles and supervises the terrorists. And the one who handles them does not know or have any dealings with the one who provides safe houses or the one who provides explosives, etc.

The databases are very large. The number of records that pertain to any kind of terrorist activity is very small. The problem is to find the needle in

S. Argamon, N. Howard (eds.), *Computational Methods for Counterterrorism*, DOI 10.1007/978-3-642-01141-2_8,

the haystack. Yet despite the difficulty, the needle we are looking for involves connections between different types of relevant data. In this chapter we discuss what dicliques are and how they can be used to find related records that constitute the needle in the haystack.

## The binary relation

Our model begins with the binary relation. A connection between information piece A and information piece B is labeled. For example, IP address A and IP address B "chatted."

$$(A, B, \text{“chatted”})$$

Let $X$ be a set. A *binary relation* $R$ is a subset of $X \times X$. Let $L$ be a set of labels. A *labeled binary relation* $R$ is a subset of $X \times X \times L$.

The set $X$ can include:

Names
Addresses
Telephone Numbers
Bank Account Numbers
Bank Names
IP Address
Passport Numbers
Places

The labels can include:

Visited
Communicated with
Was seen with
Was at
Traveled to
Telephoned
Transferred money to
Received money from

For an abstract example we consider a set $X$ defined by

$$X = \{a, b, c, d, x, y, z, t, u\}$$

The binary relation $R$ on $X$ is defined by

$$R = \{(a, x), (a, y), (a, z), (b, x), (b, y), (b, z), (b, t), (b, u), (c, x), (c, y), (d, u), (d, t), (d, y)\}$$

We can represent $R$ in a shorter list form by

| | $R$ |
|---|---|
| | $R$ |
| $a$ | $x, y, z$ |
| $b$ | $x, y, z, t, u$ |
| $c$ | $x, y$ |
| $d$ | $u, t, y$ |

and we can visualize $R$ as a digraph. In the digraph the elements of $X$ are shown as nodes. A pair $(i, j)$ in $R$ is drawn as an arrow going from $i$ to $j$. This is shown in Fig. 8.1. By changing the physical location of the nodes, different drawings of the same relation are possible.

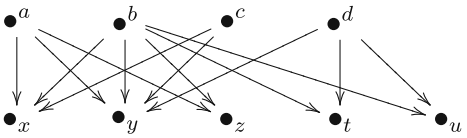



**Fig. 8.1.** Shows the digraph corresponding to the example relation $R$.

It is clear from examining the digraph that there are connections but not everything involved in the connections is connected to one another. $a, b, c, d$ have no connections to one another. $x, y, z, t, u$ have no connections to one another.

In order to form a description of what information the relation contains, we can try to reorder the digraph—redraw it as in Fig. 8.

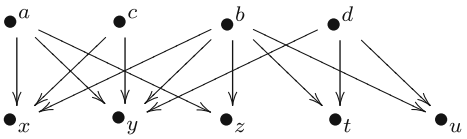



**Fig. 8.2.** Shows a reordered drawing of the digraph of Fig. 8.1.

What we want however is not a technique that is interactive, but one which is systematic and provides a description of the relation in a short and more understandable form. We want a technique that produces structures from the relation where we have in the structure itself maximally relevant information pieces. This structure is called a diclique. The diclique structure and some of its properties were first introduced by Haralick (1974).

## The diclique

A pair $(P, S)$ is called a *Diclique* of relation $R \subseteq X \times X$ if and only if

- Containment: $P \times S \subseteq R$
- Maximality: $P' \times S' \subseteq R$ and $P' \supseteq P$ and $S' \supseteq S$ imply $P = P'$ and $S = S'$

The containment condition restricts the diclique description of the relation to only include pairs actually in the relation and not include any pairs not in the relation. The maximality condition means that the diclique structure itself contains maximally relevant information pieces.

The set $\mathcal{D}$ of dicliques of the example relation $R$ are:

$$D_1 = (\{b\}, \{x, y, z, t, u\})$$
$$D_2 = (\{a, b\}, \{x, y, z\})$$
$$D_3 = (\{a, b, c\}, \{x, y\})$$
$$D_4 = (\{b, d\}, \{y, t, u\})$$
$$D_5 = (\{a, b, c, d\}, \{y\})$$

$$\mathcal{D} = \{D_1, D_2, D_3, D_4, D_5\}$$

The primary, or input, side of the diclique can be interpreted as cause, and the secondary, or output, side of the diclique can be interpreted as effect. For example consider the diclique $D_3$ of our example relation.

$$D_3 = (\{a, b, c\}, \{x, y\})$$

In the counter-terrorism application we might have:

$a$ is a funder
$b$ is a handler
$c$ is the explosive provider
$x$ is terrorist 1
$y$ is terrorist 2

But dicliques have as well other interpretations. For example they can be used to define events.

- Dicliques provide a window to events.
- Overlapping dicliques provide different views of the same event.
  - When does an event begin?
  - When does an event end?
  - What does an event include?

Dicliques can be used in document information extraction. Define the relation $R$ by

$$R = \{(word1, word2) | \text{word1 stands in relation to word2}\}$$

The concept of "stands in relation" can mean: word1 occurs within k words after word2 in j documents; or word1 is a noun subject, word2 is a noun object, of a given verb in j documents. In these applications a diclique is a word usage pattern or a meaning usage.

**Diclique cover**

Once we have the constructed the set of dicliques of a relation, we may construct from the set of dicliques a diclique cover of the relation. Let $R$ be a binary relation on $X$ and $\mathcal{D}$ be the set of dicliques of $R$. A *Diclique Cover* of $R$ is a subset $C$ of dicliques of $R$ satisfying

$$\bigcup_{(P,S)\in C} P \times S = R$$

A Diclique Cover $C$ of the example relation $R$ is $C = \{D_2, D_3, D_4\}$ where

$$D_2 = (\{a, b\}, \{x.y.z\})$$

$$D_3 = (\{a, b, c\}, \{x, y\})$$

$$D_4 = (\{b, d\}, \{y, t, u\})$$

A diclique cover $C$ of $R \subseteq X$ can be visualized by a system diagram where the rectangles represent dicliques and the connecting lines represent the elements of the set $X$. This is shown in Fig. 8.3.

The system diagram itself has multiple interpretations; many of which are centered as a cause and effect interpretation. The primary set of the diclique is the cause and the secondary set of the diclique is the effect. In the system diagram of Fig. 8.3, we have the following inferences.

Cause and Effect
$b$ works with $a$ and $c$
$x, y$, and $z$ are terrorists
$a$ and $b$ work with $x, y$, and $z$ on event 1
$a, b$, and $c$ work with $x$ and $y$ on event 2
$b$ and $d$ work with $x, u$, and $t$ on event 3
$b$ is a key person

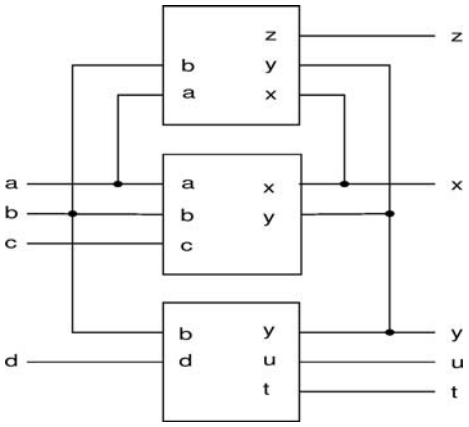



**Fig. 8.3.** Shows the set of dicliques of a relation represented by a system diagram.

## Diclique properties

In this section we summarize the basic mathematical properties of the diclique. These properties are important because they will lead us to an algorithm for computing all the dicliques of a relation. We begin with the diclique intersection theorem.

**Theorem 1** *Let $R \subseteq X \times X$. Let $(P_1, S_1)$ and $(P_2, S_2)$ be dicliques of $R$. Then*

$$\begin{array}{ll} (1) & (P_1 \cap P_2 \bigcap_{x \in (P_1 \cap P_2)} R(x)) \\ (2) & (\bigcap_{w \in (S_1 \cap S_2)} R^{-1}(y), S_1 \cap S_2) \end{array}$$

*are dicliques of $R$.*

The diclique intersection theorem leads us to see that the set of dicliques form an idempotent commutative monoid. Actually there are two idempotent commutative monoids formed. One on the basis of the input or primary sets and one on the basis of the output sets or secondary sets.

**Theorem 2** *Let $R \subseteq X \times X$ and let $\mathcal{D}$ be the set of dicliques of $R$. Let $(P_1, S_1) \in \mathcal{D}$ and $(P_2, S_2) \in \mathcal{D}$. Define the operation $\circ$ on $\mathcal{D}$ by*

$$(P_1, S_1) \circ (P_2, S_2) = (P_1 \cap P_2, \bigcup_{x \in P_1 \cap P_2} R(x))$$

*Then $(\mathcal{D}, \circ)$ is an idempotent commutative monoid.*

**Theorem 3** *Let $R \subseteq X \times X$ and let $\mathcal{D}$ be the set of dicliques of $R$. Let $(P_1, S_1) \in \mathcal{D}$ and $(P_2, S_2) \in \mathcal{D}$. Define the operation $\circ$ on $\mathcal{D}$ by*

$$(P_1, S_1) \circ (P_2, S_2) = (\bigcup_{w \in S_1 \cap S_2} R^{-1}(w), S_1 \cap S_2)$$

*Then* $(\mathcal{D}, \circ)$ *is an idempotent commutative monoid.*

The algebraic structure of the set of dicliques is more than just monoids. The set of dicliques have a partial ordering.

Let $(A, B)$ and $(C, D)$ be dicliques of $R$. Define the $\leq$ relation as

$$(A, B) \leq (C, D) \text{ if and only if } A \subseteq C$$

**Theorem 4** *Let* $\mathcal{D}$ *be the set of all dicliques of* $R$*. Then* $(\mathcal{D}, \leq)$ *is a partially ordered set.*

This partial ordering together with the two idempotent commutative monoids makes the partial ordering a lattice. Thus the set of dicliques forms a lattice.

**Theorem 5** $(\mathcal{D}, \leq)$ *is a lattice with the meet* $\wedge$ *and join* $\vee$ *operators defined by*

$$(A, B) \wedge (C, D) = (A \cap C, \bigcap_{x \in A \cap C} R(x))$$

$$(A, B) \vee (C, D) = (\bigcap_{y \in C \cap D} R^{-1}(y), B \cap D)$$

With these definitions, we see that if $\bigcap_{x \in X} R(x) = \emptyset$, then $(X, \emptyset)$ is a diclique, and if $\bigcap_{y \in X} R^{-1}(y) = \emptyset$, then $(\emptyset, X)$ is a diclique.

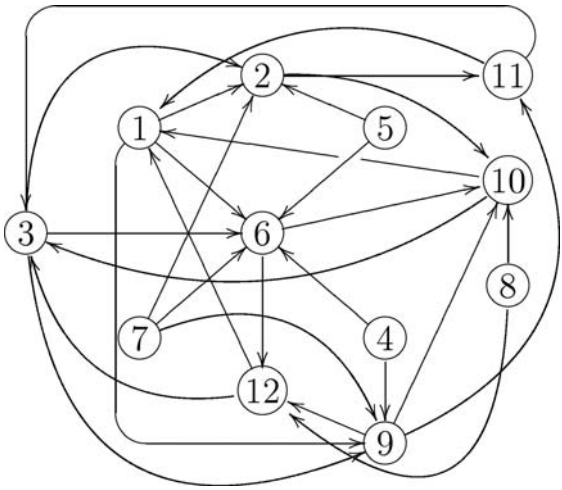



**Fig. 8.4.** Shows a digraph visualization of a more complicated relation.

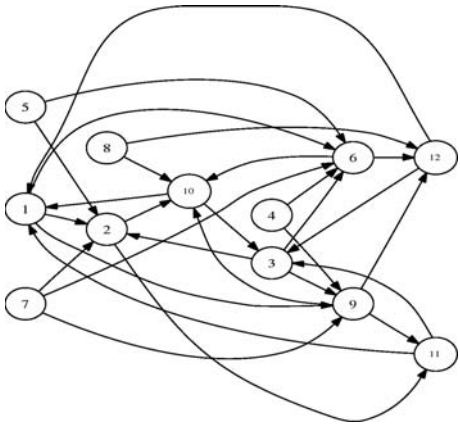



**Fig. 8.5.** Shows a different digraph visualization of a more complicated relation.

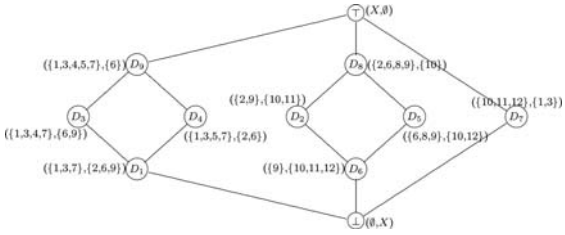



**Fig. 8.6.** Shows the lattice of the dicliques of a more complicated relation.

## Computing dicliques

In this section we summarize some theorems which establish the algorithm for constructing all the dicliques of a relation.

**Theorem 6** *If* $(A, B)$ *is a diclique of* $R$*, then*

$$A = \bigcap_{b \in B} R^{-1}(b)$$

$$B = \bigcap_{a \in A} R(a)$$

Indeed this idea of starting with the input or primary set of the dicique and from it determining the output or secondary set and vica-versa can be generalized. We can start with any subset as the input set. From that subset determine the corresponding output set and from the output set determine a new enlarged input set. Once we have weaved our way through the relation

in the forward sense and then again in the backward sense, there is no need to continue the process. We have reached a fixed point.

**Theorem 7** *Let $A \subseteq X$ be given. Define $B, A'$, and $B'$ by*

$$B = \bigcap_{a \in A} R(a)$$

$$A' = \bigcap_{b \in B} R^{-1}(b)$$

$$B' = \bigcap_{a \in A'} R(a)$$

*Then $B = B'$.*

This fixed point theorem has a dual by starting with the output sets and weaving through the relation to the input set and then again to the output set.

**Theorem 8** *Let $B \subseteq Y$ be given. Define $A, B'$, and $A'$ by*

$$A = \bigcap_{b \in B} R^{-1}(b)$$

$$B' = \bigcap_{a \in A} R(a)$$

$$A' = \bigcap_{b \in B'} R^{-1}(b)$$

*Then $A = A'$.*

The fixed point theorems lead us to the diclique representation theorem.

**Theorem 9** *$(P, S)$ is a diclique of $R \subseteq X \times X$ if and only if for some $A \subseteq X$,*

$$S = \bigcap_{x \in A} R(x)$$

$$P = \bigcap_{y \in S} R^{-1}(y)$$

The diclique representation theorem leads us to the diclique finding algorithm.

(i) $n = 0$, $T_n = \{R(x) | x \in X\}$
(ii) Repeat until no change
(iii) Compute $Q$ the set of all possible intersections between pairs of sets in $T_n$
(iv) $T_{n+1} = T_n \cup Q$
(v) At fixed point $\mathcal{D} = \{(\bigcup_{y \in S} R^{-1}(y), S), S \in T_n\}$

## Coalescing dicliques

The data collection and aggregation forming the initial relation $R$ may not be complete. This is probably the case most of the time—some facts do not get gathered. The relation $R$ we have at hand may miss some $(x, y)$ pairs. It is possible by working with the dicliques to form hypotheses about which $(x, y)$ pairs may be missing. Such hypotheses are of course an enormous benefit to the operational personnel as to where to specifically look or gather additional information to confirm the hypothesis. The technique used to form such hypotheses is called coalescing of dicliques.

Dicliques eligible for coalescing are those that have corresponding primary and secondary sets with large overlap. For example consider the following pair of dicliques.

$$D_1 = \{\{1, 3, 5\}, \{2, 4, 8\}\}$$
$$D_2 = \{\{3, 5, 6\}, \{4, 7, 8\}\}$$

Neglecting whatever else might be in the relation from which these dicliques come, if these dicliques were to be coalesced the new coalesced diclique would be $D = \{\{1, 3, 5, 6\}, \{2, 4, 7, 8\}\}$. And this would indeed be a diclique of a new relation based on the original relation but including the added pairs $(1, 7)$ and $(6, 2)$.

Now adding pairs to the relation to effect a coalescing of some dicliques affects other dicliques. To maintain the consistency between the augmented relation and the dicliques, we must determine the effect of pairs added to relation for all dicliques. Our tool for doing this is based on groupoid homomorphisms.

Let $(G, \circ)$ and $(H, *)$ be groupoids. A function $F : G \rightarrow H$ is a *Homomorphism* from $(G, \circ)$ to $(H, *)$ if and only if

$$x, y \in G \text{ implies } f(x \circ y) = f(x) * f(y)$$

We will be wanting to do everything in the framework of groupoid homomorphisms. To do so we will need a characterization of the coalescing that takes place through the homomorphism. This characterization is based on what has historically been called the substitution property partition.

Let $(G, \circ)$ be a groupoid and $\Pi = \{\pi_k\}_{k=1}^{K}$ be a partition over $G$. $\Pi$ is called a *Substitution Property Partition* if and only if for every $\pi_i$ and $\pi_j$, there exits a $\pi_k$ such that

$$x \in \pi_i \text{ and } y \in \pi_j \text{ implies } x \circ y \in \pi_k$$

Groupoid homomorphisms create substitution property partitions as stated in the following theorems.

**Theorem 10** *Let $(G, \circ)$ and $(H, *)$ be groupoids and $f : G \to H$ a homomorphism from $(G, \circ)$ to $(H, *)$. Then*

$$\Pi = \{f^{-1}(h) | h \in H\}$$

*is a substitution property partition.*

**Theorem 11** *Let $(G, \circ)$ be a groupoid and $\Pi$ be a substitution property partition on $G$. Then $(\Pi, *)$ is a groupoid where the multiplication $*$ is defined by*

$$\pi_i * \pi_j = \pi_k$$

*where if $x \in \pi_i$ and $y \in \pi_j$, $x \circ y \in \pi_k$.*

**Theorem 12** *Let $(G, \circ)$ be a groupoid and $\Pi$ be a substitution property partition on $G$. Then $f : G \to \Pi$ is a homomorphism from $(G, \circ)$ to $(H, *)$.*

**Theorem 13** *Let $(G, \circ)$ be a groupoid and $\Pi$ be a partition over $G$. $\Pi$ is a substitution property partition if and only if $x, y \in P \in \Pi$ implies for every $z \in G$*

*(i) there exists $Q \in \Pi$ such that $x \circ z, y \circ z \in Q$*
*(ii) there exists $Q' \in \Pi$ such that $z \circ x, z \circ y \in Q'$*

These theorems lead us to the coalescing algorithm in which the algorithm constructs the largest homomorphic image of a commutative groupoid $G$ having elements $a$ and $b$ coalesced.

(i) Set up a partition of $G$ in which each cell has one member except for the cell containing both $a$ and $b$
(ii) In the multiplication table replace all references of $b$ by $a$
(iii) If there exist two columns labeled the same, go to (4)
(iv) If for any row of these same labeled columns, the pair of entries is not in the same cell, coalesce the cells and in the multiplication table replace all references of the second cell by references of the first cell. Continue doing this until one pair of these same labeled columns are identical
(v) Delete one column of the pairs of identical same labeled columns. Delete the corresponding row. Go to (3)

Now, coalesce E and H by adding (2,12) to the relation $R$.

$$B = (\{2, 9\}, \{10, 11\}) \subseteq (\{2, 6, 8, 9\}, \{10, 11, 12\}) = NewE$$

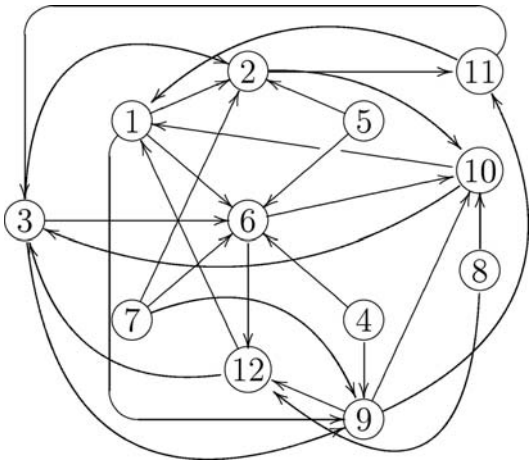



**Fig. 8.7.** The example relation $R$.

**Table 8.1.** The dicliques for the relation $R$ of Fig. 8.

| Diclique Label | Primary Set | Secondary Set |
|---|---|---|
| A | 1,3,7 | 2,6,9 |
| B | 2,9 | 10,11 |
| C | 1,3,4,7 | 6,9 |
| D | 1,3,5,7 | 2,6 |
| E | 6,8,9 | 10,12 |
| F | 9 | 10,11,12 |
| G | 10,11,12 | 1,3 |
| H | 2,6,8,9 | 10 |
| I | 1,3,4,5,7 | 6 |
| X | $\emptyset$ | X |
| $\emptyset$ | X | $\emptyset$ |

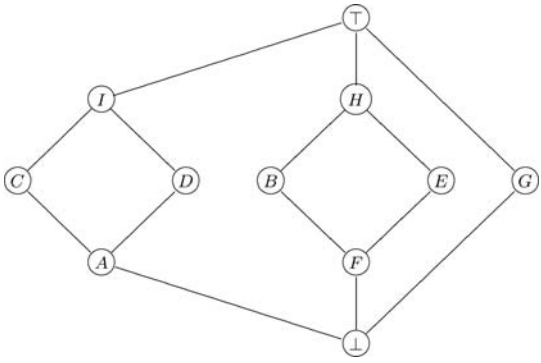



**Fig. 8.8.** The diclique lattice for the dicliques of the relation $R$ of Fig. 8.

**Table 8.2.** Multiplication table of commutative groupoid for the relation R.

| | A | B | C | D | E | F | G | H | I | X | ∅ |
|---|---|---|---|---|---|---|---|---|---|---|---|
| A | A | ∅ | C | D | ∅ | ∅ | ∅ | ∅ | I | A | ∅ |
| B | ∅ | B | ∅ | ∅ | H | B | ∅ | H | ∅ | B | ∅ |
| C | C | ∅ | C | I | ∅ | ∅ | ∅ | ∅ | I | C | ∅ |
| D | D | ∅ | I | D | ∅ | ∅ | ∅ | ∅ | I | D | ∅ |
| E | ∅ | H | ∅ | ∅ | E | E | ∅ | H | ∅ | E | ∅ |
| F | ∅ | B | ∅ | ∅ | E | F | ∅ | H | ∅ | F | ∅ |
| G | ∅ | ∅ | ∅ | ∅ | ∅ | ∅ | G | ∅ | ∅ | G | ∅ |
| H | ∅ | H | ∅ | ∅ | H | H | ∅ | H | ∅ | H | ∅ |
| I | I | ∅ | I | I | ∅ | ∅ | ∅ | ∅ | I | I | ∅ |
| X | A | B | C | D | E | F | G | H | I | X | ∅ |
| ∅ | ∅ | ∅ | ∅ | ∅ | ∅ | ∅ | ∅ | ∅ | ∅ | ∅ | ∅ |

**Table 8.3.** Coalescing dicliques C and D corresponds to adding (4,2) and (5,9) to the relation *R*.

| Diclique Label | Primary Set | Secondary Set |
|---|---|---|
| C | 1,3,4,7 | 6,9 |
| D | 1,3,5,7 | 2,6 |
| New C | 1,3,4,5,7 | 2,6,9 |

**Table 8.4.** Coalescing Algorithm: C and D overlap so replace D by C.

| | A | B | C | C | E | F | G | H | I | X | ∅ |
|---|---|---|---|---|---|---|---|---|---|---|---|
| A | A | ∅ | C | C | ∅ | ∅ | ∅ | ∅ | I | A | ∅ |
| B | ∅ | B | ∅ | ∅ | H | B | ∅ | H | ∅ | B | ∅ |
| C | C | ∅ | C | I | ∅ | ∅ | ∅ | ∅ | I | C | ∅ |
| C | C | ∅ | I | C | ∅ | ∅ | ∅ | ∅ | I | C | ∅ |
| E | ∅ | H | ∅ | ∅ | E | E | ∅ | H | ∅ | E | ∅ |
| F | ∅ | B | ∅ | ∅ | E | F | ∅ | H | ∅ | F | ∅ |
| G | ∅ | ∅ | ∅ | ∅ | ∅ | ∅ | G | ∅ | ∅ | G | ∅ |
| H | ∅ | H | ∅ | ∅ | H | H | ∅ | H | ∅ | H | ∅ |
| I | I | ∅ | I | I | ∅ | ∅ | ∅ | ∅ | I | I | ∅ |
| X | A | B | C | C | E | F | G | H | I | X | ∅ |
| ∅ | ∅ | ∅ | ∅ | ∅ | ∅ | ∅ | ∅ | ∅ | ∅ | ∅ | ∅ |

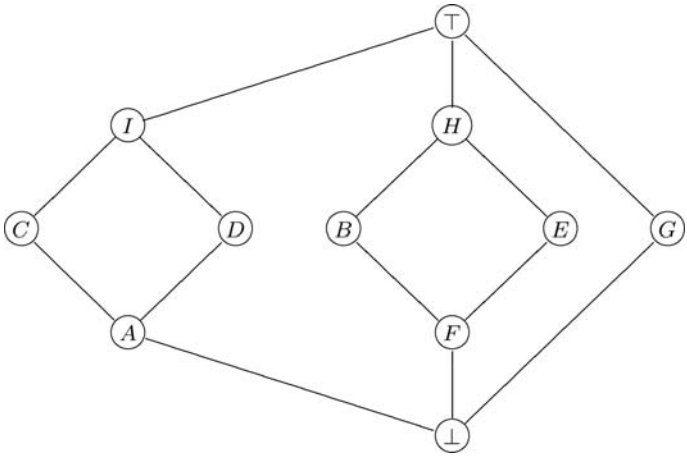



**Fig. 8.9.** The new diclique lattice after the merging of dicliques C and D.

**Table 8.5.** By the merging of C and D we are now forced to merge I and C. Replace I by C.

| | A | B | C | C | E | F | G | H | C | X | ∅ |
|---|---|---|---|---|---|---|---|---|---|---|---|
| A | A | ∅ | C | C | ∅ | ∅ | ∅ | ∅ | C | A | ∅ |
| B | ∅ | B | ∅ | ∅ | H | B | ∅ | H | ∅ | B | ∅ |
| C | C | ∅ | C | C | ∅ | ∅ | ∅ | ∅ | C | C | ∅ |
| C | C | ∅ | C | C | ∅ | ∅ | ∅ | ∅ | C | C | ∅ |
| E | ∅ | H | ∅ | ∅ | E | E | ∅ | H | ∅ | E | ∅ |
| F | ∅ | B | ∅ | ∅ | E | F | ∅ | H | ∅ | F | ∅ |
| G | ∅ | ∅ | ∅ | ∅ | ∅ | ∅ | G | ∅ | ∅ | G | ∅ |
| H | ∅ | H | ∅ | ∅ | H | H | ∅ | H | ∅ | H | ∅ |
| C | C | ∅ | C | C | ∅ | ∅ | ∅ | ∅ | C | C | ∅ |
| X | A | B | C | C | E | F | G | H | C | X | ∅ |
| ∅ | ∅ | ∅ | ∅ | ∅ | ∅ | ∅ | ∅ | ∅ | ∅ | ∅ | ∅ |

**Table 8.6.** By the merging of C and D we are now forced to merge I and C. Replace I by C.

| | A | B | C | E | F | G | H | X | ∅ |
|---|---|---|---|---|---|---|---|---|---|
| A | A | ∅ | C | ∅ | ∅ | ∅ | ∅ | A | ∅ |
| B | ∅ | B | ∅ | H | B | ∅ | H | B | ∅ |
| C | C | ∅ | C | ∅ | ∅ | ∅ | ∅ | C | ∅ |
| E | ∅ | H | ∅ | E | E | ∅ | H | E | ∅ |
| F | ∅ | B | ∅ | E | F | ∅ | H | F | ∅ |
| G | ∅ | ∅ | ∅ | ∅ | ∅ | G | ∅ | G | ∅ |
| H | ∅ | H | ∅ | H | H | ∅ | H | H | ∅ |
| X | A | B | C | E | F | G | H | X | ∅ |
| ∅ | ∅ | ∅ | ∅ | ∅ | ∅ | ∅ | ∅ | ∅ | ∅ |

**Table 8.7.** Replace diclique C by A.

| | A | B | A | E | F | G | H | X | ∅ |
|---|---|---|---|---|---|---|---|---|---|
| A | A | ∅ | A | ∅ | ∅ | ∅ | ∅ | A | ∅ |
| B | ∅ | B | ∅ | H | B | ∅ | H | B | ∅ |
| A | A | ∅ | A | ∅ | ∅ | ∅ | ∅ | A | ∅ |
| E | ∅ | H | ∅ | E | E | ∅ | H | E | ∅ |
| F | ∅ | B | ∅ | E | F | ∅ | H | F | ∅ |
| G | ∅ | ∅ | ∅ | ∅ | ∅ | G | ∅ | G | ∅ |
| H | ∅ | H | ∅ | H | H | ∅ | H | H | ∅ |
| X | A | B | A | E | F | G | H | X | ∅ |
| ∅ | ∅ | ∅ | ∅ | ∅ | ∅ | ∅ | ∅ | ∅ | ∅ |

**Table 8.8.** The coalesced table.

| | A | B | E | F | G | H | X | ∅ |
|---|---|---|---|---|---|---|---|---|
| A | A | ∅ | ∅ | ∅ | ∅ | ∅ | A | ∅ |
| B | ∅ | B | H | B | ∅ | H | B | ∅ |
| A | A | ∅ | ∅ | ∅ | ∅ | ∅ | A | ∅ |
| E | ∅ | H | E | E | ∅ | H | E | ∅ |
| F | ∅ | B | E | F | ∅ | H | F | ∅ |
| G | ∅ | ∅ | ∅ | ∅ | G | ∅ | G | ∅ |
| H | ∅ | H | H | H | ∅ | H | H | ∅ |
| X | A | B | E | F | G | H | X | ∅ |
| ∅ | ∅ | ∅ | ∅ | ∅ | ∅ | ∅ | ∅ | ∅ |

**Table 8.9.** Showing the overlap between dicliques E and H which suggests that they should be coalesced.

| Diclique Label | Primary Set | Secondary Set |
|---|---|---|
| E | 6,8,9 | 10,12 |
| H | 2,6,8,9 | 10 |
| New E | 2,6,8,9 | 10,12 |

**Table 8.10.** Replace diclique H by E.

| | A | B | E | F | G | E | X | ∅ |
|---|---|---|---|---|---|---|---|---|
| A | A | ∅ | ∅ | ∅ | ∅ | ∅ | A | ∅ |
| B | ∅ | B | E | B | ∅ | E | B | ∅ |
| A | A | ∅ | ∅ | ∅ | ∅ | ∅ | A | ∅ |
| E | ∅ | E | E | E | ∅ | E | E | ∅ |
| F | ∅ | B | E | F | ∅ | E | F | ∅ |
| G | ∅ | ∅ | ∅ | ∅ | G | ∅ | G | ∅ |
| E | ∅ | E | E | E | ∅ | E | E | ∅ |
| X | A | B | E | F | G | E | X | ∅ |
| ∅ | ∅ | ∅ | ∅ | ∅ | ∅ | ∅ | ∅ | ∅ |

**Table 8.11.** The coalesced table.

| | A | B | E | F | G | X | ∅ |
|---|---|---|---|---|---|---|---|
| A | A | ∅ | ∅ | ∅ | ∅ | A | ∅ |
| B | ∅ | B | E | B | ∅ | B | ∅ |
| A | A | ∅ | ∅ | ∅ | ∅ | A | ∅ |
| E | ∅ | E | E | E | ∅ | E | ∅ |
| F | ∅ | B | E | F | ∅ | F | ∅ |
| G | ∅ | ∅ | ∅ | ∅ | G | G | ∅ |
| E | ∅ | E | E | E | ∅ | E | ∅ |
| X | A | B | E | F | G | X | ∅ |
| ∅ | ∅ | ∅ | ∅ | ∅ | ∅ | ∅ | ∅ |

**Table 8.12.** Coalesce E and F by adding (6,11), (8,11).

| Diclique Label | Primary Set | Secondary Set |
|---|---|---|
| B | 2,9 | 10,11 |
| New E | 2,6,8,9 | 10,12 |
| F | 9 | 10,11,12 |
| New E | 2,6,8,9 | 10,12 |

**Table 8.13.** Replace F by E.

| | A | B | E | E | G | X | ∅ |
|---|---|---|---|---|---|---|---|
| A | A | ∅ | ∅ | ∅ | ∅ | A | ∅ |
| B | ∅ | B | E | B | ∅ | B | ∅ |
| A | A | ∅ | ∅ | ∅ | ∅ | A | ∅ |
| E | ∅ | E | E | E | ∅ | E | ∅ |
| E | ∅ | B | E | E | ∅ | E | ∅ |
| G | ∅ | ∅ | ∅ | ∅ | G | G | ∅ |
| E | ∅ | E | E | E | ∅ | E | ∅ |
| X | A | B | E | E | G | X | ∅ |
| ∅ | ∅ | ∅ | ∅ | ∅ | ∅ | ∅ | ∅ |

**Table 8.14.** The overlap between B and the new E suggests that they be coalesced.

| Diclique Label | Primary Set | Secondary Set |
|---|---|---|
| B | 2,9 | 10,11 |
| New E | 2,6,8,9 | 10,11,12 |

**Table 8.15.** Replace E by B.

| | A | B | B | B | G | X | ∅ |
|---|---|---|---|---|---|---|---|
| A | A | ∅ | ∅ | ∅ | ∅ | A | ∅ |
| B | ∅ | B | B | B | ∅ | B | ∅ |
| A | A | ∅ | ∅ | ∅ | ∅ | A | ∅ |
| B | ∅ | B | B | B | ∅ | B | ∅ |
| B | ∅ | B | B | B | ∅ | B | ∅ |
| G | ∅ | ∅ | ∅ | ∅ | G | G | ∅ |
| B | ∅ | B | B | B | ∅ | B | ∅ |
| X | A | B | B | B | G | X | ∅ |
| ∅ | ∅ | ∅ | ∅ | ∅ | ∅ | ∅ | ∅ |

**Table 8.16.** Coalesced

| | A | B | G | X | ∅ |
|---|---|---|---|---|---|
| A | A | ∅ | ∅ | A | ∅ |
| B | ∅ | B | ∅ | B | ∅ |
| A | A | ∅ | ∅ | A | ∅ |
| B | ∅ | B | ∅ | B | ∅ |
| B | ∅ | B | ∅ | B | ∅ |
| G | ∅ | ∅ | G | G | ∅ |
| B | ∅ | B | ∅ | B | ∅ |
| X | A | B | G | X | ∅ |
| ∅ | ∅ | ∅ | ∅ | ∅ | ∅ |

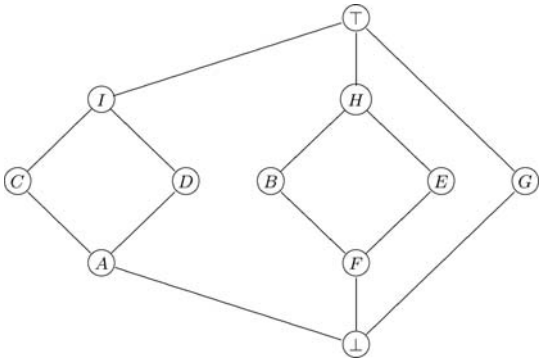



**Fig. 8.10.** Diclique lattice

## Conclusion

In counterterrorist applications, databases of possible connections and activities are very large and the number of records that pertain to terrorist activity is very small. The problem is to find (and assemble) a needle in a haystack made up of connections between different types of relevant data. The bits and pieces of information come from many sources and the pieces do not all tightly connect together. Some (possibly disconnected) pieces tightly connect to some other (possibly disconnected) pieces.

To address this problem, we have developed the concept of diclique and have shown how the dicliques of a relation maximally aggregate relation information pieces together. We have summarized the mathematical properties

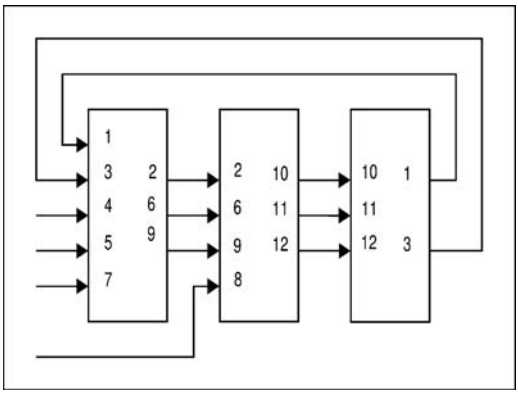



**Fig. 8.11.** In the system diagram *rectangles* correspond to dicliques; *lines* correspond to elements of $X$.

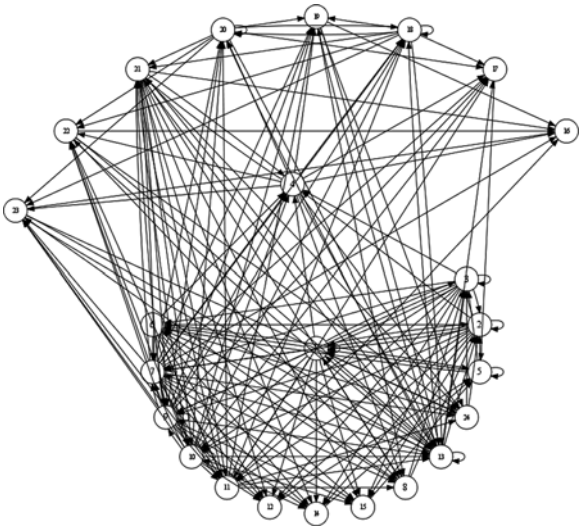


**Fig. 8.12.** The graph of a more complex relation.

of dicliques showing that the set of dicliques forms a lattice and in two different ways forms commutative idempotent monoids. Finally, by investigating groupoid homomorphisms on the monoid structure, we have shown how to coalesce overlaping dicliques and how coalesing a pair of dicliques may require coalescing other dicliques. We have stated the theorems governing these forced coalescings and have demonstrated all of this in an example.

Future work will be in the use of time local relations where we believe that the dicliques correspond to events. In general relations, dicliques correspond to subsystems. We hope to do future research by developing a formal framework

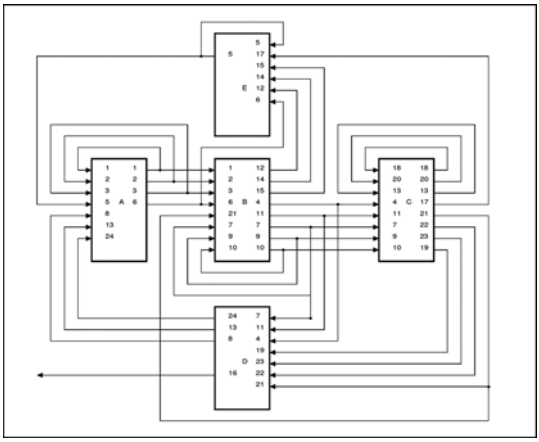


**Fig. 8.13.** The dicliques of the more complex relation represented as a system diagram.

for events and subsystems. This formal framework would then permit us to algorithmically infer the event or subsystem from the observations.

The work that we have done for simple unlabeled binary relations needs to be extended to labeled binary relations and needs to be extended to N-ary relations.

Although in practical applications in counterterrorism, it is more likely to have pairs of the relation missing then extraneous pairs in the relation, we would nevertheless like to have a theory for the removal of pairs in the relation that would simplify the diclique description of the relation. Finally, work needs to be done on developing fast algorithms for finding dicliques.

## References

Haralick, R. 1974. The diclique representation and decomposition of binary relations. *Journal of the Association for Computing Machinery* (July) 356–366.

# 9

# Information Superiority via Formal Concept Analysis

Bjoern Koester and Stefan E. Schmidt

**Summary.** This chapter will show how to get more mileage out of information. To achieve that, we first start with an introduction to the fundamentals of Formal Concept Analysis (FCA). FCA is a highly versatile field of applied lattice theory, which allows hidden relationships to be uncovered in relational data. Moreover, FCA provides a distinguished supporting framework to subsequently find and fill information gaps in a systematic and rigorous way. In addition, we would like to build bridges via a universal approach to other communities which can be related to FCA in order for other research areas to benefit from a theory that has been elaborated for more than twenty years. Last but not least, the essential benefits of FCA will be presented algorithmically as well as theoretically by investigating a real data set from the MIPT Terrorism Knowledge Base and also by demonstrating an application in the field of Web Information Retrieval and Web Intelligence.

## Introduction

This work gives the reader an introduction into Formal Concept Analysis (FCA) and also an overview via applications dealing with real data. Beyond a standard introduction into FCA, this chapter provides a category-theoretical frame of reference for FCA. More explicitly, a novel substantiation of FCA via the category of ordered sets and adjunctions between ordered sets as morphisms is presented. Within this set-up, a construction of the *diagonal* of an adjunction is given and, subsequently, a decomposition theorem for adjunctions is outlined. It says that every adjunction can be decomposed via the diagonal into an epimorphism followed by a monomorphism. The close connection between adjunctions and Galois connections is then exploited in order to utilize FCA. In particular, an invariant formulation of a concept lattice as the Galois diagonal of its induced Galois connection (between the power set lattice of the objects set and the power set lattice of the attribute set of a formal context) are given. Also, it is pointed out how the decomposition theorem for adjunctions yields the basic theorem of Formal Concept Analysis. The remaining sections focus on applications of FCA.

S. Argamon, N. Howard (eds.), *Computational Methods for Counterterrorism*, DOI 10.1007/978-3-642-01141-2_9,

Networks play an important role in uncovering and understanding relationships of human societies on various levels. In the domain of security analysis and defense studies, hierarchical aspects of networks are of crucial significance; they can be formalized via the mathematical concepts of ordered sets and lattices. Jonathan Farley, for instance, has shown how to break terrorism cells using applied lattice theory (Farley 2003). Voss and Joslyn have given an overview of ways to find relationships between people, groups, terrorist events, and areas of expertise using graph theoretic aspects as well as applied lattice theory (Voss and Joslyn). Within the demonstrative part of this chapter, we also show step-by-step how to uncover hidden relationships in data given by statistics of the MIPT Terrorism Knowledge Base using applied lattice theory. Rome and Haralick have shown how social communities can be found in a Web graph built of hyperlinks between Web pages (Rome and Haralick 2005). The Web will also play a role in the demonstrative part of our current work; we present an approach using the data contained in Web search engine results to gain new information, and hence new knowledge.

## Formal concept analysis: An introduction

Formal Concept Analysis (FCA) deals with mathematical methods to develop and analyze relationships between concepts and its hierarchical implications on given data. FCA studies how objects can be hierarchically grouped together according to their common attributes. FCA is based on a mathematization of the philosophical understanding of a concept. For the formalization of concept, an algebraic theory of binary relations and complete lattices is used. Within Computer Science, FCA is increasingly applied in Conceptual Clustering, Data Analysis, Information Retrieval, Knowledge Discovery, and Ontology Engineering (cf. Priss 2006).

The main idea of FCA can be outlined as follows: data is represented in a very basic data type which is called a formal context (a 'cross table'). Each formal context is then transformed into a mathematical structure called a concept lattice (a 'concept hierarchy'). For comprehensive analysis, rather than a formal context, the appropriate concept lattice is used. A concept lattice can then—without loss of information—be transformed back into the original formal context and vice versa. In some cases it might be appropriate to represent the concept lattice graphically to support communication, or it may be also be investigated using algebraic methods to unravel its structure. Readers interested in a deeper insight into FCA are referred to Ganter and Wille (1999). In this section, we would like to introduce the basic notions of Formal Concept Analysis along with examples. But first, we need some mathematical foundations.

**Ordered sets**

A *binary relation* $R$ on two sets $M$ and $N$ is a subset $R \subseteq M \times N$. Then, $(m, n) \in R$ means that '$m$ is related to $n$ under $R$' and is often written as $mRn$. The binary relation $R$ contains all pairs of elements that are related via $R$. If $R \subseteq M \times M$, we simply say that $R$ is a binary relation on $M$. A binary relation $R$ is called an *order relation* on a set $M$ if for all elements $x, y, z \in M$ the following hold:

$$xRx \qquad \text{(reflexivity)} \tag{9.1}$$

$$xRy,\ yRz \implies xRz \qquad \text{(transitivity)} \tag{9.2}$$

$$xRy,\ yRx \implies x = y \qquad \text{(anti-symmetry)} \tag{9.3}$$

An order relation is usually denoted by the symbol $\leq$. In this case, $x \leq y$ is read as '$x$ is less than or equal to $y$.' For $x \leq y$ we also write $y \geq x$ and say $y$ is greater or equal to $x$. If $x \leq y$ and $x \neq y$ hold, we will write $x < y$ or $y > x$ and say that $x$ is less than $y$ or that $y$ is greater than $x$.

An *ordered set* (or partially ordered set or poset) is a pair $\mathbb{M} = (M, \leq)$ consisting of a set $M$ and an order relation $\leq$ on $M$. An ordered set will be called *finite* if and only if the underlying set is finite. We will focus in this chapter on finite ordered sets. An element $a$ is a *lower neighbor* of an element $b$, denoted by $a \prec b$, if $a$ is less than $b$ and there exists no element which is greater than $a$ and less than $b$. The relational structure $(M, \prec)$ is called the *Hasse structure* of $\mathbb{M}$.

According to Schröder (2003), lattices are the most common ordered structures in mathematics. If $\mathbb{M} = (M, \leq)$ is an ordered set and $A$ is a subset of $M$, then a *lower bound* of $A$ is an element $s$ of $M$ with $s \leq a$ for all $a \in A$. An *upper bound* of $A$ is defined dually. If there is a largest element in the set of all lower bounds of $A$, it is called the *infimum* of $A$ and is denoted by $\inf(A)$. Dually, a least upper bound is called *supremum* and is denoted by $\sup(A)$.

Also, the infimum (supremum) of a family $(a_i)_{i \in I}$ in $\mathbb{M}$ is defined as the infimum (supremum) of the set $\{a_i \mid i \in I\}$. An ordered set $\mathbb{L} = (L, \leq)$ is a *lattice* if for any two of its elements the supremum and the infimum exist. $L$ is called a *complete lattice* if for each subset of $L$ the supremum and the infimum exist. Every complete lattice $L$ has a largest element, denoted by $\mathbf{1}_V$, and a smallest element $\mathbf{0}_L$. A subset $A$ of $L$ is called *supremum-dense* in $\mathbb{L}$ if for every $x \in L$ one has $x = \sup\{a \in A \mid a \leq x\}$. The dual of supremum-dense is *infimum-dense*.

Every finite ordered set $\mathbb{M} = (M, \leq)$ can be represented as a *line diagram* (i.e., a geometric representation of its Hasse structure) as shown in the later examples.

### Example data

The following example data is taken from the U.S. Customs and Border Protection. They list Border Patrol-wide gaps in fifteen areas (Department of Homeland Security 2004). For this section, we selected the following three gap areas, namely 'Biometric,' 'Training,' and 'Vehicles,' along with the following operational functions:

O-1 Detect, prevent, and characterize illegal breaches to the U.S. border
O-2 Respond to and apprehend individuals and items of interest
O-3 Process individuals in custody and items as accurately, efficiently, and effectively as possible
O-4 Deploy Agents using effective decision support systems and solutions
O-5 Assess threats and determine likely illegal border crossing scenarios, methods, and locations

These operational functions were given in a matrix by the U.S. Customs and Border Protection to correlate the gaps. Please note that we have selected only a few of the operational functions mentioned and gaps in this example as shown in Table 9.1.

**Table 9.1.** An extract of a given matrix by the U.S. Customs and Border Protection which lists gap correlations to operational functions.

| Gap Correlations | Operational Functions |
|---|---|
| Biometric | O-2, O-3, O-5 |
| Training | O-1, O-2, O-3 |
| Vehicles | O-1, O-2, O-3, O-4 |

### Formal context

A *formal context* $\mathbb{K} := (G, M, I)$ is composed of a set $G$, the elements of which are called objects, a set $M$, the elements of which are called attributes, and a binary relation $I \subseteq G \times M$. We call $I$ the *incidence relation* and read $(g, m) \in I$ as 'the object g has the attribute m.'

**Table 9.2.** Cross Table for selected sets of 'Gap Correlation to Operational Functions.'

| Gap Correlation | O-1 | O-2 | O-3 | O-4 | O-5 |
|---|---|---|---|---|---|
| Biometric | | × | × | | × |
| Training | × | × | × | | |
| Vehicles | × | × | × | × | |

A formal context $\mathbb{K}$ can be visualized as a two dimensional table, a *cross table*, as shown in Table 9.2. A cross table, as the name suggests, uses crosses to indicate the incidence relation.

### Formal concept

The polarity described by a Galois connection is one of the principal ingredients of FCA. Such a duality can be observed in relations between two ordered sets. When a given set is increased, the correlated set decreases in size, and vice versa. To define this behavior, we introduce the following derivation operators: For a subset $A \subseteq G$ of the objects we define the set of attributes common to the objects in $A$ as

$$A' := \{m \in M \mid gIm \ for\ all\ g \in A\} \tag{9.4}$$

respectively. For a subset $B \subseteq M$ of the attributes we define a set of objects which have all attributes in $B$ as

$$B' := \{g \in G \mid gIm \ for\ all\ m \in B\} \tag{9.5}$$

The pair of *derivation operators* form a Galois connection. This means the following statements are true for a given formal context $(G, M, I)$, subsets $A, A_1, A_2 \subseteq G$ of objects and subsets $B, B_1, B_2 \subseteq M$ of attributes:

$$A_1 \subseteq A_2 \Rightarrow A_2' \subseteq A_1' \quad \text{and} \quad B_1 \subseteq B_2 \Rightarrow B_2' \subseteq B_1' \tag{9.6}$$

$$A \subseteq A'' \quad \text{and} \quad B \subseteq B'' \tag{9.7}$$

$$A' = A''' \quad \text{and} \quad B' = B''' \tag{9.8}$$

$$A \subseteq B' \Leftrightarrow B \subseteq A' \quad \Leftrightarrow \quad A \times B \subseteq I \tag{9.9}$$

Both the derivation operators and the formed Galois connection now allow us to define a formal concept. In a philosophical sense, a concept is made up of two parts: the extension and the intension. The extension covers all objects belonging to this concept and the intension comprises all attributes valid for all those objects.

A *formal concept* of the corresponding formal context $\mathbb{K} := (G, M, I)$ is composed of a pair $(A, B)$ such that $A \subseteq G$, $B \subseteq M$, and

$$A' = B \text{ and } B' = A \tag{9.10}$$

$A$ is called *extent* and $B$ *intent* of the concept $(A, B)$. We denote the set of all concepts of the formal context $(G, M, I)$ as $\mathfrak{B}(G, M, I)$ and write $\mathfrak{B}(\mathbb{K})$ for short.

In general, all subsets of $G$ do not form an extent, and nor do all subsets of $M$ form an intent of some concept. Subsets $A \subseteq G$ and $B \subseteq M$ are an extent and an intent of a concept if and only if $A'' = A$ and $B'' = B$.

From a geometrical perspective, the concepts of a formal context can be interpreted as maximal rectangles filled with crosses in the cross table. Equation (9.10) implies a maximal rectangle $A \times B$ in the cross table of the incidence relation.

**From formal contexts to formal concepts**

The number of concepts in a concept lattice depends on the relation in the prior formal context. A formal context generated by $|G|$ objects and $|M|$ attributes will—in worst case—produce a concept lattice consisting of $2^{\min(|G|,|M|)}$ concepts. In particular, an $n \times n$ formal context $\mathbb{K} := (\{1, \ldots, n\}, \{1, \ldots, n\}, \neq)$ has $2^n$ concepts.

The following section presents a naive algorithm which to make the computational challenges of finding concepts clear.

As introduced earlier, a subset $A$ of the set of all objects $G$ is an extent of a concept if $A'' = A$. In this case the concept is $(A, A')$, analogous for a subset $B$ with $B$ an intent of $M$ if $B'' = B$. Accordingly, the concept is $(B', B)$.

(i) $\mathfrak{B} := \emptyset$
(ii) Select $A \in 2^G$, with $2^G$ the power set of $G$
(iii) Derive $A'$
(iv) Derive $A'' = (A')'$
(v) $\mathfrak{B} := \mathfrak{B} \cup \{(A'', A')\}$
(vi) Repeat steps 2 to 5 for all remaining $A \in 2^G$
(vii) $\mathfrak{B}$ is the set of all concepts in $\mathbb{K} = (G, M, I)$

*An example of the naive method*

For easier readability, we enumerate the objects such that B = Biometric, T = Training, V = Vehicles. Attributes are represented by their lowercase short names from o1 to o5.

Given the example cross table in Table 9.2, this method derives all the concepts as illustrated in Table 9.3. The concepts of $\mathbb{K}$ denoted by $\mathfrak{B}(\mathbb{K})$ consist of all $(A'', A')$ as listed in the last column. Thus, the resulting set of all concepts

**Table 9.3.** The naive method suggests to iterate over the complete power set of $G$ and two derivations. All concepts are listed in the last column.

| $2^G$ | $A'$ | $A''$ | $(A'', A')$ | $\mathfrak{B}(\mathbb{K})$ |
|---|---|---|---|---|
| $\emptyset$ | $\{o1, o2, o3, o4, o5\}$ | $\emptyset$ | $(\emptyset, \{o1, o2, o3, o4, o5\})$ | $\{(\emptyset, \{o1, o2, o3, o4, o5\}),$ |
| $\{B\}$ | $\{o2, o3, o5\}$ | $\{B\}$ | $(\{B\}, \{o2, o3, o5\})$ | $(\{B\}, \{o2, o3, o5\}),$ |
| $\{T\}$ | $\{o1, o2, o3\}$ | $\{T, V\}$ | $(\{T, V\}, \{o1, o2, o3\})$ | $(\{T, V\}, \{o1, o2, o3\}),$ |
| $\{V\}$ | $\{o1, o2, o3, o4\}$ | $\{V\}$ | $(\{V\}, \{o1, o2, o3, o4\})$ | $(\{V\}, \{o1, o2, o3, o4\}),$ |
| $\{B, T\}$ | $\{o2, o3\}$ | $\{B, T, V\}$ | $(\{B, T, V\}, \{o2, o3\})$ | $(\{B, T, V\}, \{o2, o3\})\}$ |
| $\{B, V\}$ | $\{o2, o3\}$ | $\{B, T, V\}$ | $(\{B, T, V\}, \{o2, o3\})$ | |
| $\{T, V\}$ | $\{o1, o2, o3\}$ | $\{T, V\}$ | $(\{T, V\}, \{o1, o2, o3\})$ | |
| $\{B, T, V\}$ | $\{o2, o3\}$ | $\{B, T, V\}$ | $(\{B, T, V\}, \{o2, o3\})$ | |

is $\mathfrak{B}(\mathbb{K}) = \{\ (\emptyset, \{o1, o2, o3, o4, o5\}),\ (\{B\}, \{o2, o3, o5\}),\ (\{T, V\}, \{o1, o2, o3\}),\ (\{V\}, \{o1, o2, o3, o4\}),\ (\{B, T, V\}, \{o2, o3\})\}$.

The first efficient method, the Next Closure Algorithm, was proposed by Ganter (1986). Today, there are many algorithms to gain all concepts of a formal context, see (Ganter and Wille 1999) and (Carpineto and Romano 2004) for further reference.

*Remark: Dicliques and concepts*

Besides, as an example of an analogy between FCA and another theory, we would like to mention Robert Haralick's pioneering work from 1974 in graph theory (Haralick 1974) which introduced a diclique OK as a maximal complete directed bipartite subgraph of a graph. Instead of extents $A$ and intents $B$, Haralick used $I$ and $O$ as inputs and outputs. Indeed, the notion of dicliques coincides with the notion of formal concepts. Haralick's diclique algorithm is shown below:

(i) First form a set $T$ of subsets $R(x), x \in X$
(ii) For all output sets of $T$, intersections are generated and added to the set $T$
(iii) For each output set $O \in T$, its corresponding input set $I = \bigcap_{o \in O} R^{-1}(o)$ is computed

Another method to generate a set of concepts, which is used in the example with Web searches is shown in Algorithm 3.

**The basic theorem of FCA**

Let $(A_1, B_1)$ and $(A_2, B_2)$ be formal concepts of a formal context $(G, M, I)$. If $A_1 \subseteq A_2$ holds, which is equivalent to $B_2 \subseteq B_1$, then $(A_1, B_1)$ is called a *subconcept* of $(A_2, B_2)$, and respectively, $(A_2, B_2)$ is a *superconcept* of $(A_1, B_1)$. In this situation we will write $(A_1, B_1) \leq (A_2, B_2)$. This defines an order relation on the set of all formal concepts.

The set of all formal concepts of $(G, M, I)$ ordered by the previously introduced order relation $\leq$ is called a *concept lattice* and is denoted by

$$\underline{\mathfrak{B}}(G, M, I) := (\mathfrak{B}(G, M, I), \leq)$$

Now we are prepared to state the basic theorem of FCA (Ganter and Wille 1999):

The concept lattice $\underline{\mathfrak{B}}(G, M, I)$ is a complete lattice in which infimum and supremum are given as in Eqs. (9.11) and (9.12). Furthermore, a complete lattice $\mathbf{V}$ is isomorphic to $\underline{\mathfrak{B}}(G, M, I)$ if and only if there are mappings $\widetilde{\gamma} : G \to V$ and $\widetilde{\mu} : M \to V$ such that $\widetilde{\gamma}(G)$ is supremum-dense in $\mathbf{V}$, $\widetilde{\mu}(M)$ is infimum-dense in $\mathbf{V}$ and $gIm$ is equivalent to $\widetilde{\gamma}g \leq \widetilde{\mu}m$ for all $g \in G$ and for all $m \in M$. In particular, $\mathbf{V} \cong \underline{\mathfrak{B}}(V, V, \leq)$.

Indeed, in a concept lattice $\underline{\mathfrak{B}}(G, M, I)$, there is always an *infimum* (greatest common subconcept) and a *supremum* (least common superconcept) given by:

$$\inf(A_t, B_t)_{t\in T} = \bigwedge_{t\in T}(A_t, B_t) = (\bigcap_{t\in T} A_t, (\bigcup_{t\in T} A_t)') \tag{9.11}$$

$$\sup(A_t, B_t)_{t\in T} = \bigvee_{t\in T}(A_t, B_t) = ((\bigcup_{t\in T} B_t)', \bigcap_{t\in T} B_t) \tag{9.12}$$

**Line diagram of a concept lattice**

*How to read extents and intents*

To read the extent or intent of a concept, one can follow the edges represented by the sub- and superconcept hierarchy in the line diagram. The extent of a concept can be obtained by joining all objects located at the respective circle and the circles that can be reached by descending paths from the circle. Analogously, obtaining the intent of a concept, one follows the ascending paths.

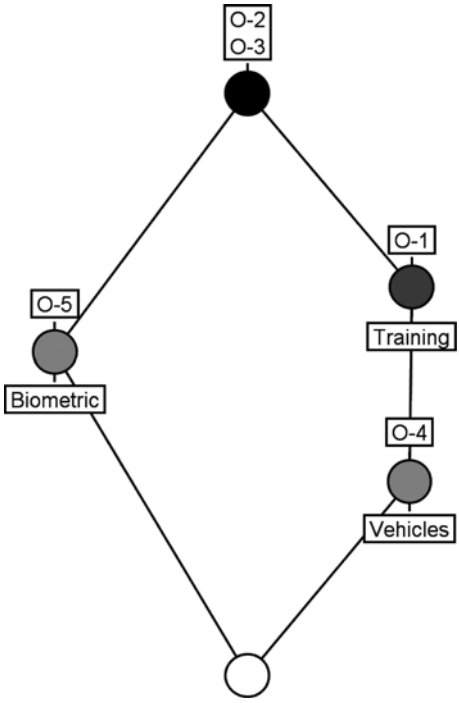



**Fig. 9.1.** Line diagram for the formal context of the Gap Correlations to Operational Functions as introduced earlier. The concept lattice shows five concepts.

In the example in Fig. 9.1, we can easily read off the extent of the concept labeled 'O-1, Training' as 'Training' and 'Vehicles'. Equivalently, the intent of this complies 'O-1,' 'O-2,' and 'O-3.'

*How to read implications*

To read implications from a line diagram, one chooses an attribute set and takes the corresponding concepts from which the infimum is built. Those attributes that are above the concept identified by the infimum are implied by the previously chosen attributes. In our example in Fig. 9.1, for instance, we choose 'O-1' and 'O-5' as the attribute set for which we are interested in obtaining its implications. We can then build the infimum of the two attribute concepts, which is the bottom concept and is not labeled at all. Accordingly, all attributes are implied by 'O-1' and 'O-5.' Implications are useful especially for finding functional dependencies and association rules.

*How to cope with uncertainty*

In contrast to the classic logic approach with two values (0 and 1), fuzzy logic introduces truth degrees. So it becomes a matter of degree to which an object is related to an attribute (Belohlavek 2002; Pollandt 1996). One other approach is the rough set theory. Rough sets have been proposed to cope with undefinable sets, which help to approximate a concept. An overview as well as an introduction of rough sets in combination with FCA terminology is provided in Yao and Chen (2004).

*How to cope with large data sets*

Stumme et al. (2002) show that adopting frequent itemsets into concept lattice theory considers only top-most concepts of a lattice. Reduction is done with the introduction of a minimum support count for the attribute set (the itemset). Then, only concepts with a frequent intent are gathered in an iceberg concept lattice. Iceberg lattices can also be used for conceptual clustering to find implications among concepts. This method also provides a good and efficient starting point for association rule mining. The latter can be done more efficiently than with a normal frequent itemset, because a reduced set of frequent concepts can be used instead. Especially in data mining, the word 'item' is commonly used for what in this chapter is called an 'attribute.' Therefore, itemsets can be understood in FCA terms as the subsets of the attribute set $M$. One native FCA method to cope with larger data sets which is called *conceptual scaling* and will be demonstrated with a data set given by the MIPT Terrorism Knowledge Base.

Those who are completely fearless of heights are invited not to skip the following section.

## Building a bridge: A universal approach to FCA

Only those who don't mind a little detour via Mt. Everest are encouraged to read the following section. As a reward, they may enjoy the beautiful view onto the basic theorem of FCA (as introduced above) and may also obtain a better understanding of formal concept lattices via category theory. To be more precise, we will present a novel insight into concept lattices via the mathematical category of ordered sets—with adjunctions as morphisms: Our main result will state that every adjunction between ordered sets can be decomposed into an epimorphism followed by a monomorphism, and that this decomposition is unique up to isomorphy.

Let's bring it on: As introduced earlier, a pair $\mathbb{P} = (P, \leq)$ is called an *ordered set* if $P$ is a set and $\leq$ is a reflexive, transitive, and anti-symmetric binary relation on $P$. An *adjunction* (between ordered sets) is a quadruple $\mathcal{A} = (\mathbb{P}, \mathbb{L}, f, g)$ consisting of ordered sets $\mathbb{P} = (P, \leq)$ and $\mathbb{L} = (L, \leq)$ and maps $f : P \to L$ and $g : L \to P$ such that for all $x \in P$ and $y \in L$ we have: $f(x) \leq y \Leftrightarrow x \leq g(y)$.

Here, the map $g$ is uniquely determined by $f$ (and vice versa) and will be denoted in the sequel by $f^+ := g$. For short, we will also write

$$\mathbb{P} \underset{f^+}{\overset{f}{\rightleftarrows}} \mathbb{L}$$

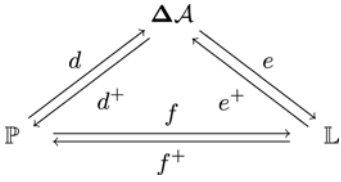
instead of $(\mathbb{P}, \mathbb{L}, f, f^+)$.

If $\mathcal{A} = (\mathbb{P}, \mathbb{L}, f, f^+)$ and $\mathcal{D} = (\mathbb{L}, \mathbb{M}, h, h^+)$ are adjunctions, then so is their *composition* $\mathcal{AD} := (\mathbb{P}, \mathbb{M}, h \circ f, f^+ \circ h^+)$. It is now obvious that ordered sets together with adjunctions between them form a category.

For an adjunction $\mathcal{A} = (\mathbb{P}, \mathbb{L}, f, f^+)$, we always have $f \circ f^+ \circ f = f$ and $f^+ \circ f \circ f^+ = f^+$. Furthermore, $\mathcal{A}$ is a *monomorphism* if and only if $f^+ \circ f = id_P$ holds. $\mathcal{A}$ is an *epimorphism* if and only if $f \circ f^+ = id_L$ holds (here, $id_P$ and $id_L$ denote the identity maps on $P$ and $L$, respectively).

The *diagonal* of $\mathcal{A}$ is given by the ordered set $\mathbf{\Delta}\mathcal{A} := (\Delta\mathcal{A}, \leq)$, where $\mathbf{\Delta}\mathcal{A} := \{(x, y) \in P \times L \mid f(x) = y \text{ and } f^+(y) = x\}$ is ordered via $(s, t) \leq (x, y)$ if and only if $s \leq x$ and $t \leq y$.

**Theorem 1 (Decomposition Theorem for Adjunctions)** *Every adjunction $\mathcal{A} = (\mathbb{P}, \mathbb{L}, f, f^+)$ can be decomposed into an epimorphism $(\mathbb{P}, \mathbf{\Delta}\mathcal{A}, d, d^+)$ followed by a monomorphism $(\mathbf{\Delta}\mathcal{A}, \mathbb{L}, e, e^+)$ as shown below:*



*Here,* $d : P \to \Delta\mathcal{A}, x \mapsto ((f^+ \circ f)(x), f(x))$ *and* $d^+ : \Delta\mathcal{A} \to P, (x, y) \mapsto x$ *and* $e : \Delta\mathcal{A} \to L, (x, y) \mapsto y$ *and* $e^+ : L \to \Delta\mathcal{A}, y \mapsto (f^+(y), (f \circ f^+)(y))$. *Furthermore,* $\mathbf{\Delta}\mathcal{A}$ *is determined up to isomorphy (that is, every other decomposition of* $\mathcal{A}$ *into an epimorphism followed by a monomorphism is isomorphic to* $\mathbf{\Delta}\mathcal{A}$*).*

This theorem has a remarkable consequence for complete lattices. As mentioned before, a complete lattice is known to be an ordered set $\mathbb{P}$ such that every subset $X$ of $\mathbb{P}$ has a *greatest lower bound*, denoted by $\inf_{\mathbb{P}}(X)$, and a *least upper bound*, denoted by $\sup_{\mathbb{P}}(X)$. If $\mathbb{P}$ and $\mathbb{L}$ are complete lattices, then for every adjunction $\mathcal{A} = (\mathbb{P}, \mathbb{L}, f, f^+)$ it follows that $f$ is sup-*preserving* and $f^+$ is inf-*preserving*, that is, $f(\sup_{\mathbb{P}}(X)) = \sup_{\mathbb{L}}(f(X))$ and $f^+(\inf_{\mathbb{L}}(Y)) = \inf_{\mathbb{P}}(f^+(Y))$ hold for all subsets $X$ of $\mathbb{L}$ and $Y$ of $\mathbb{P}$. In particular, we derive from the above theorem the following:

**Corollary 1** *For every adjunction* $\mathcal{A} = (\mathbb{P}, \mathbb{L}, f, f^+)$ *between complete lattices, its diagonal* $\mathbf{\Delta}\mathcal{A}$ *forms a complete lattice:*

$$\inf_{\mathbf{\Delta}\mathcal{A}}(x_i, y_i)_{i\in I} = (\inf_{\mathbb{P}}(x_i)_{i\in I}, f(\inf_{\mathbb{P}}(x_i)_{i\in I}))$$

*and*

$$\sup_{\mathbf{\Delta}\mathcal{A}}(x_i, y_i)_{i\in I} = (f^+(\sup_{\mathbb{L}}(y_i)_{i\in I}), \sup_{\mathbb{L}}(y_i)_{i\in I})$$

*for all* $(x_i, y_i) \in \Delta\mathcal{A}$ *(with* $i \in I$*).*

This corollary can be applied to Galois connections as follows: If $\mathbb{L} = (L, \leq)$ is an ordered set, let $\mathbb{L}^d := (L, \geq)$ denote the *dual* of $\mathbb{L}$. A *Galois connection* between ordered sets $\mathbb{P}$ and $\mathbb{L}$ is defined as an adjunction between $\mathbb{P}$ and $\mathbb{L}^d$. Every Galois connection $\mathcal{G} = (\mathbb{P}, \mathbb{L}, f, f^+)$ is associated with its *Galois diagonal* $\mathbf{\Delta}\mathcal{G}$, which is defined as the diagonal of $(\mathbb{P}, \mathbb{L}^d, f, f^+)$, that is, $\mathbf{\Delta}\mathcal{G} := \mathbf{\Delta}(\mathbb{P}, \mathbb{L}^d, f, f^+)$.

**Corollary 2** *Let* $\mathcal{G} = (\mathbb{P}, \mathbb{L}, f, f^+)$ *be a Galois connection. Then the Galois diagonal* $\mathbf{\Delta}\mathcal{G}$ *forms a complete lattice and for all* $(x_i, y_i)$ *in* $\mathbf{\Delta}\mathcal{G}$ *(with* $i \in I$*) we have:*

$$\inf_{\mathbf{\Delta}\mathcal{G}}(x_i, y_i)_{i\in I} = (\inf_{\mathbb{P}}(x_i)_{i\in I}, f(\inf_{\mathbb{P}}(x_i)_{i\in I}))$$

*and*

$$\sup_{\mathbf{\Delta}\mathcal{G}}(x_i, y_i)_{i\in I} = (f^+(\inf_{\mathbb{L}}(y_i)_{i\in I}), \inf_{\mathbb{L}}(y_i)_{i\in I})$$

This last corollary applies to FCA nicely: Every formal context $\mathbb{K} = (G, M, I)$ (note: here, the set $I$ is a binary relation, the incidence relation, on $G$ and $M$) is associated with a Galois connection $\mathcal{G}\mathbb{K} = (\mathbf{2}^G, \mathbf{2}^M, f_{\mathbb{K}}, f_{\mathbb{K}}^+)$ where $\mathbf{2}^G = (2^G, \subseteq)$ and $\mathbf{2}^M = (2^M, \subseteq)$ denote the power set lattices of $G$ and $M$, respectively, and $f_{\mathbb{K}} : 2^G \to 2^M, X \mapsto X'$ and $f_{\mathbb{K}}^+ : 2^M \to 2^G, Y \mapsto Y'$. Indeed, the Galois diagonal of $\mathcal{G}\mathbb{K}$ is nothing else but the concept lattice of $\mathbb{K}$, that is, $\underline{\mathfrak{B}}\mathbb{K} = \mathbf{\Delta}(\mathcal{G}\mathbb{K})$. The basic theorem of FCA is now reflected in Corollary 2.

This is the punch-line and we have completed our little excursion. Now, let's harvest some fruits on the application side of FCA.

## Analysis of target data from the MIPT Terrorism Knowledge Base

In this section, we would like to show how FCA can help to better analyze and understand large data sets. The Memorial Institute for the Prevention of Terrorism (MIPT) is a non-profit organization dedicated to preventing terrorism on U.S. soil or mitigating its effects. The MIPT provides a public Terrorism Knowledge Base (TKB) available on the Web,[1] which can be used for comprehensive research and analysis on global terrorist activities. The TKB covers the history, affiliations, locations, and tactics of terrorist groups operating across the world, with more than a hundred group and leader profiles as well as trials.

The MIPT Terrorism Knowledge Base offers a public Web interface to list, retrieve, and filter terrorist incidents since 1968. In this section, we will investigate information on terrorism targets. The TKB differentiates between domestic and international targets and lists all target incidents along with each the number of injuries and fatalities. Domestic are those incidents which were perpetrated by local nationals against a purely domestic target, whereas international incidents are defined as being those in which terrorists go abroad to strike the targets, select domestic targets associated with a foreign state, or create an international incident by attacking airline passengers, personnel, or equipment. Although data provided by the TKB contains only international incidents for years 1968–1997, both domestic and international incidents do exist for the years 1998 to present. Thus, in this section, we are primarily focussing on data beginning 1998–2006.

The TKB data for incidents to targets is provided in a numerical chart and it allows the reader to naively compare incidents via targets and also get an idea whether the casualties involved in an incident are more likely to be domestic or international. MIPT provides the possibility to retrieve the data interactively by selecting a time interval of one's choice. One striking challenge for an FCA analysis, which we will adopt in this section, is the question how to rigorously discover potential dependencies between targets and their type of impact as well as their origin.

### A first glance at MIPT's TKB target data

In Table 9.4, two retrieval results (one domestic, one international) of the MIPT target data set have been connected to one large data table. Target categories are listed in the first column. The next three columns show the domestic incidents, injuries, and fatalities. In contrast, the last three columns provide the international incidents, injuries, and fatalities. Instead of partially filling the columns with crosses (as done in earlier sections with cross tables in FCA), here we are facing a different type of data, namely, quantitative data.

[1] `http://www.tkb.org`

**Table 9.4.** Statistical data categorized by targets of the MIPT Terrorism Knowledge Base. Data is filtered by both domestic and international targets ranging from 01/01/1998 to 11/27/2006.

| Target | Domestic | | | International | | |
|---|---|---|---|---|---|---|
| | Incidents | Injuries | Fatalities | Incidents | Injuries | Fatalities |
| Abortion Related | 5 | 2 | 2 | 0 | 0 | 0 |
| Airports & Airlines | 88 | 104 | 66 | 40 | 121 | 32 |
| Business | 1471 | 4443 | 1424 | 357 | 3408 | 3105 |
| Diplomatic | 42 | 18 | 24 | 351 | 6487 | 483 |
| Educational Institutions | 529 | 1279 | 521 | 13 | 60 | 11 |
| Food or Water Supply | 6 | 0 | 0 | 0 | 0 | 0 |
| Government | 4566 | 8788 | 4404 | 124 | 648 | 457 |
| Journalists & Media | 376 | 200 | 161 | 63 | 106 | 57 |
| Maritime | 7 | 32 | 14 | 4 | 51 | 18 |
| Military | 99 | 984 | 457 | 23 | 47 | 9 |
| NGO | 92 | 84 | 66 | 100 | 94 | 153 |
| Other | 1199 | 1694 | 1508 | 287 | 386 | 499 |
| Police | 3602 | 12160 | 6914 | 81 | 386 | 134 |
| Private Citizens & Property | 4702 | 17211 | 9355 | 328 | 2456 | 686 |
| Religious Figures/ Institutions | 854 | 4819 | 1888 | 45 | 264 | 118 |
| Telecommunication | 149 | 77 | 54 | 6 | 0 | 0 |
| Terrorists/Former Terrorists | 125 | 74 | 153 | 2 | 1 | 6 |
| Tourists | 63 | 238 | 23 | 33 | 942 | 376 |
| Transportation | 936 | 4951 | 1571 | 72 | 1371 | 450 |
| Unknown | 615 | 866 | 322 | 57 | 47 | 26 |
| Utilities | 936 | 464 | 367 | 25 | 17 | 17 |

When modeling the data provided by the table shown in Table 9.4 for further investigation and analysis in FCA, we first need to assign the sets for our formal objects and formal attributes. The formal context $\mathbb{K}_{TKB}$ in this case will be built up assigning targets as objects $G$, and the casualties as attributes $M$, and an incidence relation $I$ as follows:

$$\mathbb{K}_{TKB} := (G, M, I) \tag{9.13}$$

Using the allocation of attributes $m \in M$ that are assigned to a target object $g \in G$, the incidence relation $I$ for the formal context $\mathbb{K}_{TKB}$ can be constructed. However, in this data set, we face the problem that we see quantitative data. FCA, or more specifically its basic data type, the formal context, however, needs a one-valued relationship between objects and attributes. The next section will show how we can cope with these situations.

**Scaling of data provided by the MIPT Terrorism Knowledge Base**

In the following, we will show how Formal Concept Analysis is an adequate tool to investigate database-type of information. This can be achieved by a powerful FCA-method called *scaling*. In most real-world situations, data is given in many-valued contexts, that is, instead of (one-valued) cross tables we have a (many-valued) database. In order to make such databases accessible to FCA, introducing a scaling process is necessary which involves special interpretation. Most frequently used are nominal, ordinal, multi-ordinal, contranominal, and dichotomic scales. The explicit—and thus transparent—act of interpretation consists of the choice of the type of scale and the subsequent setting of thresholds. As a result, each many-valued attribute is replaced by scale attributes.

Since interpretative work is involved, a suggestive selection of conceptual scaling, in this case an ordinal scaling, and a setting of thresholds needs to be performed. In this investigative analysis, we have chosen to scale two kinds of information given in the data table shown in Table 9.4: first, the total number of incidents per target, and also the comparison between domestic or international majority in the amount of events (incidents, injuries, and fatalities).

Based on the data given, we have focused on the sum of all incidents per target, summing up the domestic incidents with the international incidents. The scaling thresholds have then been set as follows: 'small' means '$\geq 0$,' 'medium' means '$\geq 1,000$,' 'large' means '$\geq 2,500$,' and 'very large' means '$\geq 5,000$.' Then, the original matrix can be transformed into a one-valued cross table, representing a formal context, as follows: mark a cross in every position in which the target fulfills the newly introduced attribute ($\geq 0$, $\geq 1,000$, $\geq 2,500$, and $\geq 5,000$). For example, we consider the target, 'Military.' The sum of domestic and international incidents is $99 + 23 = 122$. Hence, this target as a particular formal object fulfills only the first formal attribute 'small': $122 \geq 0$, but $122 \ngeq 1,000$. The object 'Military' has no crosses under the attributes 'medium,' 'large,' and 'very large.'

The second scaling distinguishes between predominantly domestic and international events. Each domestic column, we compare with its international counterpart as follows: '# domestic incidents $\geq$ # international incidents,' '# domestic injuries $\geq$ # international injuries,' and '# domestic fatalities $\geq$ # international fatalities.' The scaling heuristics can then be set as follows: 'incidents: mainly domestic,' 'injuries: mainly domestic,' and 'fatalities: mainly domestic.' With the international dominance vice versa. For example, if we consider the target 'Airports & Airlines,' we first compare the incidents: the number of domestic incidents are 88, whereas the international number of incidents are only 40. Here, we can make a cross under the new attribute 'incidents: mainly domestic' in the cross table. Analogously, we compare the number of domestic injuries for the target. Here, we have 104 in a domestic environment, whereas internationally, the are 121 injuries. In this case,

we cannot make a cross under the new attribute 'injuries: mainly domestic.' Finally, we compare the fatalities for the target. We face 66 domestic fatalities and only 32 international fatalities. Hence, we can make a cross under the new attribute 'fatalities: mainly domestic.' For the next three new attributes, namely, 'incidents: mainly international,' 'injuries: mainly international,' and 'fatalities: mainly international,' we take the inverse relation. Accordingly, we can simplify our work by making crosses with those attributes related to their counterparts in the domestic columns where there is no cross. Now, we have translated the many-valued database into a one-valued formal context as shown in Table 9.5.

**Table 9.5.** Cross table for a scaled data set of the MIPT Terrorism Knowledge Base.

| Targets | small # | medium # | large # | very large # | incidents: mainly domestic | injuries: mainly domestic | fatalities: mainly domestic | incidents: mainly international | injuries: mainly international | fatalities: mainly international |
|---|---|---|---|---|---|---|---|---|---|---|
| Abortion | × | | | | × | × | × | | | |
| Airports & Airlines | × | | | | × | | × | | × | |
| Business | × | × | | | × | × | × | | | |
| Diplomatic | × | | | | | | | × | × | × |
| Educational Institutions | × | | | | × | × | × | | | |
| Food or Water Supply | × | | | | × | × | × | | | |
| Government | × | × | × | | × | × | × | | | |
| Journalists & Media | × | | | | × | × | × | | | |
| Maritime | × | | | | × | | | | × | × |
| Military | × | | | | × | × | × | | | |
| NGO | × | | | | | | | × | × | × |
| Other | × | × | | | × | × | × | | | |
| Police | × | × | × | | × | × | × | | | |
| Private Citizens & Property | × | × | × | × | × | × | × | | | |
| Religious Figures/Institutions | × | | | | × | × | × | | | |
| Telecommunications | × | | | | × | × | × | | | |
| Terrorists/Former Terrorists | × | | | | × | × | × | | | |
| Tourists | × | | | | × | | | | × | × |
| Transportation | × | × | | | × | × | × | | | |
| Unknown | × | | | | × | × | × | | | |
| Utilities | × | | | | × | × | × | | | |

This section has shown that conceptual scaling is not an automatic process, it is rather an act of interpretation. Accordingly, we could have chosen any other kind of scale in order to interpret the many-valued MIPT Terrorism Knowledge Base. However, we have made certain decisions in a way we wanted to analyze the given data. After the scaling process, we now have the basic data type for FCA, the formal context. The reason why we have scaled the data is the benefit for further investigating the data using the earlier introduced visualization technique to present data in a line diagram which we call the concept lattice. The next section will deal with the interpretation of the resulting concept lattice.

### Interpreting the concept lattice of the MIPT TKB

After the scaling process of the original data which leads to a one-valued formal context $\mathbb{K}_{TKB} := (G, M, I)$, we are now able to transform the formal context $\mathbb{K}_{TKB}$ into a concept lattice $\mathfrak{B}(\mathbb{K}_{TKB})$ without losing any kind of information. The concept lattice as obtained from ToscanaJ[2] is shown in Fig. 9.2.

Let us recall the basics of reading a concept lattice. The circles represent the concepts with optional labels above and below. Labels above a concept represent attributes and labels below represent objects. An extent of a concept can be read off by joining all objects located at a concept and the concepts that can be reached by descending paths from the concept. Analogously, reading off the intent of a concept, one follows the ascending paths. Let us consider the concept labeled by the attribute 'injuries: mainly international.' What is the meaning of this concept? First of all, we read off all its intents. We see that the impact is rather small (though the only ascending edge leads us to the attribute 'small #'). Then, the attribute label of the concept indicates all injuries are mainly international. Now, we are interested in the appropriate extent to the concept that describes only a small impact on incidents, but it reveals a higher rate of injuries internationally than domestically. We can read off by following the descending paths. First, we go down left twice and find a concept with the object 'Airports & Airlines.' Then, we go back to the beginning. We go down left once and then down right once. We find a concept labeled by two objects: 'Maritime' and 'Tourists.' Then, we go back to the beginning once more and go down right twice. Again, we find a concept labeled by two objects, here: 'Diplomatic' and 'NGO.' Now, we have found all objects below the concept labeled by 'injuries: mainly international.' The extent of that concept is the set of all objects below, which are: 'Airports & Airlines,' 'Maritime,' 'Tourists,' 'Diplomatic,' and 'NGO.' To put it in a nutshell: what does that concept mean? Intentionally, we have a small impact on

[2] The ToscanaJ suite is an open source re-implementation of a commercial product called TOSCANA and can be obtained at `http://toscanaj.sourceforge.net/`. ToscanaJ is part of the TOCKIT framework (an effort to providing a framework for Conceptual Knowledge Processing).

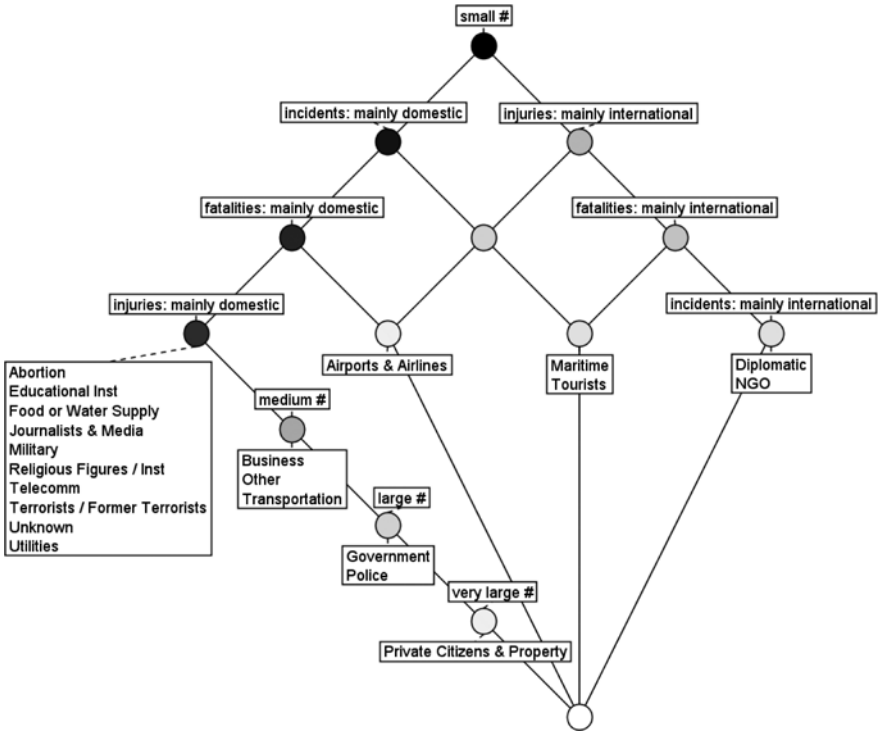



**Fig. 9.2.** A concept lattice visualizing the scaled data of the MIPT Terrorism Knowledge Base. The concept lattice has been visualized using the ToscanaJ suite (Becker and Correia 2005).

incidents, but those incidents have a higher impact on injuries internationally than domestically. The targets are located in the aero, maritime, touristic, diplomatic, and NGO area. We can also see that for the maritime and touristic area, we also face a higher fatality rate in international terrorism than in a domestic situation. Moreover, in the diplomatic and NGO sector even more incidents happen internationally.

By looking at the concept labeled by the attribute 'Airports & Airlines,' we can see that although there are more domestic incidents, the injury rate is higher internationally. However, in contrast, the fatality rate is higher domestically.

If we followed the concepts down left beginning from the top concept, we find a concept labeled by very many objects, namely, 'Abortion,' 'Educational Institutions,' 'Food or Water Supply,' 'Journalists & Media,' 'Military,' 'Religious Figures/Institutions,' 'Telecommunications,' 'Terrorists/Former Terrorists,' 'Unknown,' and 'Utilities.' All these objects have in common that the number of incidents are rather small, but the incidents, injuries, and fatalities

are more likely in a domestic situation. Now, we can navigate further down from this concept by looking at the subconcepts. The next concept represents an at least medium impact on incidents with the objects in its extent. The next concept reveals an even large impact on incidents, and the last one an even larger impact on incidents. Interestingly, we can clearly see that the highest impact is indeed on the private citizens and their property. It can be assumed that this fact has enormously been influenced by the tragic events of September 11, 2001.

## Analysis of snippet data returned by a Web search engine

In this section, we would like to present a way to analyze search engine results via FCA by looking only at the snippet data. The presentation is partly enriched by pseudo-code. First, however, we need to introduce some basic notions and formalizations. For a deeper insight into the following work, we would like to suggest further reading in Koester (2006c) for a complete coverage as well as an investigation on the quality of snippets. In addition, the prior works in Koester (2005) and Koester (2006a) put a stronger emphasis on the contextual retrieval of data. Koester (2006b) primarily focuses on the visualization of Web retrieval data in a concept lattice.

### Web information retrieval

To give a brief overview, the following tasks can be identified as general steps in Web Information Retrieval:

(i) Crawling for Web objects to build up an index.
(ii) A search query is entered into a text field by the user. Depending on the Web search engine, the expression in terms of a query can range from entering solely keywords over a complex query language to forming questions in natural language.
(iii) The search engine looks up potential matches within the index that suits the entered query.
(iv) The potential matches are sorted after relevance and presented to the user in an ordered retrieval result list.

When submitting a search query to a search engine, the results not only consist of simple Web pages. They also consist of PDF documents, music, ASCII text files, directory indexes of Web servers, error pages, etc. Hence, when considering search results, we have to cope with different types of resulting objects. The term 'Web object' seems to be more appropriate than just 'Web document' or 'Web page.' Thus, we use the term *Web object* in this chapter and denote a Web object by $w$. $W$ denotes the set of all Web objects

on the Web. We understand a *search query* as a finite sequence of elements called terms, forming a request for data from a database.

When looking at current Web search engines, we observe that nearly all major search engines return similar result lists. They usually consist of references to matching Web objects, each one at least comprising of

- a title,
- a short description, called a snippet,
- and a URL.

*Title*

The title $w_t$ of a result shows the name given to the Web object by its publisher. For instance, in HTML objects, the tag `<TITLE>...</TITLE>` defines the title. Search engines extract these titles which are then presented as part of the result lists.

*Snippet*

A snippet is a short text excerpt that appears right below a Web object's title in search results of current Web-based search engines. It shall give the user a short description of the content of a Web object. As thoroughly investigated in Koester (2006c), in most search engines, however, other factors besides an excerpt-based composition of a snippet text have an influence on what appears within the short description. With today's search engines, a snippet is composed from various description sources for the short text fragment, not solely relying on original text from the source of the Web object.

*URL*

A Uniform Resource Locator, in short URL, identifies an internet object. It is made up of characters conforming to a standardized format[3] in the form:

`<scheme>://<authority>[<path>[<query>][<fragment>]]`

**A formalization for Web search results**

We can now formalize a retrieved list of Web search results returned from a standard Web search engine.

[3] A detailed introduction into the format of a URL is given in RFC 3986 describing the more general URIs (Uniform Resource Identifiers) at `http://www.ietf.org/rfc/rfc3986.txt`. However, the notion of 'URL' is yet more common and will be preferred throughout this chapter. A URL is the 'address' of a Web object and will be addressed as a primary identifier for Web objects.

As previously introduced, let $W$ denote a set of all Web objects accessible on the internet (at a certain time). Further let $Q$ be a set of feasible query terms. Search results are returned by the search engine in a ranked order. After receiving a user query $q \in Q$, the search engine computes Web objects $w \in W$ that are supposed to match the users' query. If a Web object $w$ is matched by a query $q$, we will formally denote it by the *matching relation* $w \vDash q$. In the following, we will give a more formal description of the latter.

Firstly, we define three more sets: the set $\mathfrak{T}$ of all titles for Web objects, the set $\mathfrak{S}$ of all snippets derived from a Web object and the set $\mathfrak{U}$ of all URLs. Secondly, we introduce four maps: the ranking map $\varrho : W \times Q \to \mathbb{N} \cup \{\bot\}$, the title map $\tau : W \to \mathfrak{T}$, the snippet map $\sigma : W \times Q \to \mathfrak{S}$, and the URL map $\upsilon : W \to \mathfrak{U}$.

For the ranking map $\varrho$ we assume that $\varrho(w, q) = \bot$ if and only if the Web object $w$ does not match the query $q$. Thus, $W_q := \{w \in W \mid w \vDash q\}$ is the set of all Web objects returned by the query $q$. In case that $W_q$ is non-empty, we assume that there exists a natural number $n(q)$ such that $\varrho_q : W_q \to \{1, \ldots, n(q)\}$ is a bijection.

The *retrieval list* is the set $R := \{(\varrho(w, q), \tau(w), \sigma(w, q), \upsilon(w)) \mid (w, q) \in \vDash\}$. For later reference, the result can be simplified. For every $q \in Q$ with a non-empty $W_q$ we define the *title list* $T_q$, the *snippet list* $S_q$, and the *result list* $R_q$ as follows:

$$T_q : \{1, \ldots, n(q)\} \to \mathfrak{T}, i \mapsto \tau(\varrho_q^{-1}(i)) \tag{9.14}$$

$$S_q : \{1, \ldots, n(q)\} \to \mathfrak{S}, i \mapsto \sigma(\varrho_q^{-1}(i), q) \tag{9.15}$$

$$U_q : \{1, \ldots, n(q)\} \to \mathfrak{U}, i \mapsto \upsilon(\varrho_q^{-1}(i)) \tag{9.16}$$

$$R_q : \{1, \ldots, n(q)\} \to \mathfrak{T} \times \mathfrak{S} \times \mathfrak{U}, i \mapsto (T_q(i), S_q(i), U_q(i)) \tag{9.17}$$

In the situation of a fixed query $q \in Q$, we agree upon the following abbreviations, which will be used in the pseudeo-code algorithms later:

For every $i \in \{1, \ldots, n(q)\}$, let $r_i := (t_i, s_i, u_i)$ with $(i, t_i, s_i, u_i) \in R$, that is, $t_i := T_q(i)$, $s_i := S_q(i)$, and $u_i := U_q(i)$.

### FooCA: Retrieval via Google™ and Yahoo!™

The featured approach is to launch a search request using the official programming interfaces provided by the search engines. The returned set of ranked items is then analyzed by means of Web Mining and FCA.

New ways of controlling and obtaining an overview of information need to be established to guide and assist the user instead of ignoring human skills, such as the intuitive understanding of a concept.


**Fig. 9.3.** FooCA is an interface to search results of other Web-based search engines.

This work proposes a vision as well as a working prototype, called FooCA, that shows how traditional, sequential lists can indeed be replaced by conceptual structures.

In a query result, we are interested in the general items returned by the Web search engine to a normal user: the title of the Web object, its URL, and the snippet that represents the semantic context in which the search query (or parts of it) occurred.

FooCA enables the user to interact with the search engine. The user has a specific but informal concept in mind that he or she wants to search for. He or she then enters an approximate query that represents his or her concept, along with chosen options for later FooCA processing. FooCA then receives that information, evaluates the personal options for that user and forwards his or her query to the search engine which interprets the query as if it were a normal search and processes it, returning the results in a ranked order to FooCA. Using the personal options and the search results retrieved, FooCA now generates its internal representation of the formal context retrieved and presents it to the user in a visualized form. From this point on, the user can refine his or her search with the FooCA interface.

**Pre-processing and construction of the formal context**

In order to process attributes within FooCA, we need to identify word tokens. Therefore, some basic feature extraction operations are applied to the standard Web search engine retrieval results:

- Segmentation of tokens divided by space or punctuation.
- Stripping of all HTML format tags.
- Transforming all characters from upper- to lower-case.
- Removing all special characters except for '-'.

For other languages than English, a character mapping can be defined. Using these basic operations, a list of useful word tokens originating from the snippet can easily be generated. The algorithm used by FooCA is outlined in Algorithm 1.

FooCA lets the user enter a query $q$ which is directly passed on to the Web search engine without modification. The main idea is to use that snippet as a starting point since it provides us with a short, non-formal context in which the search query (or parts of it) is embedded. In cases where no snippet is retrieved, the page title is used instead. After extracting feature terms from the retrieved snippets for a specific query $q$, we gain a formal context $\mathbb{K}_{Snippet}$ considering the URLs as objects $G$, the extracted feature terms as attributes $M$ and an incidence relation $I$ as follows:

$$\mathbb{K}_{Snippet}(q) := (G, M, I) \tag{9.18}$$

Using the allocation of attributes $m \in M$ that appear within a snippet belonging to an object $g \in G$, the incidence relation $I$ for the formal context $\mathbb{K}_{Snippet}$ can be constructed.

The algorithm used by FooCA is shown in Algorithm 2. The incidence relation will be represented by a context matrix $C$ with $c_{gm} \in \{0, 1\}$ for all $g \in G, m \in M$, and $c_{gm} = 1 \Leftrightarrow gIm$.

**Representing the formal context in a cross table**

FooCA gives the searcher the possibility to gain more control over the decision-making process by using methods and operations that are usually processed automatically in general search engines without human intervention.

Those methods and operations are: Choice of search engine, interval of the search result retrieval, language restriction, removal of stop words, stemming, clarification of the formal context, user-based query refinement, limiting by an object count for attributes, minimum attribute length, attribute ranking, and exporting of the retrieved formal context. For readers interested in more detail, the methods mentioned as well as the FooCA system are thoroughly explained in Koester (2006c).

Once the user has enabled or disabled specific search strategies (as briefly mentioned above), and entered the search query, FooCA presents the retrieved

**Algorithm 1** Pre-Processing of the Search Results

**Input:** fixed query $q \in Q$, $R_q : \{1, \ldots, n(q)\} \to \mathfrak{T} \times \mathfrak{S} \times \mathfrak{U}, i \mapsto r_i = (t_i, s_i, u_i)$ as introduced above

**Output:** ordered attribute set $M$ consisting of $|M|$ identified tokens, ordered object set $G$ consisting of $|G|$ URLs and a feature extracted snippet list $S_q^*$

$M^{(0)} := \emptyset$;
$G^{(0)} := \emptyset$;
**for all** $1 \leq i \leq n(q)$ **do**
  $G^{(i)} := G^{(i-1)} \cup \{u_i\}$;
  **if** $s_i = \emptyset$ **then**
    $s_i := t_i$;
  **end if**
  $\langle$*Remove all HTML format tags from* $s_i\rangle$;
  $\langle$*Map characters in* $s_i\rangle$;
  $\langle$*Transform characters to lower-case in* $s_i\rangle$;
  $\langle$*Identify the set of tokens* $X_i$ *divided by space or punctuation in* $s_i\rangle$;
  **for all** $x_j \in X_i$ **do**
    $\langle$*Remove all special characters except '-' in* $x_j\rangle$;
    $M^{(i)} := M^{(i-1)} \cup \{x_j\}$;
    $S_q^{*(i)} := M^{(i)}$;
  **end for**
**end for**

**Algorithm 2** Building Context Matrix $C$

**Input:** fixed query $q \in Q$, pre-processed, modified result list $R_q^* : \{1, \ldots, n(q)\} \to \mathfrak{T} \times \mathfrak{S} \times \mathfrak{U}, i \mapsto r_i^* := (t_i, s_i^*, u_i)$ with $s_i^* := S_q^*(i)$, and the attribute set $M$ as gained in Algorithm 1

**Output:** Context matrix $C = \{c_{ij}\}$

$\langle$*Fill matrix* $C$ *with '0'*$\rangle$;
**for all** $1 \leq i \leq n(q)$ **do**
  **for all** $m_j \in M$ such that $1 \leq j \leq |M|$ **do**
    **if** $m_j \in s_i^*$ **then**
      $c_{ij} := 1$;
    **end if**
  **end for**
**end for**

results in an interactive two-dimensional cross table. The row headers are object names, which are clickable numbers in our case, representing the ranked Web object search results. Columns are headed by attributes which are the extracted feature terms of the snippets. The incidence relation $I$ of the formal context $\mathbb{K}_{Snippet}$ between the Web objects $G$ and its attributes $M$ is marked by a cross '$\times$' in the table. The $i$th Web object possesses the $j$th attribute indicated by a cross in the $(i, j)$ position.

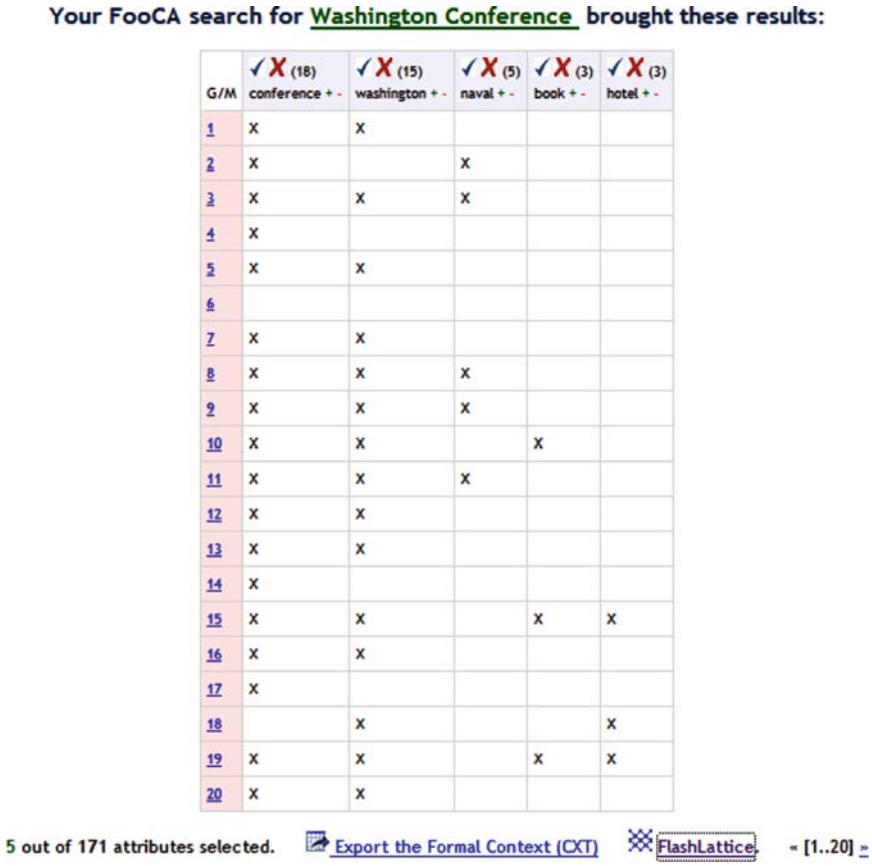



**Fig. 9.4.** The contextual view of a FooCA-based Web search for the query 'Washington Conference' using Yahoo!™ with the first 20 retrieval results in the English language.

The cross table can be navigated using the mouse. As the mouse cursor moves along the table, the underlying row is highlighted indicating the Web object. The user can click anywhere inside the table and is promptly directed to the related Web object.

Apart from navigating inside the table by ways of the incidence relation of the formal context, another navigation method using query refinement is offered. The user can click on any listed attribute name in order to either search for that name directly, launching a new query with that attribute only, or he can include or exclude an attribute by further qualifying the previous query.

**Visualization of the retrieval in a concept lattice**

The concept lattice of our search task, as shown in Fig. 9.5, provides us with the following information: the top concept reveals us that the search returned 20 objects, from which 15 results had snippets featuring the attribute 'washington,' and 18 the attribute 'conference.' It shows that the dominance of the result snippets is rather the term 'conference' instead of the location Washington. Since we were querying for 'Washington Conference,' it can also be seen that there is a strong correlation between the concepts mentioning 'washington' and 'conference.' Actually, the meet of 'washington' and 'conference' was our intended search query. The combination of both query tokens involves 14 out of 20 results. It is not surprising that this concept indeed represents conferences based in Washington (however, both DC as well as the state of Washington). Right below, we see a concept labeled by the attribute 'book.' We can also see that the attribute 'book' implicates 'washington' and 'conference.'

However, the number of extents does not satisfy the amount of information a user might be interested in. When opening the list of extents as additional information, as shown in Fig. 9.6, we might be surprised that both the concept itself as well as the two subconcepts of the concept labeled by the attribute

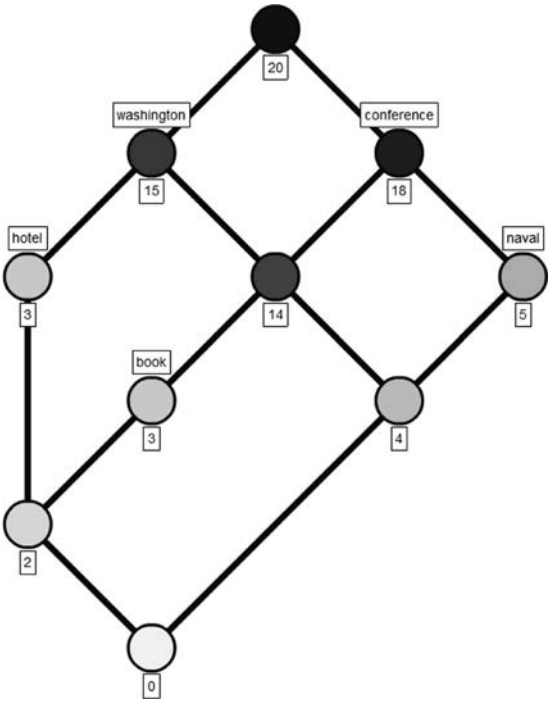



**Fig. 9.5.** The concept lattice of 20 processed research results with FooCA. Here the extent labels show the amount of objects below a concept.

'book' do not refer to a printed book such as a proceeding to a conference, but rather to the process of 'booking' a flight or a hotel room for the conference. This example demonstrates impressively the power of the duality of both intents and extents labeled in the concept lattice.

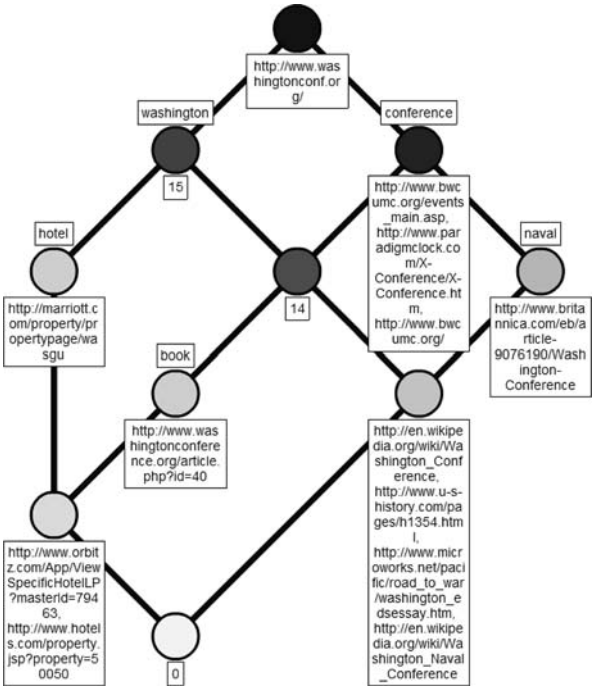



**Fig. 9.6.** The concept lattice of 20 processed research results with FooCA. Here the extent labels show the objects of concept and provide the user with even more information.

Astonishingly, the attribute 'naval' appears quite often within the search results, since it has an extent of five. As the concept lattice suggests, there must be a strong connection between 'naval' and 'conference.' We were wondering what 'naval' means in the informal context of Washington and Conference, so we had to find out by clicking on the concept labeled by 'naval.' A website of the Encyclopedia Britannica opened, explaining the term 'Washington Conference' and turned out to be a synonym for 'Washington Naval Conference,' an 'international conference called by the United States to limit the naval arms race and to work out security agreements in the Pacific area. Held in Washington, DC, the conference resulted in the drafting and signing of several major and minor treaty agreements' (Encyclopedia Britannica). In the line diagram it can be seen that the concept below the concept labeled by 'naval' additionally includes the attribute 'washington.' The snippet data of those search results obviously contained more data considering these terms.

Also surprising is the fact that the website pointing to a Web object called 'WashingtonConf'[4] is located right at the top concept of the lattice. This means that it is not related to the search terms 'washington' or 'conference.' The explanation is rather simple: the website is temporarily unavailable. It does not list conferences in the location of Washington. Instead, when visiting the website, it gives a standard message saying that the website is not configured at this address.

*Similarity of FCA to clustering*

When looking at the concept lattice returned by FooCA, one might immediately think of FCA as being a good technique for clustering. Indeed, FCA has already been applied in both ways, as an algorithmic method for clustering (e.g., Carpineto and Romano 2004) as well as a graphical representation method for clustering (e.g., Hotho et al. 2003). In the latter approach it is shown how a common (non-conceptual) clustering technique such as k-means can be combined with the intensional descriptions provided by a conceptual clustering approach using FCA. For common clustering approaches, the main aim is to cluster data into disjoint sets. When applying FCA, however, it is the strength of the concept lattice to present objects and attributes at the same time. In contrast to disjoint clustering efforts, the idea of FCA is the hierarchical ordering of concepts consisting both of extensions and intensions that allow to read implications and relationships between concepts directly of the concept lattice. Additionally, 'overlapping clusters' can be represented. The clusters need not be disjoint. Hotho et al. (2003) have summarized the advantages using a lattice structure as provided by FCA in comparison to a common clustering effort as follows:

- FCA shows relationships between clusters,
- helps to identify inconsistencies,
- and allows the ability for intentional descriptions of clusters.

**Algorithm 3** Generating a set of all concepts using the intersection method

**Input:** formal context $\mathbb{K} := (G, M, I)$
**Output:** set $\mathfrak{B}(G, M, I)$ of all concepts of $\mathbb{K}$
1: $\mathfrak{B}(\mathbb{K}) := \{ (M', M) \}$;
2: **for all** $g \in G$ **do**
3: **for all** $(X, Y) \in \mathfrak{B}(\mathbb{K})$ **do**
4: Intersection $:= Y \cap \{g\}'$;
5: **if** Intersection is different from any concept intent in $\mathfrak{B}$ **then**
6: $\mathfrak{B}(\mathbb{K}) := \mathfrak{B}(\mathbb{K}) \cup \{(\text{Intersection}', \text{Intersection})\}$;
7: **end if**
8: **end for**
9: **end for**

[4] `http://www.washingtonconf.org`

## Summary and outlook

While Formal Concept Analysis (FCA) has been developed into a widely-spread and well-known theory with annual conferences—namely the International Conference on Formal Concept Analysis (ICFCA), the International Conference on Conceptual Structures (ICCS), and the Conference on Lattices and their Applications (CLA)—all of them entirely devoted to the subject, Haralick's pioneering work on dicliques, though it has been introduced earlier, has obviously (not up to this work) appeared on the radar of the FCA community. This means that we have adopted the terminology of FCA rather than graph-theoretical notions. However, we feel highly inspired by both approaches, since they come from different paradigmatic backgrounds and provide different potential applications for modeling relationships of human societies on various levels. Our understanding of research is rather interdisciplinary. This work aims at building bridges between yet differing scientific methods and fields. For example, the diclique approach gave rise to coalescing dicliques, which turned out to be fruitful within FCA research as well (as has recently been shown in Rome and Haralick 2005). Over the past 25 years of FCA research, thousands of theoretical as well as applied contributions in FCA, ranging from abstract algebra to knowledge processing, have created a sound platform for future work.

We propose Formal Concept Analysis as a meta-theory for integrating disciplines in the area of conceptual analysis, providing a framework for turning data first into information and then into knowledge.

## References

Becker, P., and J. H. Correia. 2005. The ToscanaJ suite for implementing conceptual information systems. In *Formal Concept Analysis, Foundations and Applications*. Ed. B. Ganter, G. Stumme, and R. Wille. Berlin: Springer-Verlag.

Belohlavek, R. 2002. *Fuzzy Relational Systems: Foundations and Principles*. New York: Kluwer Academic.

Carpineto, C., and G. Romano. 2004. *Concept Data Analysis: Theory and Applications*. Chichester: John Wiley & Sons.

Department of Homeland Security, Customs and Border Protection, U.S. Border Patrol, America's Shield Initiative Industry Day, at 18 (Aug. 5, 2004). `http://www.epic.org/privacy/surveillance/spotlight/`

Encyclopedia Britannica. Washington Conference. `http://www.britannica.com/eb/article-9076190`

Farley, J. D. 2003. Breaking Al Qaeda cells: A mathematical analysis of counterterrorism operations (a guide for risk assessment and decision making). In *Studies in Conflict and Terrorism*, 26:399–411(13). Oxford: Taylor & Francis.

Ganter, B. 1986. Algorithmen zur formalen Begriffsanalyse. In *Beiträge zur Begriffsanalyse* (in German). Ed. B. Ganter, R. Wille, and K. E. Wolff. BI Wissenschaftsverlag, Mannheim.

Ganter, B., and R. Wille. 1999. *Formal Concept Analysis: Mathematical Foundations*. Berlin: Springer-Verlag.

Haralick, R. M. 1974. The diclique representation and decomposition of binary relations. *Journal of the Association for Computing Machinery*, 21(3): 356–366.

Hotho, A., S. Staab, and G. Stumme. 2003. Explaining text clustering results using semantic structures. In *Principles of Data Mining and Knowledge Discovery, 7th European Conference, PKDD 2003* 217–228. Berlin: Springer-Verlag.

Koester, B. 2005. Conceptual knowledge processing with Google. In *Lernen, Wissensentdeckung und Adaptivität*. Ed. M. Bauer, J. Fürnkranz, A. Kröner, et al. LWA 2005, GI Workshops, Saarbrücken, October 10–12.

Koester, B. 2006a. FooCA: Enhancing Google information research by means of Formal Concept Analysis. In *Contributions to ICFCA 2006*. Ed. B. Ganter and L. Kwuida. Verlag Allgemeine Wissenschaft, Mühltal.

Koester, B. 2006b. Conceptual knowledge retrieval with FooCA: Improving Web search engine results with contexts and concept hierarchies. In *Advances in Data Mining, Proceedings of ICDM 2006, Leipzig, Germany* 176–190. Ed. P. Perner. Berlin: Springer-Verlag.

Koester, B. 2006c. *FooCA – Web Information Retrieval with Formal Concept Analysis*. Verlag Allgemeine Wissenschaft, Mühltal.

Pollandt, S. 1996. *Fuzzy-Begriffe: Formale Begriffsanalyse unscharfer Daten* (in German). Berlin: Springer-Verlag.

Priss, U. 2006. Formal concept analysis in information science. In *Annual Review of Information Science and Technology*, 40:521–543. Ed. B. Cronin.

Rome, J. E., and R. M. Haralick. 2005. Towards a formal concept analysis approach to exploring communities on the World Wide Web. In *Formal Concept Analysis, ICFCA 2005, Lens, France, February 14–18, 2005, Proceedings*. B. Ganter and R. Godin. Berlin: Springer-Verlag.

Schröder, B. S. W. 2003. *Ordered Sets: An Introduction*. Boston: Birkhäuser, Boston.

Stumme, G., Y. Bastide, L. Lakhal, N. Pasquier, and R. Taouil. 2002. Computing iceberg concept lattices with TITANIC. In *Data Knowledge Engineering*, vol. 42 pp. 189–222. Amsterdam: Elsevier Science Publishers B.V.

Voss, S., and C. Joslyn. Advanced knowledge integration in assessing terrorist threats LDRD-DR. *Los Alamos National Laboratory Internal Report*. LAUR 02-6557.

Yao, Y. Y., and Y. Chen. 2004. Rough set approximations in formal concept analysis. In *Proceedings of 23rd International Meeting of the North American Fuzzy Information Processing Society, NAFIPS 04* 73–78.

# 10

# Reflexive Analysis of Groups

Vladimir A. Lefebvre

**Summary.** This chapter develops further a model I previously introduced, of an agent facing a choice between the positive and the negative poles. Here I will consider agents whose individual behavior depends on a 'society' compounded by all of them. Four ideas underlie the theory. The first idea is to consider relationships between the subgroups of agents, not just pairs of agents; this idea allows us to represent a decomposable graph corresponding to an agent or a group of agents as a tree of subgraphs. The second idea is to establish a correspondence between decomposable graphs and polynomials, allowing us to replace a tree of subgraphs with a tree of polynomials representing a computational process. The third idea consists of the interpretation of the tree of polynomials as an agent who has images of the self, which can have images of the self, etc. Finally, the fourth idea is putting an equation into correspondence to the agent, allowing us to find out the agent's state. The theory is illustrated here with several examples from modern geopolitics, including scenarios of current interest.

## Introduction

In this work, I develop the ideas described in my book *Algebra of Conscience* (Lefebvre 1982, 2001). I have introduced there a model of an agent facing a choice between the positive and the negative poles. Several predictions of the models have already passed experimental tests (Lefebvre 1980; Lefebvre et al. 1986; Adams-Webber 1997; Grice et al. 2005). Now I will consider agents whose individual behavior depends on a 'society' compounded by all of them.

Imagine a group of agents, each pair of which is either in the relationship of union or that of conflict. Let the group members be involved in work over a certain task and each one has to choose between the active and passive lines of behavior. The active behavior is valued as the positive pole, and the passive behavior as the negative pole. Every agent may experience an influence from the other agents and the source inside the self. In the framework of this scheme, the agent can be in one of four states. In the *first* state, the agent is free to choose any line of behavior, active or passive, depending on circumstances. In

S. Argamon, N. Howard (eds.), *Computational Methods for Counterterrorism*, DOI 10.1007/978-3-642-01141-2_10,

this state, the agent is able to realize his strategic thinking. In the *second* state, the agent is deprived of the freedom of choice and always chooses the active line of behavior, even if the passive behavior could be more advantageous. In the *third* state, the agent is also deprived of the freedom of choice, but always chooses the passive line of behavior even when it is harmful. In the *fourth* state, the agent is not capable of making a choice at all: it is either inactive or rushing about between two lines of behavior.

The theory described in this work will answer: How to find the states of the agents by knowing the structure of agents' relationships and their influences on one another.

Four ideas underlie the theory. The essence of the first idea is to transfer 'the relationships between the agents' to 'the relationships between the subgroups of agents,' if the agents of one subgroup have the same relationship with the agents of the other subgroup. This idea allows us to represent a decomposable graph corresponding to an agent or a group of agents as a tree of subgraphs.

The second idea consists of establishing a correspondence between decomposable graphs and polynomials. This allows us to replace a tree of subgraphs with a tree of polynomials representing a computational process.

The third idea consists of the interpretation of the tree of polynomials as an agent who has images of the self, which can have images of the self, etc.

Finally, the fourth idea is putting an equation into correspondence to the agent, and this equation allows us to find out the agent's state.

## Completed graphs

We presume that the reader is knowledgeable about general definitions in graph theory. Further, we will consider only completed and elementary graphs. A graph is called *completed* if any two nodes $a$ and $b$ are connected by a link $(a, b)$. Links $(a, b)$ and $(b, a)$ are equivalent. A graph is called *elementary* if it consists of one node. We divide a set of all links of a non-elementary graph into two disjoint subsets (one of them can be empty) and call them *relations* $R$ and $\bar{R}$. If $(a, b) \in R$, we say that $a$ and $b$ are connected by link $R$, which is recorded as $aRb$. If $(a, b) \in \bar{R}$ then $a$ and $b$ are connected by $\bar{R}$, which is recorded as $a\bar{R}b$. All further definitions for $R$ hold for $\bar{R}$ as well. If two nodes, $a$ and $b$, can be connected by a sequence of $R$-type links, we say that $a$ and $b$ are connected in $R$. If any two nodes of a graph are connected in $R$, we say that a graph is connected in $R$. If every node of a graph $A$ is connected with every node of graph $B$ by link $R$, we write $ARB$. If graph $G$ consists of subgraphs which are in relation $R$ two-by-two, we say that graph $G$ is divided to these subgraphs. In this case, we will write $G = A_1RA_2R...RA_n$, where $A_1, A_2, ..., A_n$ are subgraphs. The expression $\{a, b, ...\}$ designates a graph with nodes $a, b...$ .

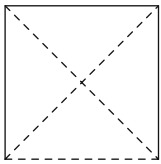


**Fig. 10.1.** Graph $S_{(4)}$. *Solid lines* correspond to $R$. *Dotted lines* correspond to $\bar{R}$.

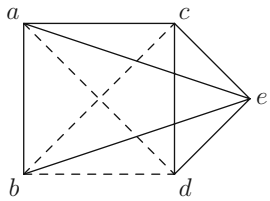



**Fig. 10.2.** An example of a stratified graph which is not totally stratified.

Main definitions:

**Definition 1** A completed graph $G$ is *stratified* in $R$, if it can be represented as $ARB$. The graphs $A$ and $B$ are called *strata* of graph $G$ in $R$.

**Definition 2** A complete graph $G$ is totally *stratified*, if every one of its non-elementary subgraphs is stratified in $R$ or in $\bar{R}$.

**Definition 3** If graph $A$ is a *stratum* of $G$ in $R$ and is non-stratified in $R$, $A$ is a *minimal stratum* in $R$.

## A theorem of total stratification

A graph with four nodes, presented in Fig. 10.1, is called $S_{(4)}$. It is easy to see that this graph is connected both in $R$, and in $\bar{R}$.

**Theorem 1** *A completed graph $G$ is totally stratified if and only if among its subgraphs there is no one isomorphic to $S_{(4)}$.*

(For proofs see Batchelder and Lefebvre 1982 or Lefebvre 2001.)

Consider, for example, the graph in Fig. 10.2. It is easy to see that this graph is stratified because it can be represented as $\{a, b, c, d\}R\{e\}$. But it contains subgraph $\{a, b, c, d\}$ isomorphic to $S_{(4)}$. Therefore, the graph in Fig. 10.2 is not totally stratified.

None of the 4-node subgraphs of the graph shown in Fig. 10.3 is isomorphic to $S_{(4)}$.

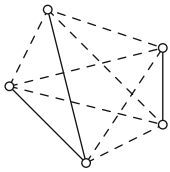


**Fig. 10.3.** Graph $G$, each 4-node subgraph of which is not isomorphic to $S_{(4)}$.

It follows from Theorem 1 that the graph in Fig. 10.3 is totally stratified. Therefore, the main rule of how to find out if a graph is totally stratified is to check that among its 4-node subgraphs there are no graphs isomorphic to $S_{(4)}$.

## A theorem of decomposition

The following statement holds: if a completed graph $G$ is stratified in $R$, there is only one division to minimal strata (to within the numeration of minimal strata) (see proofs in Lefebvre 2001). This statement underlies a decomposition procedure of the completed graph $G$ ($D$-procedure). It consists of consecutive division of the graph $G$ and its subgraphs into minimal strata. Each stratum obtained in this way belongs to a particular level of division with its ordinal number. We investigate each subgraph located at the level $k$, in order to find out if it is stratified. If it *is not*, its examination stops. If it *is stratified*, the subgraph is divided into minimal strata belonging to the level $k + 1$. $D$-procedure generates trees of the type shown in Fig. 10.4.

Every circle corresponds to a subgraph of graph $G$. Symbols $R$ and $\bar{R}$ correspond to relations between minimal strata. If a circle is an *end*, that is, there are no branches starting there, then the corresponding graph is either elementary (consisting of one node), or double-connected. By virtue of the statement given at the beginning of this section, the decomposition tree is unique (to within the order of branches coming from each node). In accordance with the definition of a minimal stratum, relations between the strata at the neighboring levels are different.

If the ends of a graph's decomposition are the graph's nodes, then we call the graph decomposable.

**Theorem 2** *A completed graph $G$ is decomposable if and only if it is totally stratified.*

(See proof in Lefebvre 2001.)

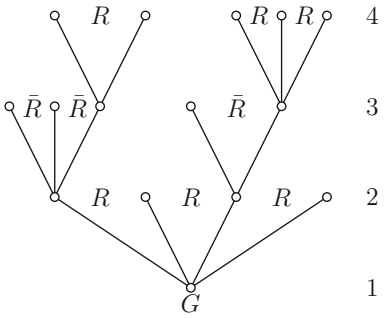



**Fig. 10.4.** An example of a decomposition tree. The numbers on the right are the levels of division.

It follows from Theorems 1 and 2 that if among a graph's 4-node subgraphs there are no graphs isomorphic to $S_{(4)}$ then the graph is decomposable.

## Graphs, polynomials, and polynomial trees

Decomposable graphs can be represented in an analytical form that facilitates their analysis. Let us introduce a procedure for transitioning from a decomposable graph to its analytical form. First, a decomposition tree is constructed. The ends of its branches are designated with the same letters as the nodes on the graph which correspond to them, but the tree ramifications are designated with other letters. Then we construct a grammar tree isomorphic to the tree of decomposition. The ramifications and branches' ends of the two trees are designated with the same letters. Symbols $R$ and $\bar{R}$ are also at the same places. Letters corresponding to the branches' ends will be called *terminal*, and all other letters will be called *intermediate*. Each *intermediate* letter designates a group of letters located directly above a given ramification together with symbols $R$ or $\bar{R}$ and brackets. As a result we obtain the rule for substitution: each intermediate letter can be substituted with a group of symbols (letters, brackets, and signs $R$ or $\bar{R}$ taken into parentheses) located above it. After a series of such substitutions we obtain a *word* which represents an analytical form of the graph.

Consider the completed graph in Fig. 10.5. It is decomposable because it does not contain any subgraph isomorphic to $S_{(4)}$.

The transition of this graph to its analytical form is shown in Fig. 10.6. The solid lines correspond to $R$, the dotted ones to $\bar{R}$.

At the next step, we consider the letters in the graph's analytical form as variables defined on a certain set and symbols $R$ and $\bar{R}$ as a designation of two different binary operations, each one *associative* and *commutative*: $+$ and $\bullet$. With such an interpretation, a graph's analytical form turns into a polynomial. $R$ and $\bar{R}$ may designate $+$ and $\bullet$ in two ways: $R \leftrightarrow \bullet$, $\bar{R} \leftrightarrow +$ or $R \leftrightarrow +$, $\bar{R} \leftrightarrow \bullet$.

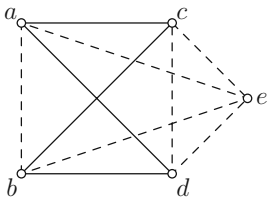



**Fig. 10.5.** A decomposable graph.

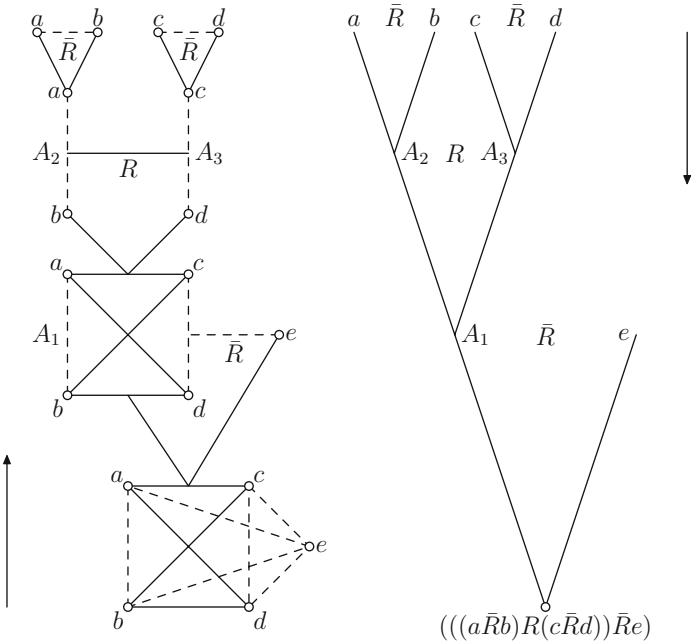



**Fig. 10.6.** An example of a transition from a decomposable graph to its analytical form. A tree of the graph decomposition is on the *left*. A grammatical tree is on the *right*.

Consider formulas (10.1), (10.2), and (10.3):

$$(((a\bar{R}b)R(c\bar{R}d))\bar{R}e) \tag{10.1}$$

$$(((a+b)\bullet(c+d))+e) \tag{10.2}$$

$$(((a\bullet b)+(c\bullet d))\bullet e) \tag{10.3}$$

Formula (10.1) is the analytical form of the decomposable graph in Fig. 10.5; (10.2) is a polynomial if $R$ corresponds to $\bullet$, and $\bar{R}$ to $+$; (10.3) is

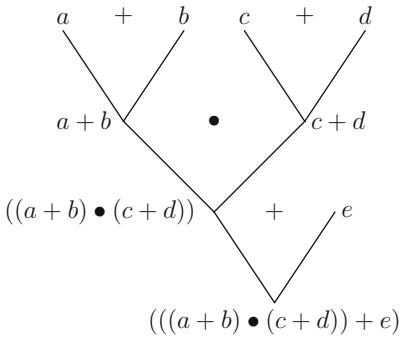



**Fig. 10.7.** A polynomials tree, when • corresponds to $R$, and + to $\bar{R}$.

a polynomial if $R$ corresponds to +, and $\bar{R}$ to •. We call '+' *addition*, and '•' *multiplication*. Polynomial $P$ corresponding to decomposable graph $G$ is represented either as a sum or as a product of other polynomials, each of which corresponds to a minimal stratum of graph $G$ in the same relation $R$. We will call these polynomials *minimal* or *minimal summands* and *minimal multipliers*, respectively. It is impossible to represent a minimal summand as a *sum* of other polynomials and a minimal multiplier as a *product* of other polynomials. A decomposition of polynomial $P$ into minimal polynomials is unique (to within numeration of the minimal polynomials) as follows from one-to-one correspondence between the minimal strata of graph $G$ and minimal polynomials of $P$. A polynomial which contains only one letter will be called *elementary*. Every non-elementary polynomial is a composition of one or two binary operations; we do not assume in advance that they are distributive, hence, multiplying out is not allowed. Every subgraph of the graph $G$ in Fig. 10.6 can also be represented as a polynomial. As a result we obtain a tree of polynomials (see Fig. 10.7).

After the polynomial tree is constructed, we put the polynomials at the ramifications and ends into brackets.

Let us make some simplifications: expressions of the type $[(A)]$ may be changed to $[A]$, and $(A \bullet B)$ to $A \bullet B$ or $AB$. Now, the polynomial tree in Fig. 10.7 can be represented as a conventional tree of polynomials (Fig. 10.8).

Let us agree to write a conventional polynomial tree in a diagonal form and omit the tree branches. For example, the tree in Fig. 10.8 is represented in the following diagonal form:

$$\begin{array}{rcccl} & & [a]+[b] & & [c]+[d] \\ & [a+b] & & \bullet\ [c+d] & \\ [(a+b)(c+d)] & & & & +[e] \\ [(a+b)(c+d)+e] & & & & \end{array} \tag{10.4}$$

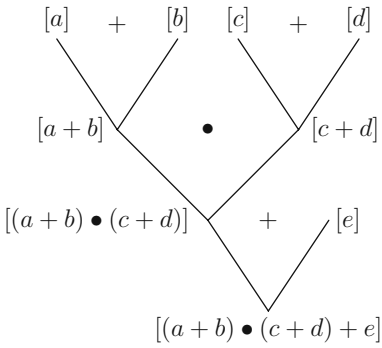



**Fig. 10.8.** A conventional polynomial tree.

If we know a polynomial corresponding to the graph, we can construct its diagonal form without analyzing the graph. A procedure is as follows.

(i) Put the polynomial into brackets and simplify it.
(ii) If it is elementary, the procedure is over.
(iii) If it is a sum of minimal polynomials, we write to the right-up-diagonally the minimal summands in brackets and connected by +.
(iv) If it is a product of minimal polynomials, we write to the right-up-diagonally the minimal multipliers in brackets and connected by •.
(v) With every polynomial written diagonally we repeat the procedure starting step 2.
(vi) A diagonal form is completed when each non-elementary polynomial has a diagonal polynomial.

Let the following polynomial correspond to a graph:

$$((a + b) \bullet c)$$

We put it into the brackets:

$$[((a + b) \bullet c)]$$

and simplify it:

$$[(a + b) \bullet c]$$

Then, we find that it is a product, and write to the right-up-diagonally its multipliers in the brackets and connected by •:

$$\begin{array}{r} [a + b] \bullet [c] \\ [(a + b) \bullet c] \end{array}$$

Consider now the polynomials at the second tier. Polynomial $[a+b]$ is a sum; we write to the right-up-diagonally the sum $[a]+[b]$; $[c]$ is an elementary polynomial, so we do not write anything to the right-up and obtain

$$\begin{array}{ccc} & & [a]+[b] \\ & [a+b] & \bullet\ [c] \\ [(a+b)\bullet c] & & \end{array} \tag{10.5}$$

Finally, by taking away $\bullet$, we obtain

$$\begin{array}{ccc} & & [a]+[b] \\ & [a+b] & [c] \\ [(a+b)c] & & \end{array} \tag{10.6}$$

## A diagonal form as a computational scheme

Let the polynomials introduced above be defined on the set of two Boolean elements $\{0,1\}$. Symbol $+$ corresponds to the Boolean addition (disjunction), and $\bullet$ to the Boolean multiplication (conjunction). Thus, each polynomial is a Boolean formula. Consider a Boolean function that is called *implication*. It is given by the following equation: $f(a,b)=a+\bar{b}$. Let us write it in a diagonal form as

$$f(a,b)=a^b$$

It is easy to see that

$$1^1=1, 1^0=1, 0^1=0, 0^0=1$$

We will also consider that

$$a^{b^c}=a^{(b^c)}$$

Suppose the relation 'up-right-diagonally' has the meaning of implication. Now, every diagonal form is a designation of Boolean function:

$$\Phi=\Phi(a_1,...,a_k,...,a_n) \tag{10.7}$$

Having given values to variables $a_1,...,a_k,...,a_n$, we can find the value of Boolean function $\Phi$, remembering that $[g]=g$. Consider, for example, form (10.6) for $a=0$, $b=0$, $c=1$ and make computations:

$$\begin{array}{ccc} & & [0]+[0] \\ & [0+0] & [1] \\ [(0+0)1] & & =0 \end{array}$$

Further we will use Boolean equations for $a_k$,

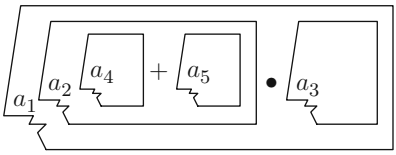



**Fig. 10.9.** An example of a reflexive structure. Subject $a_1$ has images of subjects $a_2$ and $a_3$, whose relationship is •. In its turn, subject $a_2$ has images of subjects $a_4$ and $a_5$ being in relationship +.

$$a_k = \Phi(a_1, ..., a_k, ..., a_n), k \in \{1, 2, ..., n\} \tag{10.8}$$

where the values of variables $a_i$, $(i \neq k)$ are fixed. The right-hand part of (10.8) may equal to: (I) variable $a_k$, (II) Boolean 1, (III) Boolean 0, (IV) variable $\bar{a_k}$. Therefore, Eq. (10.8) can have one of the four forms:

(I) $a_k = a_k$
(II) $a_k = 1$
(III) $a_k = 0$
(IV) $a_k = \bar{a_k}$

In case (I) the equation has two solutions: 0 and 1; in case (II) one solution: 1; in case (III) one solution: 0; and in case (IV) the equation does not have a solution.

## Reflexive structures and a model of the agent

One of the main metaphors in the mentalistic psychology is that the subject's cognition is represented as a structure of images of the self and others inserted into each other together with the relationships between them (Fig. 10.9).

It was shown earlier (Lefebvre 1982) that cartoons of this type may underlie a description of computational processes in human mental domain which are connected with the choice between 'positive' and 'negative' alternatives. In the same work, a concept of a diagonal form was introduced without any special connection to a representation of a polynomial tree. For example, the reflexive structure in Fig. 10.9 corresponds to the following diagonal form:

$$\begin{array}{ccc} & a_4 + a_5 & \\ a_2 & & \bullet\ a_3 \\ a_1 & & \end{array} \tag{10.9}$$

where $a_1$, $a_2$, $a_3$, $a_4$, and $a_5$ are Boolean variables, + is Boolean addition, • is Boolean multiplication, and diagonal operation $a^b$ corresponds to the implication: $b \rightarrow a = a + \bar{b}$. Therefore, a diagonal form of a polynomial tree is

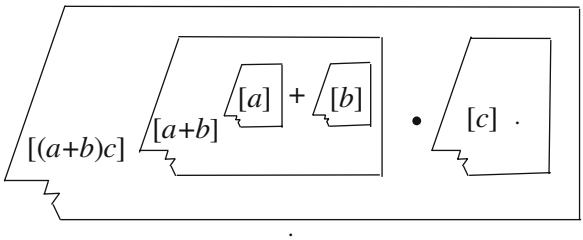



**Fig. 10.10.** Reflexive structure corresponding to the agent's diagonal form.

a *special* case of the diagonal form. We will interpret such a tree as a picture of an agent's reflexive structure. For example, reflexive structure in Fig. 10.10 corresponds to the diagonal form of the polynomial tree (10.5).

In the most general case, a reflexive structure describes a hierarchical system of images of the self and others. We will consider a special case, in which each image is the subject's image of the self. For example, the subject in Fig. 10.10, has two images of the self connected by •. The left image, in turn, has two images of the self connected by +.

Imagine a group of agents immersed in solving a certain problem. Each agent has to make a choice between *active* and *passive* lines of behavior, so each agent faces a choice between two poles. From the point of view of every one of them, each pair of agents has either a + or a • relationship. Therefore, there is a completed graph of relationships between the agents. Let this graph be *decomposable*. We will assume that the diagonal form of the polynomial tree of this graph represents the reflexive structure of an agent from the group. *This statement corresponds to the psychological assumption, that in the agent's mental domain, the process of the graph's decomposition is going on and that process completely determines the reflexive structure.* At the first step, each subgraph corresponds to an image of the self; at the second step, each subgraph corresponds to an image of the self belonging to the images of the self, and so on, until the graph is completely divided into elementary subgraphs.

To make our further consideration more clear, let us look at the diagonal form corresponding to the reflexive structure in Fig. 10.10 and suppose it represents agent $c$:

$$\Phi = \begin{matrix} & & [a]+[b] & \\ & [a+b] & & [c] \\ [(a+b)c] & & & \end{matrix} \tag{10.10}$$

where one of the symbols $(+, \bullet)$ represents a union and the other one a conflict.

The value of each variable is the influence that $c$ experiences from another agent or the self. We designate an agent with the same letter as the corresponding variable. The value of '1' means a pressure toward choosing the

active line of behavior (which is the positive pole), and '0' means a pressure toward the passive line of behavior (the negative pole). For example, $a = 1$ means that $a$ inclines $c$ to choose the active behavior, and $a = 0$ means that $a$ inclines $c$ to choose the passive behavior.

Each polynomial represents an influence of the entire group or its subgroup on the agent or on the agent's image of the self. The value of polynomial $[(a+b)c]$ is the influence of the entire group on agent $c$. The value of polynomial $[a + b]$ is the influence of the group consisting of $a$ and $b$ on the agent's left image of the self. The value of $[c]$ is the influence of the self on the agent's right image of the self. A group consisting of agents $a$ and $b$ is in relation with the group consisting of only one agent $c$. The agent depicted by (10.10) has separate images of the self corresponding to each of these groups. *We assume that the relation between images is the same as the relation between the groups which influence the images.* If the groups are in conflict, then the images under their influence are also in conflict; if the groups are in union, the images are in union.

The value of diagonal form $\Phi = \Phi(a, b, c)$ has the meaning of the agent's choice: $\Phi = 1$ means that $c$ chose the active line of behavior, and $\Phi = 0$ means that $c$ chose the passive line. Having assigned the values of $a, b$, and $c$ influences, we can find the value of $\Phi$.

There is, however, the essential difference between the values of $a$ and $b$, on the one hand, and the value of $c$, on the other. The influence of the other agents can be observed, but self-influence is unobservable. It requires the introduction of a concept of *intention*. So, we assume that variable $c$ represents $c$'s intention; $c = 1$ means that the agent has intention to choose the active line of behavior, and $c = 0$ means that the agent has intention to choose the passive behavior. Thus, we identify intention and self-influence.

Let us make the next step and assume that the agent makes a *conscious* choice, i.e., *the agent's choice always coincides with his intention.* This assumption corresponds to the condition

$$\Phi = c \tag{10.11}$$

Under condition (10.11), equality (10.10) turns into equation:

$$\begin{array}{c} \qquad\qquad [a] + [b] \\ [a + b] \qquad\quad [c] \\ c = [(a + b)c] \end{array} \tag{10.12}$$

Let $a = 1$ and $b = 1$. By substituting these values to (10.12), we obtain $c = 1$. This means that with these influences, agent c chooses the active line of behavior and this choice coincides with the agent's intention.

Consider now a case when $a$ and $b$ incline $c$ to choose the passive line of behavior: $a = 0$ and $b = 0$. Then, we obtain the equation

$$c = \bar{c} \tag{10.13}$$

This equation has no roots; this means that the agent cannot generate the choice which would coincide with his intention. We interpret this fact as evidence that agent $c$ is not capable of making a decision.

In general, an agent can be represented by an equation of the type (10.8). When it has two solutions, 1 and 0, we say that the agent is in the state (I) and has freedom to choose between active and passive lines of behavior. When (10.8) has one solution and it is 1, the agent is in the state (II) and can choose only active behavior, and when the solution is 0, the agent is the state (III) and can choose only passive behavior. Finally, when the equation does not have a solution, we say that the agent is in the state (IV), in which the agent is not able to choose the line of behavior.

If a graph of relationships is not decomposable, we suppose that each agent consecutively excludes other group members from consideration. This procedure is as follows. First, the agent establishes an order of significance on the set of group members (without himself) starting with the least significant one. The agent excludes the least significant member and checks whether the graph becomes decomposable. If it does, the procedure ends, if it does not, the next least significant member is excluded and the graph is checked for decomposition, etc. This procedure always has an end, because any three-node graph is decomposable. The resulting graph serves as a basis for constructing the diagonal form, writing the equation, and finding the agent's state.

In the scheme described, the influences of agents on one another do not depend on the agents' states or on their relationships.

## The ethical systems

In earlier works (Lefebvre 1980, 1982, 2001) we brought arguments in favor of the assumption that in the human cognitive domain, there is a specialized processor which operates with codes 'positive' and 'negative.' The work of the processor can be described in the language of Boolean algebra, where 1 corresponds to 'positive,' and 0 to 'negative.' This processor allows the subject to evaluate the results of integration and disintegration of positive and negative values. We have demonstrated that there are two different systems of evaluation, which were called *the first and the second ethical systems.*

In the first ethical system, the result of integration of the codes 'positive' and 'negative' is coded 'negative,' and the result of disintegration of such codes is coded 'positive.' The first ethical system works, for example, in negative evaluation of using bad means for achieving good goals, and positive evaluation for restraining from such use.

In the second ethical system, the result of integration of the codes 'positive' and 'negative' is coded 'positive,' and the result of disintegration of such codes is coded 'negative.' Joining a good goal with bad means is evaluated positively, and their separation is evaluated negatively.

We have also demonstrated that Boolean operations + and • in the diagonal forms can be interpreted as operations of integration and disintegration or as *union* and *conflict*. In the first ethical system, • corresponds to *union*, and + to *conflict*. In the second ethical system, + corresponds to *union*, and • to *conflict*. Note that all the agents belonging to the same ethical system and involved in a situation with a decomposable graph have the same diagonal form.

In the further analysis of real situations we will take into consideration the agents' ethical systems. The agents connected with western cultural tradition belong to the first ethical system. The agents involved into the life of extremist regimes belong to the second ethical system. We have shown earlier that the Soviet culture was based on the second ethical system. In our current analysis, we assume that in the countries which appeared after the Soviet Union collapsed, the cultures continue to be based on the second ethical system. We ascribe Fascist Germany and Communist countries to the second ethical system as well.

## A scheme of the analysis of a situation

In this section, we describe a scheme of using the concepts introduced in the previous sections to analyze situations, in which several agents interact with each other and everyone faces a choice between active and passive lines of behavior. Some agents may belong to the first ethical system, while others to the second one. The process of the analysis is as follows:

1. Selecting a set of agents with their ethical systems.
2. Defining a graph of the agents' relationships (union, conflict).
3. Constructing a Boolean matrix, $\|\alpha_{ij}\|$, of the agents' influences on each other, where $\alpha_{ij}$ is the influence of agent $i$ on agent $j$; when $i \neq j$, $a_{ij}$ is either 1, or 0; when $i = j = k$, $\alpha_{kk}$ is an unknown value.
4. Check if the graph is decomposable.
5. If it is, the diagonal form is constructed for each agent $k$ (taking into consideration the agent's ethical system) and an equation for $\alpha_{kk}$ is written:

$$\alpha_{kk} = \Phi(\alpha_{1k}, ..., \alpha_{kk}, ..., \alpha_{nk}), k = 1, ..., n \tag{10.14}$$

   where $\alpha_{kk}$ is intention or self-influence. This equation is analyzed and the agent's state is found. Note that (10.14) is just another form of Eq. (10.8).
6. If the graph is not decomposable, then for each agent, a list of other agents in the order of increasing significance for the given agent is compiled. The procedure of sequential elimination is applied and a graph of the relationships specific for the particular agent is found. Then taking into consideration the agent's ethical system, we construct a diagonal form pertinent to the new graph and write an equation similar to (10.14).

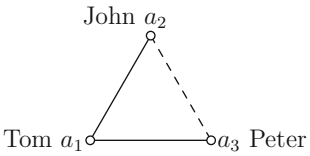



**Fig. 10.11.** A graph of relationships between Tom ($a_1$), John ($a_2$), and Peter ($a_3$).

| | | $a_1$ | $a_2$ | $a_3$ |
|---|---|---|---|---|
| Tom | $a_1$ | $\alpha_{11}$ | 0 | 0 |
| John | $a_2$ | 0 | $\alpha_{22}$ | 1 |
| Peter | $a_3$ | 0 | 1 | $\alpha_{33}$ |

**Fig. 10.12.** A matrix of mutual influences.

Let us consider an example with a decomposable graph. Tom, John, and Peter work on a common project. John and Peter got into a conflict, but Tom keeps good relationship with each of them. Tom pushes John and Peter toward passive behavior; John and Peter incline each other to activity and Tom to passivity. A question: in what state are the members of this group?

Suppose that Tom belongs to the first ethical system, and John and Peter to the second. A graph corresponding to the situation is given in Fig. 10.11, where solid lines designate union, and a dotted line designates conflict. This graph is decomposable. Since the agents in the group belong to different ethical system, we construct diagonal forms for the first and the second ethical systems.

The following form is written for the first ethical system:

$$\begin{matrix} & & [a_2]+[a_3] \\ & [a_1][a_2+a_3] & \\ [a_1(a_2+a_3)] & & \end{matrix} \tag{10.15}$$

And the next one for the second ethical system:

$$\begin{matrix} & & [a_2][a_3] \\ & [a_1]+[a_2a_3] & \\ [a_1+a_2a_3] & & \end{matrix} \tag{10.16}$$

Using the situation description we construct the matrix of mutual influences in Fig. 10.12.

The unknown values of the agents' intentions are given on the diagonal. To write an equation for $a_1$, we use form (10.15) and the first matrix column:

$$\begin{matrix} & & [0]+[0] \\ & [\alpha_{11}][0+0] & \\ \alpha_{11}=[\alpha_{11}(0+0)] & & \end{matrix}$$

After transformation we obtain:

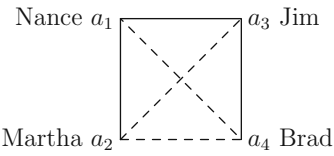



**Fig. 10.13.** A graph of relationships between Nance ($a_1$), Martha ($a_2$), Jim ($a_3$), and Brad ($a_4$).

$$\alpha_{11} = \bar{\alpha}_{11}$$

This equation does not have solutions, therefore, agent $a_1$ is in state (IV).

To write an equation for $a_2$ we use (10.16) and the second matrix column:

$$\begin{array}{r} [\alpha_{22}][1] \\ [0] + [\alpha_{22}1] \\ \alpha_{22} = [0 + \alpha_{22}1] \end{array} \tag{10.17}$$

and obtain after the transformation:

$$\alpha_{22} = \alpha_{22}$$

Therefore, $a_2$ is in state (I).

For agent $a_3$ we use form (10.16) and the third matrix column:

$$\begin{array}{r} [1][\alpha_{33}] \\ [0] + [1\alpha_{33}] \\ \alpha_{33} = [0 + 1\alpha_{33}] \end{array} \tag{10.18}$$

and find

$$\alpha_{33} = \alpha_{33}$$

Therefore, agent $a_3$ is in state (I).

Our analysis allowed us to extract the following information from a description of the situation: John and Peter being in conflict have freedom of choice, while their friend Tom has difficulty making a choice between the active and passive lines of behavior.

Consider now a case in which a graph of relationships is not decomposable. The members of the interaction are Nance, Martha, Jim, and Brad. Nance is in union with Martha and Jim and in conflict with Brad. Martha is in conflict with Jim and Brad. Jim and Brad are in union (see Fig. 10.13). Let all of them belong to the first ethical system.

Let the matrix of mutual influences be as in Fig. 10.14.

We see that the graph of the situation is not decomposable because it is isomorphic to $S_4$. For this reason, in the mental domains of all agents, there is a process of elimination of non-significant members of the group. Let Brad be the least significant for Nance, for Martha it is Nance, for Jim it is Martha,

| | | $a_1$ | $a_2$ | $a_3$ | $a_4$ |
|---|---|---|---|---|---|
| Nance | $a_1$ | $\alpha_{11}$ | 0 | 1 | 1 |
| Martha | $a_2$ | 0 | $\alpha_{22}$ | 0 | 0 |
| Jim | $a_3$ | 0 | 0 | $\alpha_{33}$ | 0 |
| Brad | $a_4$ | 1 | 0 | 1 | $\alpha_{44}$ |

**Fig. 10.14.** The matrix of mutual influences.

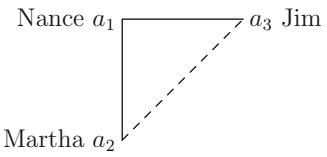



**Fig. 10.15.** Relationships graph for Nance after excluding Brad from consideration.

and for Brad it is Nance. Let us begin with Nance. The least significant for her is Brad, so she excludes him first. It is enough, because any graph with three nodes is decomposable. After Brad is excluded, graph becomes Fig. 10.15.

The following diagonal form corresponds to this graph:

$$\begin{matrix} & & [a_2]+[a_3] \\ & [a_1][a_2+a_3] & \\ [a_1(a_2+a_3)] & & \end{matrix} \tag{10.19}$$

To this form, we substitute the influences on Nance ($a_1$) from Martha ($a_2$) and Jim ($a_3$), given in column ($a_1$) of the table in Fig. 10.14, and obtain the following equation

$$\begin{matrix} & & [0]+[0] \\ & [\alpha_{11}][0+0] & \\ \alpha_{11}=[\alpha_{11}(0+0)] & & \end{matrix} \tag{10.20}$$

It follows from (10.20) that $\alpha_{11} = \bar{\alpha}_{11}$. Nance is in state (IV), in which she cannot choose between active and passive lines of behavior.

Martha excludes Nance from the consideration, resulting in the graph represented in Fig. 10.16.

The diagonal form for this graph is as follows:

$$\begin{matrix} & & [a_3][a_4] \\ & [a_2]+[a_3a_4] & \\ [a_2+a_3a_4] & & \end{matrix} \tag{10.21}$$

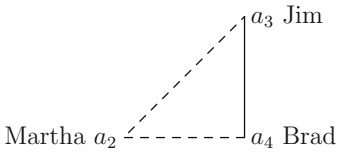



**Fig. 10.16.** Relationships graph for Martha after excluding Nance from consideration.

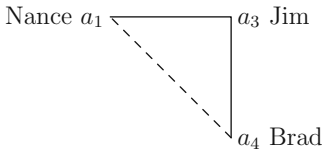



**Fig. 10.17.** Relationships graph for Jim after excluding Martha from consideration.

We substitute Jim's ($a_3$) and Brad's ($a_4$) influences on Martha ($a_2$) from column $a_2$ in Fig. 10.14 to this form and obtain the following equation:

$$\begin{array}{r} [0][0] \\ [\alpha_{22}] + [00] \\ \alpha_{22} = [\alpha_{22} + 00] \end{array} \tag{10.22}$$

It follows from (10.22) that $\alpha_{22} = \alpha_{22}$. Therefore, Martha is in state (I), where she has the freedom of choice between active and passive lines of behavior.

Jim excludes Martha from consideration and his graph looks like Fig. 10.17.

The following diagonal form corresponds to this graph:

$$\begin{array}{r} [a_1] + [a_4] \\ [a_3][a_1 + a_4] \\ [a_3(a_1 + a_4)] \end{array} \tag{10.23}$$

After substitution of the values from column $a_3$ in Fig. 10.14, we obtain the following equation:

$$\begin{array}{r} [1] + [1] \\ [\alpha_{33}][1 + 1] \\ \alpha_{33} = [\alpha_{33}(1 + 1)] \end{array} \tag{10.24}$$

It follows from this equation that $\alpha_{33} = 1$. Jim chooses the active line of behavior. He is in state (II).

Consider now Brad. He excludes Nance from consideration. The relationship graph for Brad is given in Fig. 10.16. It coincides with the graph for Martha. That graph corresponds to the form:

$$\begin{array}{c} [a_3][a_4] \\ [a_2] + [a_3 a_4] \\ [a_2 + a_3 a_4] \end{array} \tag{10.25}$$

and equation

$$\begin{array}{c} [0][\alpha_{44}] \\ [0] + [0\alpha_{44}] \\ \alpha_{44} = [0 + 0\alpha_{44}] \end{array} \tag{10.26}$$

It follows from this equation that $\alpha_{44} = 0$, i.e., Brad is in state (III) and chooses the passive line of behavior. Therefore, all four states are represented in the group.

## Two theorems about compatibility of the states

In the previous section, we analyzed a group in which the four possible states were realized and each agent had its own relationship graph because the initial graph was not decomposable. Let us pose a question: are there such groups, in which all four states are present and the relationship graph is decomposable? Theorem 3 gives us the answer.

The second question, to which Theorem 4 gives the answer, is as follows: is there a group with a decomposable relationship graph, one member of which can be in each of the four different states depending on the influences from other agents?

The proofs of these theorems are done by demonstration of the appropriate examples.

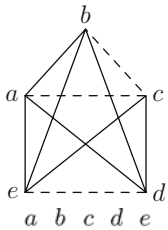
**Theorem 3** *There is a group of agents belonging to the same ethical system and with a decomposable graph of relationships and a matrix of mutual influences such that there are four members in the four different states.*

*Proof* Consider a group of agents with the relationship graph shown in Fig. 10.18 and a matrix of mutual influences shown in Fig. 10.19.

Graph $G$ is decomposable, and the following diagonal form corresponds to each agent:

$$\begin{array}{c} [a][b] \\ [a\ b] \quad +[c] \quad [d] + [e] \\ [a\ b + c] \quad [d + e] \\ [(a\ b + c)(d + e)] \end{array} \tag{10.27}$$

1. Consider agent $d$. Column $d$ in Fig. 10.19. corresponds to the influences on him from other agents and the self. Let us find an equation corresponding to $d$. To do so, we substitute the values of the variables from this column to (10.27) and equate the expression to $\alpha_{44}$:




| | a | b | c | d | e |
|---|---|---|---|---|---|
| a | | • | + | • | • |
| b | • | | + | • | • |
| c | + | + | | • | • |
| d | • | • | • | | + |
| e | • | • | • | + | |

**Fig. 10.18.** Graph and matrix of relationships in a group of five agents. *Solid lines* represent •, and *dotted lines* +.

| | a | b | c | d | e |
|---|---|---|---|---|---|
| a | $\alpha_{11}$ | 0 | 1 | 0 | 0 |
| b | 1 | $\alpha_{22}$ | 1 | 0 | 0 |
| c | 0 | 0 | $\alpha_{33}$ | 1 | 1 |
| d | 0 | 0 | 0 | $\alpha_{44}$ | 1 |
| e | 0 | 0 | 0 | 0 | $\alpha_{55}$ |

**Fig. 10.19.** Matrix of mutual influences.

$$\alpha_{44} = [(0\ 0+1)(\alpha_{44}+0)]\ [0\ 0+1]\ [0][0]\ +[1]\ [\alpha_{44}+0]\ [\alpha_{44}]+[0]$$

As a result, we obtain the equation:

$$\alpha_{44} = \alpha_{44}$$

This equation has two roots, thus $d$ is in the state (I).

2. Let us write the equation for $e$. Column $e$ in Fig. 10.19 corresponds to this agent.

$$\alpha_{55} = [(0\ 0+1)(1+\alpha_{55})]\ [0\ 0+1]\ [0][0]\ +[1]\ [1+\alpha_{55}]\ [1]+[\alpha_{55}]$$

After transformation, we obtain:

$$\alpha_{55} = 1$$

Thus, $e$ is in the state (II).

3. The equation for c is as follows:

$$\alpha_{33} = [(1\ 1 + \alpha_{33})(0 + 0)]^{[1\ 1 + \alpha_{33}]}_{[1\ 1]}{}^{[1][1]}_{[0+0]} + [\alpha_{33}]^{[0]+[0]}$$

We obtain

$$\alpha_{33} = 0$$

Therefore, $c$ is in the state (III).

4. Finally, we write an equation for $a$:

$$\alpha_{11} = [(\alpha_{11}\ 1 + 0)(0 + 0)]^{[\alpha_{11}\ 1 + 0]}_{[\alpha_{11}\ 1]}{}^{[\alpha_{11}][1]}_{[0+0]} + [0]^{[0]+[0]}$$

and obtain

$$\alpha_{11} = \bar{\alpha}_{11}$$

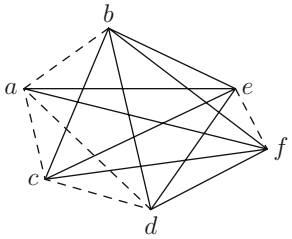
Agent $a$ is in the state (IV).

Therefore, agents $a, b, c$, and $d$ are in the different states.

**Theorem 4** *There is a group of agents with a decomposable relationships graph and four sets of influences on one of the group members such that each of the four different states can be realized in this one agent.*

*Proof* Consider a group of agents in Fig. 10.20. Let $b, c, d, e$, and $f$ have the ability to influence $a$ with four different sets represented in Fig. 10.21. The graph in Fig. 10.20 is decomposable, and agent $a$ corresponds to the following diagonal form:

$$[(a + b(c + d))(e + f)]^{[a+b(c+d)]}_{[a]+[b(c+d)]}{}^{[b][c+d]}_{[e+f]}{}^{[c]+[d]}_{[e]+[f]} \tag{10.28}$$




| | a | b | c | d | e | f |
|---|---|---|---|---|---|---|
| a | | + | + | + | • | • |
| b | + | | • | • | • | • |
| c | + | • | | + | • | • |
| d | + | • | + | | • | • |
| e | • | • | • | • | | + |
| f | • | • | • | • | + | |

**Fig. 10.20.** Relationships graph in a group of six agents and the matrix of their connections.

| | 1 | 2 | 3 | 4 |
|---|---|---|---|---|
| b | 1 | 1 | 1 | 0 |
| c | 0 | 1 | 1 | 1 |
| d | 0 | 0 | 1 | 1 |
| e | 1 | 1 | 0 | 0 |
| f | 0 | 1 | 0 | 0 |

**Fig. 10.21.** The four influences to agent $a$ from agents $b$, $c$, $d$, $e$ and $f$. Each column corresponds to one state of $a$.

1. Let us find $a$'s state when the set of influences is given by column 1 in Fig. 10.21. To do so, we substitute letters $b, c, d, e$, and $f$ in (10.28) with the values from column 1, change variable $a$ to unknown value $\alpha_{11}$, and obtain the following equation:

$$\alpha_{11} = [(\alpha_{11} + 1(0+0))(1+0)]^{[\alpha_{11}+1(0+0)]^{[\alpha_{11}]+[1(0+0)]^{[1][0+0]^{[0]+[0]}}}}_{[1+0]^{[1]+[0]}}$$

   After transformation, we find that $\alpha_{11} = \alpha_{11}$. Thus, the agent is in the state (I).
2. When the set of influences is given in column 2, the equation is

$$\alpha_{11} = [(\alpha_{11} + 1(1+0))(1+1)]^{[\alpha_{11}+1(1+0)]^{[\alpha_{11}]+[1(1+0)]^{[1][1+0]^{[1]+[0]}}}}_{[1+1]^{[1]+[1]}}$$

   Thus $\alpha_{11} = 1$. The agent is in the state (II).

3. With the set of influences from the third column, the equation is

$$\alpha_{11} = [(\alpha_{11} + 1(1+1))(0+0)]^{[\alpha_{11}+1(1+1)]^{[\alpha_{11}]+[1(1+1)]^{[1][1+1]^{[1]+[1]}}}}_{[0+0]^{[0]+[0]}}$$

   Thus $\alpha_{11} = 0$, that is, the agent is in the state (III).
4. The fourth set corresponds to the fourth column in Fig. 10.21, and the equation is

$$\alpha_{11} = [(\alpha_{11} + 0(1+1))(0+0)]^{[\alpha_{11}+0(1+1)]^{[\alpha_{11}]+[0(1+1)]^{[0][1+1]^{[1]+[1]}}}}_{[0+0]^{[0]+[0]}}$$

   giving $\alpha_{11} = \bar{\alpha}_{11}$. The agent is in the state (IV). Thus, agent a can be in any of the four states.

## Superactivity

We will call the agent superactive, if for any set of influences from other agents, he is in the active state. From the formal point of view, this definition means that the equation

$$a_k = \Phi(a_1, ..., a_k, ..., a_n), k \in \{1, 2, ..., n\} \tag{10.29}$$

corresponding to the agent, has the only solution $a_k = 1$ for any set of values $a_1, ..., a_{k-1}, a_{k+1}, ..., a_n$. This definition is equivalent to the condition that, for any set of values $a_1, ..., a_k, ..., a_n$ the following identity holds:

$$\Phi(a_1, ..., a_k, ..., a_n) \equiv 1 \tag{10.30}$$

Let us prove the above statement. It is clear that (10.29) follows from (10.30). We have to show that (10.30) follows from (10.29). Indeed, the two identities which follow from (10.29):

$$\Phi(a_1, ..., a_k = 1, ..., a_n) \equiv 1$$
$$\Phi(a_1, ..., a_k = 0, ..., a_n) \equiv 1$$

are equivalent to (10.30). The latter holds because if for variable $a_k$ and at least one set of values of other variables the following equation held,

$$\Phi(a_1, ..., a_k = 0, ..., a_n) = 0$$

it would mean that the corresponding Eq. (10.29) had two solutions, which would contradict the initial condition.

We will call a group *superactive* if every agent in the group is superactive.

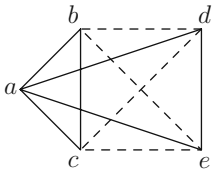



**Fig. 10.22.** Graph corresponding to a group.

Consider the following example. Figure 10.22 shows a graph of a group. Solid lines represent union, and dotted lines represent conflict. Suppose the group includes agents of both the first and second ethical systems.

This graph is decomposable. The following diagonal form corresponds to each agent of the first ethical system:

$$\Phi_1(a,b,c,d,e) = \begin{matrix} & & & & & & [b][c] & & [d][e] \\ & & & & [b\ c] & & + & [d\ e] & \\ & & & [a][b\ c+d\ e] & & & & & \\ [a(b\ c+d\ e)] & & & & & & & & \end{matrix} \tag{10.31}$$

and the next one to each agent of the second ethical system:

$$\Phi_2(a,b,c,d,e) = \begin{matrix} & & & & & & [b]+[c] & & [d]+[e] \\ & & & & [b+c] & & + & [d+e] & \\ & & & [a]+[(b+c)(d+e)] & & & & & \\ [a+(b+c)(d+e)] & & & & & & & & \end{matrix} \tag{10.32}$$

After transformations, we find that

$$\Phi_1(a,b,c,d,e) \equiv 1, \quad \Phi_2(a,b,c,d,e) \equiv 1 \tag{10.33}$$

Therefore, (10.30) holds for each agent independently from the ethical system; thus, Eq. (10.29) has a root equal to 1 with any influences. In other words, every agent is superactive; that is, the group is superactive.

Consider two identities:

$$\begin{matrix} & & [a_1]+...+[a_k]+...+[a_n] \\ [a_1+...+a_k+...+a_n] & & \equiv 1 \end{matrix} \tag{10.34}$$

$$\begin{matrix} & & [a_1]\bullet...\bullet[a_k]\bullet...\bullet[a_n] \\ [a_1\bullet...\bullet a_k\bullet...\bullet a_n] & & \equiv 1 \end{matrix} \tag{10.35}$$

The left-hand sides of these identities are diagonal forms corresponding to agents from *uniform* groups, i.e., such that all of their members are either in union with each other or in conflict. For the first ethical system, (10.34)

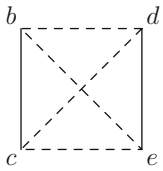



**Fig. 10.23.** Conflicting groups. The *solid lines* represent union, and the *dotted lines* represent conflict.

represents conflict and (10.35) union. For the second ethical system, the representation is opposite. Groups with uniform relationships can be called *crowd* where the activity is self-generating.

In superactive groups, the agents' states do not depend on their influences on each other. The states are predetermined by the graphs of the agents' relationships. We may hypothesize that some *outbursts of activity* which reveal themselves as political or social crises are connected with such graphs.

## A paradox of a peacemaker

It is known that the attempts to reconcile conflicting groups lead, sometimes, to increasing tension instead of detente. Let us analyze a process of a reconciler's involvement in the situation. Figure 10.23 shows a relationship graph between four members of a conflict. Agents $b$ and $c$ are in union with each other and in conflict with agents $d$ and $e$ who are also in union with each other.

For the agent of the first ethical system, the diagonal form is as follows:

$$\begin{array}{ccc} & [b][c] \quad [d][e] & \\ [b\ c] & +[d\ e] & \\ [bc+de] & & \not\equiv 1 \end{array} \tag{10.36}$$

The form is not identically equal to one, because for $b = 0$ and $d = 0$ it is equal to 0.

For the agent of the second ethical system the diagonal form is:

$$\begin{array}{ccc} & [b]+[c] \quad [d]+[e] & \\ [b+c] & [d+e] & \\ [(b+c)(d+e)] & & \not\equiv 1 \end{array} \tag{10.37}$$

The form is not identically equal to one, because for $b = 0$ and $c = 0$ it is equal to 0.

Therefore, the members of the subgroups being in conflict are not superactive.

Let a reconciler $a$ appear and become an ally of each agent. The resulting situation was analyzed previously. The graph in Fig. 10.22 corresponds to this new situation, and the diagonal forms (10.31) and (10.32) correspond to the agents. We see that after a reconciler appeared, the agents became superactive, independently from the ethical system. A move to superactivity may mean an outburst of confrontation between the agents. Thus, the reconciler's intervention led to aggravation of the conflict and an increase in tension.

Let us pose a question: in the framework of the theory being developed in this work, may a reconciler reconcile? Yes, it is possible. Let us consider a conflict between two agents instead of two conflicting groups; the agent from the first ethical system corresponds to the diagonal form

$$\begin{matrix} & [a]+[b] \\ [a+b] & \equiv 1 \end{matrix} \tag{10.38}$$

and the agent from the second ethical system to the diagonal form

$$\begin{matrix} & [a][b] \\ [ab] & \equiv 1 \end{matrix} \tag{10.39}$$

We see that independently from the ethical system, the agents are in a superactive state. Now, a peacemaker $c$ appears and sets the relationship of union with each of the conflicting agents $a$ and $b$. The agent of the first ethical system corresponds to inequality

$$\begin{matrix} & & [a]+[b] \\ & [c][a+b] & \\ [c(a+b)] & & \not\equiv 1 \end{matrix} \tag{10.40}$$

and the agent from the second ethical system to inequality

$$\begin{matrix} & & [a][b] \\ & [c]+[a\ b] & \\ [c+ab] & & \not\equiv 1 \end{matrix} \tag{10.41}$$

The agents are not superactive any longer, so, the reconciler fulfilled his task.

*The success or failure of a reconciler depends on whom he tries to reconcile; conflicting individuals or conflicting groups. In the first case, his work is successful, in the second, it is not.*

This statement was concluded from the four general correlations given below.

There are several groups of agents, any two of which are in conflict, but inside each of the groups the agents are in union.

Every agent from the first ethical system corresponds to the diagonal form in Fig. 10.24(a), where $k_i \geq 2, i = 1, ..., n, n \geq 2$..

After a peacemaker $z$ intervenes, the form changes to that in Fig. 10.24(b).

$$[(a_1...a_{k1}) + ... + (r_1...r_{kn})]^{[a_1...a_{k1}]^{[a_1]...[a_{k1}]} + ... + [r_1...r_{kn}]^{[r_1]...[r_{kn}]}} \not\equiv 1$$

(a)

$$[z[(a_1...a_{k1}) + ... + (r_1...r_{kn})]]^{[a_1...a_{k1}]^{[a_1]...[a_{k1}]} + ... + [r_1...r_{kn}]^{[r_1]...[r_{kn}]}} \equiv 1$$

(b)

$$[(a_1 + ... + a_{k1})...(r_1 + ... + r_{kn})]^{[a_1 + ... + a_{k1}]^{[a_1] + ... + [a_{k1}]} ...[r_1 + ... + r_{kn}]^{[r_1] + ... + [r_{kn}]}} \not\equiv 1$$

(c)

$$[z + (a_1 + ... + a_{k1})...(r_1...r_{kn})]^{[z] + [(a_1 + ... + a_{k1})...(r_1 + ... + r_{kn})]^{[a_1 + ... + a_{k1}]^{[a_1] + ... + [a_{k1}]} ...[r_1 + ... + r_{kn}]^{[r_1] + ... + [r_{kn}]}}} \equiv 1$$

(d)

**Fig. 10.24.** Diagonal forms for the peacemaker analysis. (**a**) The diagonal form for the first ethical system before the peacemaker intervenes. (**b**) After the peacemaker intervenes. (**c**) The initial form for the second ethical system. (**d**) After the peacemaker intervenes.

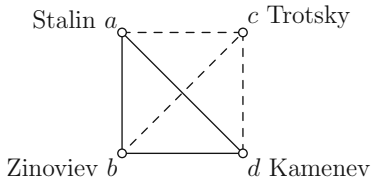



**Fig. 10.25.** Political disposition in the leadership of the Soviet Communist party in 1923–24 (*solid lines* mean union, *dotted lines* mean conflict).

| | | $a$ | $b$ | $c$ | $d$ |
|---|---|---|---|---|---|
| Stalin | $a$ | $\alpha_{11}$ | 1 | 0 | 1 |
| Zinoviev | $b$ | 0 | $\alpha_{22}$ | 0 | 1 |
| Trotsky | $c$ | 1 | 1 | $\alpha_{33}$ | 1 |
| Kamenev | $d$ | 0 | 1 | 0 | $\alpha_{44}$ |

**Fig. 10.26.** Matrix of influences.

We see that before the peacemakers' intervention, the participants of the conflict were not superactive, but after that they become superactive.

Let us conduct a similar analysis for the second ethical system. Before the peacemakers' appearance, each agent corresponds to the diagonal form in Fig. 10.24(c).

After the peacemaker intervention, the form changed to that in Fig. 10.24(d).

Therefore, in the second ethical system, intervention of a reconciler also results in superactivity of the agents which may lead to the aggravation of the conflict.

## Examples of situations' analysis

Let us consider the conflict between Leo Trotsky, on the one side, and Joseph Stalin, Gregory Zinoviev, and Leo Kamenev, on the other, which took place in the Soviet Union in 1923–24. We find its description in Volkogonov's book (1989). It was Stalin who initiated this conflict. The very fact that Trotsky occupied a high position in the party leadership mobilized others to fight with him for power. Stalin was pushing Zinoviev and Kamenev to take actions against Trotsky, and the latter was demoralized by Stalin's activity (in the 1930s, Stalin killed Zinoviev and Kamenev, as well). In spite of their conflict with Trotsky, the two made influence toward actions only on each other. Figure 10.25 represents the described situation.

The situation description given above allows us to construct a matrix of mutual influences as in Fig. 10.26.

Stalin ($a$) induces Zinoviev ($b$) and Kamenev ($d$) to actions and makes oppressive influence on Trotsky ($c$). Zinoviev ($b$) stimulates only Kamenev ($d$),

and Kamenev ($d$) stimulates only Zinoviev ($b$). Trotsky ($c$) stimulates all others: Stalin ($a$), Zinoviev ($b$) and Kamenev ($d$) to activity.

We assume that all participants belong to the second ethical system. Their corresponding diagonal form is as follows:

$$\begin{matrix} & & [a]+[b]+[d] \\ & [c][a+b+d] & \\ [c(a+b+d)] & & \end{matrix}$$

We compose equations and investigate them.

(1) For $a$: $\alpha_{11} = \alpha_{11}$. This equation has two roots, 1 and 0. Agent $a$ is in the state (I), in which he is free to choose either active or passive line of behavior.
(2) For $b$: $\alpha_{22} = 1$. The agent is in the state (II). He can choose only the active line of behavior.
(3) For $c$: $\alpha_{33} = \bar{\alpha}_{33}$. The agent is in the state (IV). He cannot choose any line of behavior.
(4) For $d$: $\alpha_{44} = 1$. The agent is in the state (II). He can choose only the active line of behavior.

The analysis above demonstrates that agent $a$ (Stalin) is able to choose either an active or passive line of behavior. This corresponds to the real profile of Stalin's behavior in the analyzed situation: he alternated soft moves with sudden hits. Agents $b$ and $d$ (Zinoviev and Kamenev respectively) can choose only the active line of behavior. In reality, Zinoviev and Kamenev did not show any flexibility, they only attacked Trotsky. Finally, agent $c$ (Trotsky) cannot make a decision. This also coincides with reality: Trotsky has not made any decisive action to preserve his power. It is practically impossible to find any particular line, active or passive, in his behavior. Volkogonov called his position 'gutta-percha' (Volkogonov 1989, 180).

Further we will use the model developed in this work to analyze a few international crises. In the framework of this analysis, an agent is not an individual person but a country or a group of countries. In this way, we assume that the activity of social systems can be described with concepts and models that appeared in the framework of psychological analysis of individual behavior.

A. Consider the situation during the Finland-Soviet war of 1940. On the one hand, there was German-Soviet collaboration, on the other, Germany did not encourage Soviet aggression toward Finland, which did not call Germany for help for fear of losing independence. The Finnish outburst of patriotism helped to stop the Soviet army and Finland was not occupied. The reason for this war was that the Soviet Union feared a German offensive and wanted to move the Soviet-Finland border farther from Leningrad. A graph of relationships between Finland, Germany, and the Soviet Union is given in Fig. 10.27, and their matrix of mutual influences in Fig. 10.28.

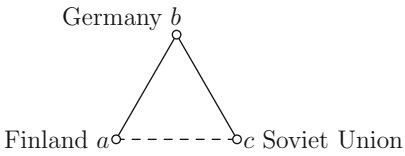



**Fig. 10.27.** Relationships between the countries during the Finland-Soviet war of 1940.

| | | $a$ | $b$ | $c$ |
|---|---|---|---|---|
| Finland | $a$ | $\alpha_{11}$ | 0 | 0 |
| Germany | $b$ | 1 | $\alpha_{22}$ | 1 |
| Soviet Union | $c$ | 1 | 1 | $\alpha_{33}$ |

**Fig. 10.28.** Matrix of influences.

Finland ($a$) does not push toward activity Germany ($b$), nor the Soviet Union ($c$). Germany ($b$) pushes toward activity both Finland ($a$), and the Soviet Union ($c$), which pushes toward activity Finland ($a$) and Germany ($b$).

We assume that Finland belongs to the first ethical system, and Germany and the Soviet Union to the second. In the first ethical system, the situation is presented by the diagonal form in (10.42).

$$\begin{matrix} & & [a]+[c] \\ & [b][a+c] & \\ [b(a+c)] & & \end{matrix} \tag{10.42}$$

The equation for $a$ is $\alpha_{11} = 1$. This means that Finland chooses the active line of behavior.

In the second ethical system, the situation is represented by the form:

$$\begin{matrix} & & [a][c] \\ & [b]+[a\ c] & \\ [b+a\ c] & & \end{matrix} \tag{10.43}$$

The equation for $b$ is $\alpha_{22} = \alpha_{22}$; that is, Germany is capable of both active and passive behavior. It has the freedom of choice. The equation for $c$ is $\alpha_{33} = 1$; thus, the Soviet Union chooses the active line of behavior.

B. Let us analyze now a zone of contemporary international relationships related to North Korea's desire to possess nuclear weapons. The countries which are most involved into this problem are South Korea, USA, China, Russia, and Japan. The graph of their relationships in the Spring of 2006 is given in Fig. 10.29.

There are two conflicting groups: one consists of USA, Japan, and South Korea, and the second of China and North Korea. Russia takes a special

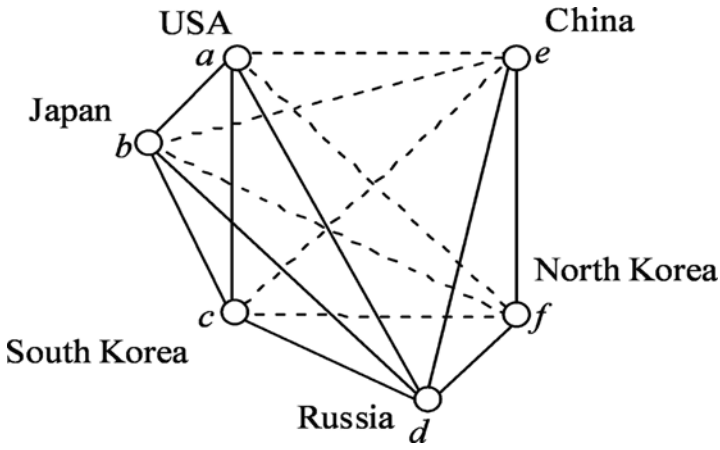



**Fig. 10.29.** Zone of tension in the Korean Peninsula.

position: it maintains the relationship of cooperation with all other participants of the interaction.
We consider that USA, Japan, and South Korea belong to the first ethical system, and China, North Korea, and Russia to the second.

The diagonal form in (10.44) corresponds to the first ethical system:

$$\begin{array}{l} \qquad\qquad\qquad\qquad\quad [a][b][c] \qquad [e][f] \\ \qquad\qquad\qquad\quad [abc] \qquad\quad +[ef] \\ \qquad\qquad\quad [d][abc+ef] \\ [d(abc+ef)] \qquad\qquad\qquad\qquad\quad \equiv 1 \end{array} \tag{10.44}$$

This identity arises from the more general equation (Fig. 10.24(b)).

The second ethical system is represented by the following form:

$$\begin{array}{l} \qquad\qquad\qquad\qquad\quad [a]+[b]+[c] \qquad [e]+[f] \\ \qquad\qquad\qquad\quad [a+b+c] \qquad\quad [e+f] \\ \qquad\qquad\quad [d]+[(a+b+c)(e+f)] \\ [d+(a+b+c)(e+f)] \qquad\qquad\qquad\qquad\quad \equiv 1 \end{array} \tag{10.45}$$

This identity arises from Fig. 10.24(d).

It follows from the above forms that all agents are superactive independently from the ethical system. Thus, the situation in this region is in risk of an outburst. There is no need to investigate how the agents influence each other; with any set of influences, every agent is in the active state. The entire information about the states of the agents is in the graph in Fig. 10.29.

C. Consider the situation that had formed in the world by 1941. Germany was in conflict with England, which was supported by USA. The Soviet

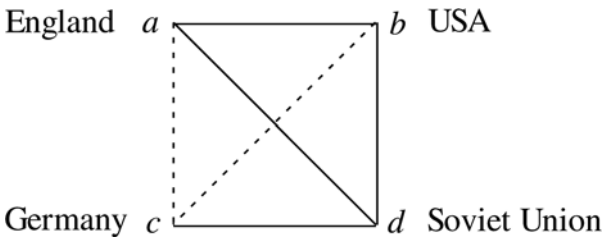



**Fig. 10.30.** Relationships between the countries in 1941.

Union signed a pact of non-aggression with Germany, but at the same time it kept relatively friendly relationships with England and USA. Germany stimulated all other countries to activity. England and USA stimulated each other to activity, but did not stimulate Germany or the Soviet Union, which did not stimulate England or the USA, but indirectly stimulated Germany because they seemed to be easy prey.

We assume that England and the USA belong to the first ethical system, and Germany and the Soviet Union to the second. The situation is described by the following graph in Fig. 10.30.

The following form corresponds to the situation in the first ethical system:

$$\begin{array}{rrr} & & [a][b] \\ & [c]+[a\ b] & \\ [d][c+a\ b] & & \\ [d(c+a\ b)] & & \equiv 1 \end{array} \tag{10.46}$$

We see that the diagonal form for the agents in the first ethical system (England and USA) is identically equal to 1, which means that they are *superactive* and their states are invariant to the pressure changes from the other countries.

In the second ethical system, the diagonal form is as follows:

$$\begin{array}{rrr} & & [a]+[b] \\ & [c][a+b] & \\ [d]+[c(a+b)] & & \\ [d+c(a+b)] & & \not\equiv 1 \end{array} \tag{10.47}$$

To find the states of the countries which belong to the second ethical system (Germany and the Soviet Union), we have to construct the matrix of the influences according to the description given above (Fig. 10.31).

By substituting the values from column $c$ to diagonal form (10.47), we obtain the equation $a_{33} = 1$. Thus, Germany is in the active state. To find the state of the Soviet Union we substitute the values from column $d$ to form (10.47) and obtain equation $a_{44} = 1$. Thus, the Soviet Union is also in the active state.

| | | $a$ | $b$ | $c$ | $d$ |
|---|---|---|---|---|---|
| England | $a$ | $\alpha_{11}$ | 1 | 0 | 0 |
| USA | $b$ | 1 | $\alpha_{22}$ | 0 | 0 |
| Germany | $c$ | 1 | 1 | $\alpha_{33}$ | 1 |
| Soviet Union | $d$ | 0 | 0 | 1 | $\alpha_{44}$ |

**Fig. 10.31.** Matrix of influences in 1941.

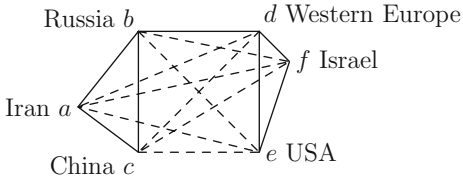



**Fig. 10.32.** The Iranian crisis.

Therefore, at the beginning of 1941 all participants of the interaction, England, USA, Germany, and the Soviet Union were in the active states, and the former two were superactive.

D. At the time of this writing (June 6, 2006), the Iranian crisis is rapidly developing. The relationship graph of the countries involved in it is illustrated in Fig. 10.32.

The graph in Fig. 10.32 is *not decomposable* due to the fact that subgraph $\{c, b, d, e\}$ is isomorphic to $S_{(4)}$. Thus, each agent must exclude some of other agents from the consideration.

Let the significance of the others for USA be increasing in the following order: China, Russia, Western Europe, Iran, and Israel. The least important agent for USA is China, and most important is Israel. The cognitive process of USA begins with excluding China from consideration. As a result, the initial graph transforms into a graph with five nodes: $\{a, b, d, e, f\}$. This graph is not decomposable, either. Then, USA excludes Russia from consideration and obtains a four-node graph which is decomposable:

We assume that USA belongs to the first ethical system. The following equation corresponds to the graph in Fig. 10.33:

$$e = [a + def] \quad \begin{matrix} [a] + [def] \\ [d][e][f] \end{matrix} \tag{10.48}$$

Western Europe ($d$) inclines USA to passivity ($d = 0$), Iran ($a$) pushes USA toward activity ($a = 1$), and Israel ($f$) inclines USA to activity ($f = 1$). Taking variable $e$ as an unknown value we obtain equation $e = 1$. Thus, USA is in the state (II), in which they choose the active line of behavior.

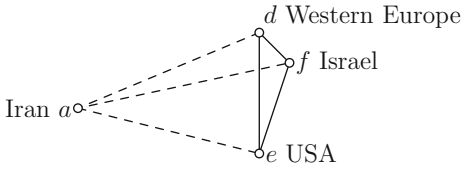



**Fig. 10.33.** Graph with which USA operates.

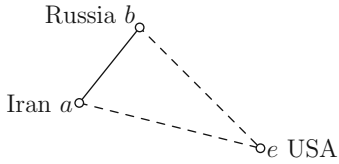



**Fig. 10.34.** Graph with which Russia operates.

For Russia, let the significance of other countries involved in this crisis rise as follows: Israel, China, Western Europe, Iran, and USA. The least important agent in this crisis is Israel, and the most important is USA. The cognitive process for Russia begins with excluding Israel from consideration and results in the graph $\{a, b, c, d, e\}$, which is not decomposable. Then, China is excluded and the graph $\{a, b, d, e\}$ appears, which is not decomposable, either. Exclusion of Western Europe results in the three-node graph $\{a, b, e\}$, as in Fig. 10.34.

The graph in Fig. 10.34 is decomposable, so there is no need to exclude any more agents from consideration. We assume that Russia belongs to the second ethical system. The following equation corresponds to the graph in Fig. 10.34:

$$b = [e(a+b)]\begin{matrix}[a+b]\\ [e][a+b]\end{matrix} \tag{10.49}$$

Iran ($a$) and USA ($e$) incline Russia to activity, so $a = 1$ and $e = 1$. By substituting these values to (10.49) and considering $b$ an unknown value, we obtain equation $b = 1$. Thus, Russia is also in the state (II), i.e., it chooses the active line of behavior.

For Iran, the list of agents in the order of increasing significance is as follows: Western Europe, China, Russia, Israel, and USA. Western Europe is the least important, and USA the most important. The process of excluding begins with Western Europe and results in the graph $\{a, b, c, e, f\}$, which is decomposable (Fig. 10.35).

We believe that Iran belongs to the second ethical system. The following equation corresponds to the graph in Fig. 10.35.

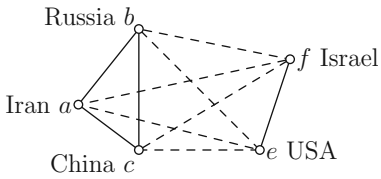



**Fig. 10.35.** Graph with which Iran operates.

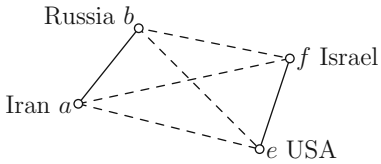



**Fig. 10.36.** Graph with which Israel operates.

$$a = \begin{matrix} [(a+b+c)(e+f)] \\ [a+b+c] \end{matrix} \begin{matrix} [a]+[b]+[c] \\ (e+f) \end{matrix} \begin{matrix} [e]+[f] \\ \end{matrix} \tag{10.50}$$

Russia ($b$), China ($c$), and USA ($e$) push Iran toward passivity ($b = 0, c = 0, and\ e = 0$) and Israel ($f$) stimulates Iran to activity ($f = 1$). By considering variable $a$ as an unknown value, we obtain equation $a = a$. Thus, Iran is free to choose either the active or passive line of behavior; it is in the state (I).

For Israel, the significance of the other agents increases in the following order: China, Western Europe, Russia, Iran, and USA. Cognitive process for Israel begins with excluding China from the consideration that results in the graph $\{a, b, d, e, f\}$. It is not decomposable, so Western Europe is excluded. A new graph is shown in Fig. 10.36.

Let us assume that Israel belongs to the first ethical system. Then, it corresponds to the following equation:

$$f = \begin{matrix} [ab+ef] \\ [ab] \end{matrix} \begin{matrix} [a][b] \\ +\,[ef] \end{matrix} \begin{matrix} [e][f] \\ \end{matrix} \tag{10.51}$$

Russia inclines Israel toward passivity ($b = 0$). Iran inclines Israel toward activity ($a = 1$). USA inclines Israel toward passivity ($e = 0$). Taking variable $f$ as an unknown value, we obtain equation $f = 0$. Thus, Israel is in the state (III) and chooses the passive line of behavior.

## Isolated agents

The model introduced in the previous sections requires knowledge about agents' influence on each other. There are situations, however, when the agent is isolated from other agents' influences. The model can be extended to this case by assuming that for an isolated agent, the variables corresponding to the other agents take on the values 1 and 0, with the probability 0.5, and they are independent from each other. We can now find the probability of each state (I, II, III, and IV) appearance and represent them in a table:

| I | II | III | IV |
|---|---|---|---|
| $p_1$ | $p_2$ | $p_3$ | $p_4$ |

where $p_1$, $p_2$, $p_3$, and $p_4$ are probabilities of the corresponding states ($p_1 + p_2 + p_3 + p_4 = 1$). Distribution $[p_1\ p_2\ p_3\ p_4]$ will be called a *spectrum of the agent's states*.

As an example, we will find a spectrum for agent $a$, represented by the following equation:

$$\begin{array}{c} \qquad\qquad [a]+[b] \qquad\quad [c]+[d] \\ [a+b] \qquad\quad [c+d] \\ a=[(a+b)(c+d)] \end{array} \tag{10.52}$$

After transformations, it looks like:

$$a = (a+b)(c+d) \tag{10.53}$$

Then, by varying the values of $b$, $c$, and $d$, we find an equation for each set of values and find the agent's state, for example, for $b = 0$, $c = 1$, and $d = 0$, Eq. (10.53) becomes

$$a = a$$

which means that agent $a$ is in the state (I). Figure 10.37 shows all $a$'s states for different values of variables $b$, $c$, and $d$:

Out of eight equiprobable combinations of $b$, $c$, and $d$ values, state (I) appears three times, state (II) also three times, state (III) two times, and state (IV) not a single time. Thus, agent $a$ corresponds to the following spectrum: $[\frac{3}{8}\ \frac{3}{8}\ \frac{2}{8}\ 0]$. If the agent is superactive, then with any set of the variables' values he is in the state (II); therefore, he corresponds to spectrum [0 1 0 0].

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| $b$ | 1 | 1 | 1 | 1 | 0 | 0 | 0 | 0 |
| $c$ | 1 | 1 | 0 | 0 | 1 | 1 | 0 | 0 |
| $d$ | 1 | 0 | 1 | 0 | 1 | 0 | 1 | 0 |
| S | II | II | II | III | I | I | I | III |

**Fig. 10.37.** Relation between the states and the sets of variables' values.

## Conclusion

How can the model introduced in this chapter be tested? It is fed by information obtained from experts, and its evaluation should be done by the experts, as well. They must choose the situations whose outcome is already known. Based on his or her image of the situation, an expert supplies the following information to the model: (1) a list of the agents with their ethical systems; (2) a graph of the relationships between them; (3) a matrix of mutual influences; (4) the order of significance of 'others' (for each agent). On these data the model computes the agents' states. The essence of the testing is that an expert interprets the agents' states, received from the model in the terms of his or her image of the situation and evaluates the model's production as *satisfactory* or *unsatisfactory*. A statistically significant 'satisfactory' evaluation means that the experts accept the model. The absence of such an evaluation means that the experts reject it.

Ideally, this experiment has to be conducted with a large number of experts in various areas of group behavior. Practically, it is close to impossible to gather all of them in the framework of one project. A real method of testing described here is the method which has already been taken by the game theory. From the moment of its appearance, it was clear that there was no small number of the experiments which could falsify it. So it was used *before* testing, or more precisely, its use was the process of verification spread over decades. As a result, certain areas were selected where the game theory predictions were evaluated as successful and others where its use was not found effective. As a rule, the researchers have not declared their evaluations of success or its absence in the use of the game theory. In case of success, they kept using it in a given area; in case of failure, they have forgotten about it. Similarly, the model described in this work can be tested. If people do not lose the interest in it after starting its use, this would mean that the theory successfully passed the testing. Speaking more generally, the model's *survival* may be considered as evidence of its *acceptability* for analysis and predictions.

## References

Adams-Webber, J. 1997. Self-reflexion in evaluating others. *American Journal of Psychology* 119:527–541.

Batchelder, W. H., and V. A. Lefebvre. 1982. A mathematical analysis of a natural class of partitions of a graph. *Journal of Mathematical Psychology* 26:124–148.

Grice, J. W., B. L. McDaniel, and D. Thompsen. 2005. Testing an algebraic model of self-reflexion. *Perceptual and Motor Skills* 100:1036–1048.

Lefebvre, V. A. 1980. An algebraic model of ethical cognition. *Journal of Mathematical Psychology* 22:2.

Lefebvre, V. A. 1982. *Algebra of Conscience*. Dordrecht: Reidel.

Lefebvre, V. A. 2001. *Algebra of Conscience*, 2nd enlarged edition (includes a new second part, 'Moral Choice,' not published before). Dordrecht: Kluwer.

Lefebvre, V. A., V. D. Lefebvre, and J. Adams-Webber. 1986. Modeling an experiment on construing self and others. *Journal of Mathematical Psychology* 30:3.

Volkogonov, D. A. 1989. *Triumf i tragedia. Politichesky portret I.V.Stalina*. Moscow: APN Press.

# 11

# Evaluating Self-Reflexion Analysis Using Repertory Grids

James Grice* and Brenda L. McDaniel

**Summary.** Vladimir Lefebvre's Algebraic Model of Self-Reflexion is a formal and flexible system for modeling the processes that underlie human judgments. The model has received notable empirical support from studies within personal construct psychology using the repertory grid approach. In this chapter we briefly review this research and the repertory grid method, and then report the results of two new studies on the model conducted in our laboratory. Lastly, we will briefly discuss the implications of our findings with regard to future research on self-reflexion.

## Introduction

Vladimir Lefebvre's (1990, 1992, 2001) Algebraic Model of Self-Reflexion (AMS-R) is a formal and flexible system for modeling the processes that underlie human judgments. Indeed, Lefebvre's approach has been successfully used in domains as diverse as clinical psychology (Schwartz 1997; Schwartz et al. 2002), psychophysics (Lefebvre 2002), and literary analysis (Bogatyrev 1998). It has been within the realm of *personal construct psychology* (Kelly 1955), however, that AMS-R has received some of its most impressive empirical support from studies employing the repertory grid (Adams-Webber 1990, 1998; Grice, McDaniel, and Thompsen 2005; Lefebvre, Adams-Webber, and Lefebvre 1986). In this chapter we will briefly review this research, as well as the repertory grid method, and then report the results of two unpublished studies conducted in our laboratory. Lastly, we will briefly discuss the implications of our findings with regard to future studies of Lefebvre's AMS-R.

## Repertory grid research

The repertory grid technique was first introduced by George A. Kelly in his two volume book *The Psychology of Personal Constructs* (1955). The original

* Correspondence should be addressed to James Grice.

S. Argamon, N. Howard (eds.), *Computational Methods for Counterterrorism*, DOI 10.1007/978-3-642-01141-2_11,

technique was developed as a tool for clinical psychologists that is administered in three, integrated steps. First, the therapist asks the client to elicit the names of approximately 20 people who fit specific role titles (e.g., a romantic partner, a disliked person, a respected teacher). Second, the therapist prompts the client to elicit his or her own *personal constructs* that are used to make sense of oneself as well as the other 20 named people. In Kelly's theory, personal constructs are bipolar dimensions of discrimination (e.g., happy–sad, trustworthy–untrustworthy) that are arranged hierarchically to form an individual's unique *personal construct system*. Third, the client rates himself or herself and the 20 people on bipolar scales constructed from the personal constructs. This three-step process thus results in a 2-dimensional matrix, or grid, of binary data that can be subjected to a host of statistical analyses.

Over the years, Kelly's technique has seen many modifications so that the term "repertory grid" has truly come to refer to a body of techniques that produce 2-dimensional matrices of rating data. It is quite common, for instance, to find personal construct psychologists providing a list of bipolar adjectives to clients (or participants in a study) rather than eliciting them uniquely from each person, or to find therapists and researchers employing multi-point rating scales on which judgments of self and others are made. The former modification has particularly been used in studies of Lefebvre's AMS-R. In a landmark paper by Lefebvre et al. (1986), for instance, participants in two studies rated themselves and other individuals on bipolar scales constructed from adjective pairs (e.g., generous–stingy, pleasant–unpleasant) taken from Osgood et al. (1957) semantic differential. Studies using similar methods have been conducted by Adams-Webber (1998) and Grice et al. (2005). In all of these studies, the AMS-R was used—with mixed success—to predict the frequencies which participants chose the positive adjective terms (e.g., generous, pleasant) as descriptions of themselves and others in the repertory grids.

A detailed exposition of how the AMS-R was used in each of these previous studies is beyond the scope of this chapter, but a general description of the approach is nonetheless possible and necessary. Consider a reflexive agent (i.e., a person) in dialogue with another person, as shown in Fig. 11.1. The largest head in the figure represents the reflexive agent. Within the reflexive agent's consciousness is an image of self ($a_2$) and an image of the other person ($b_2$). These two images represent the agent's immediate, non-reflective mental constructions of self and of the other person. In other words, these images represent reality from the reflexive agent's point of view. Subordinate to these images are the agent's reflexive image of self ($a_3$) and the other person ($b_3$).

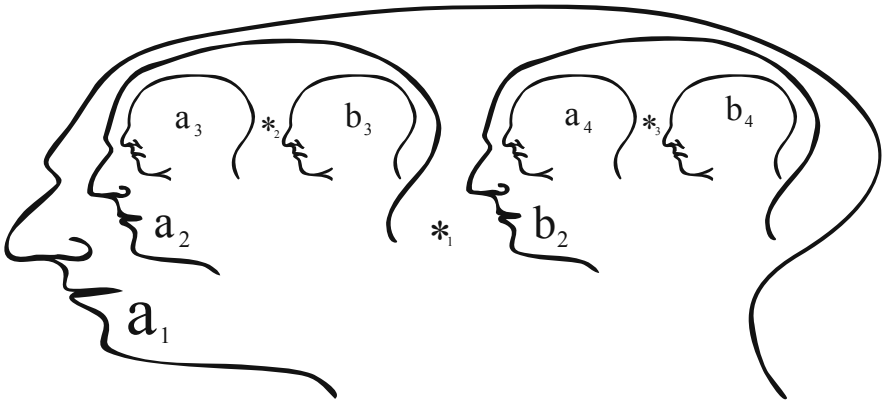



**Fig. 11.1.** Pictorial representation of Lefebvre's algebraic model of self-reflexion. Adapted from Lefebvre (2001, 10).

It is with these images that the agent is said to be "aware" or "conscious" of self and of the other person. The notion of awareness if also conferred upon the image of the other person who is construed as having a reflexive image of the agent ($a_4$) and a reflexive image of his or her self ($b_4$).

An example will help to further clarify the components of the model. Consider a content husband as the reflexive agent in conversation with his disgruntled wife (who he naturally recognizes as another reflexive agent). Referring to Fig. 11.1, the husband constructs a mental image of his wife ($b_2$) that is in relation to his constructed image of himself ($a_2$). In his mind's eye he may reflect on (i.e., be conscious of) himself ($a_3$) as being content and also reflect on his wife as being disgruntled ($b_3$). He may also imagine, in his mind's eye, that his wife sees him as content ($a_4$) and sees herself as disgruntled ($b_4$). Finally, the husband may form an immediate, non-reflective image of the relationship ($\star_1$) as one of cooperation or conflict, he may reflect on the relationship ($\star_2$) as one of cooperation or conflict, and he may imagine that his wife is reflecting on the relationship ($\star_3$) as one of cooperation or conflict. The asterisks thus represent a constructed image of the relationship between the two persons from the reflexive agent's point of view.

Stripping away the figure and removing several of the parameters (i.e., the letters with subscripted numbers) from the model, yields the following equation that serves as the base model for many repertory grid studies of Lefebvre's AMS-R:

$$\begin{aligned} & a_3 \qquad\quad a_4 \\ & a_2 \cdots \star \cdots b_2 \\ A_1 \equiv\ & a_1 \end{aligned} \tag{11.1}$$

The lowercase letters are variables that can take on real values between 0 and 1. The asterisk is strictly a Boolean variable that can represent

cooperation (•) or conflict (⊕). This hierarchical equation (it is not a type of power function) is solved using Lefebvre's (2001) *gamma algebra*, which yields a single real value, $A_1$, that ranges between 0 and 1. A program for Windows was recently written by Grice (2006) that greatly simplifies the computations for solving the reflexive models common to repertory grid researchers. When a participant is rating himself or herself in a repertory grid study, he or she is the reflexive agent and the other person ($b_2$) in the model is considered to be the watchful eye of the experimenter. Also, the reflexive agent is assumed to be in a state of cooperation with the experimenter so that the $\star$ Boolean variable is set equal to •. Applying (11.1) to multiple repertory grid responses obtained from multiple individuals would yield a result for $A_1$ that indicates the expected proportion of positive self ratings (i.e., the frequency which individuals select the positive adjective terms as valid descriptors of themselves).

A similar base model can be used to predict the proportions which individuals will rate other people positively in the repertory grid; namely,

$$\begin{array}{rl} & a_3 \cdots \star \cdots b_3 \qquad\qquad a_4 \cdots \star \cdots b_4 \\ a_2 \cdots\cdots\cdots & \star \cdots\cdots\cdots b_2 \\ A_1 \equiv a_1 & \end{array} \tag{11.2}$$

The experimenter is dropped from the model altogether when the participant is rating other people, and the parameters denoted by the letter "b" now represent the particular person being rated in the grid on the bipolar scales. Applying (11.2) to multiple repertory grid responses obtained from numerous individuals would yield a result for $A_1$ that indicates the expected proportion of positive other ratings, that is, the frequency with which individuals select the positive adjective terms as valid descriptors of other people in the grid.

Using models like those found in (11.1) and (11.2), Lefebvre et al. (1986) made point predictions regarding the proportions which individuals would rate themselves and other people positively on bipolar adjective scales while role-playing a positive, negative, or neutral mood. In the neutral mood condition, for example, Lefebvre et al. predicted the following proportions: (a) positive judgments about self, 0.719; (b) positive judgments about others, 0.628; (c) like-self judgments (i.e., judgments in which the same semantic pole was assigned to the self and others), 0.619; (d) positive judgments of others when self is first rated on the positive pole, 0.672; and (e) positive judgments about others when self is first rated on the negative pole, 0.516. Most of these predictions were supported in Lefebvre et al.'s study, providing impressive support for the point predictions produced by Lefebvre's model. In a series of related studies conducted for different age groups, Adams-Webber (1998) also confirmed predictions derived from Lefebvre's AMS-R. Most recently, Grice et al. (2005) noted several limitations to the procedures employed in previous studies and conducted a modified repertory grid study. They asked participants to rate themselves and then 18 other individuals (9 positive and 9 negative) on bipolar scales anchored by terms from the semantic differential

(e.g., pleasant–unpleasant, rugged–delicate). Their observed proportions were: (a) positive self ratings, 0.757; (b) positive ratings of others, 0.606; (c) like-self ratings, 0.615; (d) positive ratings of others when the self was first rated on the positive pole, 0.647; and (e) negative ratings of others when the self was first rated on the negative pole, 0.517. Several of the predicted proportions (0.719, 0.628, 0.619, 0.672, and 0.516, respectively) matched the observed proportions closely, although even small differences between proportions were judged as statistically significant due to extremely high statistical power.

Considered together these repertory grid studies offer promising evidence for the validity of Lefebvre's AMS-R. Nonetheless, the results are far from conclusive as a fair number of point predictions made in these studies have not been confirmed (see also, Adams-Webber 1997a, b). The overall number of published validity studies can also be considered as far too low compared to many popular theoretical models in psychology. Consequently, we have pursued several additional investigations of the AMS-R in our research laboratory using the repertory grid. In the first study we modified the procedures to allow participants to determine which adjective term from each pair they personally considered to be positive. In all previous studies, the experimenters determined which adjectives were positive. We also employed two different sets of adjective pairs and permitted participants to generate their own bipolar scales constructed from their personal constructs. In the second study we used this same design, but controlled the order in which the participants rated themselves and the other people in the grid. Algebraic models of self-reflexion were constructed to represent the different features of these studies, and point predictions were made and tested regarding the proportions of positive ratings of self and others in the repertory grids.

## Study 1

One hundred and fifty seven female and 63 male ($N = 223$) undergraduate students participated in this study in exchange for course credit. Three individuals did not report their gender. The participants' ages ranged from 18 to 46 years ($M = 20.03$; $Mdn = 19.00$; $SD = 3.40$). Eighty-three percent of the individuals reported their ethnicity as Caucasian, 6% as Native American, 4% as African American, 4% as Asian, 1% as Hispanic, and 2% as "Other."

The participants were divided into three groups: semantic differential ($N = 76$), big five ($N = 73$), and personal constructs ($N = 74$). The participants in the first group were presented with 15 bipolar rating scales constructed from the *evaluative* (e.g., generous–stingy, pleasant–unpleasant), *potency* (e.g., strong–weak, bold–timid), and *activity* (e.g., active–passive, energetic–lethargic) dimensions of the semantic differential scale. Participants in the second group were presented with 15 bipolar scales constructed from the Big Five personality factors: *neuroticism* (e.g., calm–worrying, even tempered–temperamental), *introversion* (e.g., quiet–talkative, loner–

joiner), *openness to experience* (e.g., down to earth–imaginative), *agreeableness* (e.g., stingy–generous, irritable–good natured), and *conscientiousness* (e.g., negligent–conscientious, lazy–hardworking). Finally, 15 bipolar scales were built from the personal constructs of the participants in the third group. Specifically, the participants in this group completed a series of sentences (e.g., Typically, a person who is dishonest is also _____. If I could change one thing about myself, I would be more _____.) using their own descriptive terms or short phrases. The opposites of these terms or phrases were also elicited, and the personal construct pairs were then used to anchor the unique bipolar rating scales for each participant.

Participants completed the repertory grid using Idiogrid (Grice 2002), software for administering and scoring a wide variety of self-report inventories. Participants in the semantic differential and big five groups first provided the names or titles (e.g., "mom," "dad") of people who fit 18 roles. Nine of these roles were clearly negative (e.g., a person whom you consider to be unethical or immoral) and nine were clearly positive (e.g., a person who upholds high ethical and moral standards—other than yourself). The participants in these two groups then indicated which of the terms in each of the pairs (e.g., generous–stingy, lazy–hardworking) they considered to be positive or most desirable. Lastly, the participants completed the repertory grid by rating themselves and the other 18 people on the semantic differential or big five bipolar scales. The rated individuals (self and others) and bipolar scales were presented in an order that was randomly determined for each participant, and an "uncertain/does not apply" option was provided below the bipolar scale. The participants were instructed, however, to only select the uncertain option as a last resort. Participants in the sentence-completion group followed a similar set of procedures: first, providing names or titles for the 18 roles; second, completing the sentence-completion task; third, indicating the preferred pole of each personal construct; and finally rating themselves and the 18 people on their unique bipolar scales. Again, the rated individuals and bipolar personal construct scales were presented in an order that was randomly determined for each participant.

The participants' responses were recorded in 15 (rated terms) × 19 (self and 18 others) matrices, or grids, comprised of 0's and 1's as well as missing data for the "uncertain/does not apply" option. The grids for each group were then concatenated vertically, and the observed frequencies of positive ratings for the self and for the other people in the grids were tallied in Idiogrid and converted to proportions.

In the current study, the participants indicated which of the terms in each of the bipolar pairs they considered to be positive or more desirable. Based on our previous work (Grice et al. 2005), we added another tier to the baseline models presented in (11.1) and (11.2) above to compute the predicted proportions; specifically,

$$\begin{array}{ccccc} & & a_4 & & \\ & & a_3 \cdots \star \cdots b_3 & & a_5 \cdots \star \cdots b_4 \\ & a_2 \cdots\cdots\cdots \star \cdots\cdots\cdots b_2 & & & \\ A_1 \equiv a_1 & & & & \end{array} \tag{11.3}$$

The additional level ($a_4$) accounts for the fact that individuals are aware of their judgments regarding the valence (positive or negative) of the descriptive terms. Lefebvre (1990, 36–37) first presented this 4-tiered model and pointed out that it may be appropriate for semantic differential judgments like those found in this study. Consider a person confronted with the bipolar choice of judging himself as "happy" or "sad" in the repertory grid, and the participant has already indicated that "happy" is the positive pole. The $a_1$ parameter indicates the immediate environmental pressure to choose one pole over the other (see Lefebvre 1990); $a_2$ represents the person's non-reflexive image of self as happy or sad; $a_3$ represents the person's reflexive image of self (viz., "I evaluate myself as sad"); and $a_4$ represents the person's meta-reflexive image of self (viz., "I am aware that I am evaluating myself negatively in this study"). If we consider an instance in which the participant is rating another person in the grid, $b_2$ represents the non-reflexive image of the other person, $b_3$ represents the reflexive image of the other person (viz., "I evaluate John as happy"), and $b_4$ represents the other person's reflexive image of himself or herself from the perspective of the participant (viz., "I think John evaluates himself as happy"). Lastly, $a_5$ represents the other person's reflexive image of the participant, again from the participant's perspective (viz., "I think John evaluates me assad").

Following the conventions of previous studies (Lefebvre et al. 1986; Grice et al. 2005), the other person in the model is considered to be the experimenter when the person is rating himself or herself. Moreover, the values for most of the parameters in the model are set equal to 0.5, and the relationship between the participant and experimenter is seen as one of cooperation:

$$\begin{array}{ccccc} & & .5 & & \\ & & .5 & & 1 \\ & .5 \cdots\cdots \bullet \cdots\cdots .5 & & & \\ A_1 \equiv .5 & & & & = .844 \end{array} \tag{11.4}$$

It should be noted the values of 0.5 in (11.4) indicate the lack of any a priori predictions regarding the true values of the parameters for each person in the study. The presentation of the bipolar items were also neutral in the sense that participants were not overtly pressured to choose one pole over the other, and exactly half (9/18) of the individuals in the grid fit positive role titles. Batchelder (1990) discussed and cautiously defended the reasonableness of assuming the values to equal 0.5 in similar types of studies. The single parameter ($a_5$) set equal to 1 indicates the participant is likely to imagine the

experimenter as expecting a socially desirable (i.e., positive) self-evaluation. It is well documented that undergraduate student participants in psychological studies have a tendency to respond in ways they perceive are desired by the experimenter or by society in general (e.g., see Paulhus 2002). Solving the equation with gamma algebra yields 0.844, the predicted proportion of positive self ratings across all participants in the study.

When rating other people in the grid, the base (11.3) model above is modified as follows:

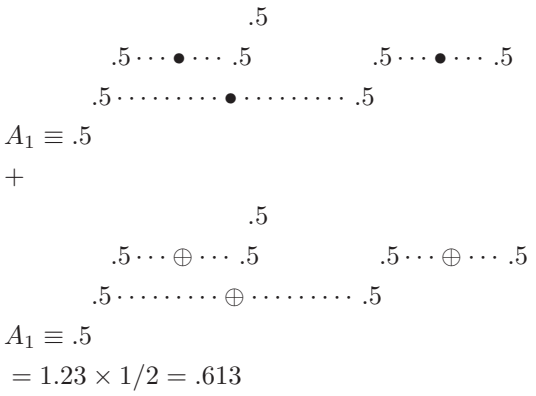
$$\begin{array}{ccc} & b_5 & \\ a_3 \cdots \star \cdots b_3 & & a_4 \cdots \star \cdots b_4 \\ a_2 \cdots\cdots\cdots \star \cdots\cdots\cdots b_2 & & \\ A_1 \equiv a_1 & & \end{array} \tag{11.5}$$

The other person in the equation is not considered to be the experimenter, but the person being rated in the repertory grid. The addition of the $b_5$ parameter furthermore represents the participant's meta-reflexive image of the other person (viz., "I am aware that I am evaluating John negatively in this study"). The parameters are all set equal to 0.5, the equation is solved for cooperation and conflict since no prediction can be made on the actual state of the relationships between the participants and the people whom they are rating, and the results are averaged:

**Table 11.1.** Study 1: Predicted and observed proportions for repertory grid ratings.

| Scales/Proportions | Observed | $CI_{.95}$ |
|---|---|---|
| Semantic Differential | | |
| Positive self ratings | 0.876∗ | 0.856, 0.895 |
| Positive other ratings | 0.607 | 0.600, 0.614 |
| Big Five Markers | | |
| Positive self ratings | 0.810∗ | 0.786, 0.833 |
| Positive other ratings | 0.620 | 0.613, 0.627 |
| Personal Constructs | | |
| Positive self ratings | 0.867∗ | 0.847, 0.888 |
| Positive other ratings | 0.616 | 0.609, 0.623 |
| All Grids | | |
| Positive self ratings | 0.851 | 0.838, 0.863 |
| Positive other ratings | 0.614 | 0.610, 0.618 |

*Note.* Predicted proportions for positive self ratings and positive other ratings were 0.844 and 0.613, respectively. Asterisks indicate proportions that are significantly different from their respective, predicted proportions ($p < 0.05$, two-tailed).




(11.6)

The resulting value, obtained from gamma algebra, indicates the predicted proportion of positive ratings for the other 18 people in the repertory grid.

The two predicted values can be compared to those computed from the grids for each of the three groups (semantic differential, big five, and personal construct), and the differences tested for statistical significance. The results are presented in Table 11.1 and show generally close agreement between the predicted proportions and the observed proportions. For each group, however, the predicted proportion of positive self ratings (0.844) was significantly different from the observed proportion, thus indicating the model was not highly accurate. With regard to the predicted proportion of positive other ratings (0.613), the three observed proportions matched closely and the differences (all $< 0.008$ in absolute magnitude) were not statistically significant. Moreover, when all 223 grids were combined, the predicted and observed proportions of positive self (0.844 vs. 0.851) and positive other (0.613 vs. 0.614) ratings match closely, and the differences were not statistically significant.

## Study 2

When all of the repertory grids were combined in Study 1, the data supported the predictions of Lefebvre's AMS-R. In a second study, we employed the same procedures, but made one small change. Specifically, the participants always rated themselves first on the adjective terms, and then rated the other 18 people in the grid (recall in Study 1 the presentation of the self and others was randomly determined for each participant). This modification lead to changes in the predictions generated from the AMS-R that could be specifically tested.

One hundred and sixty-three female and 78 male ($N = 241$) undergraduate students participated in the study in exchange for course credit. The participants' ages ranged from 18 to 46 years ($M = 20.02$; $Mdn = 19.00$; $SD = 3.21$). Eighty-five percent of the individuals reported their ethnicity as

Caucasian, 7% as Native American, 1% as African American, 3% as Asian, 1% as Hispanic, and 3% as "Other." The participants were again divided into three groups: semantic differential ($N = 78$), big five ($N = 83$), and personal constructs ($N = 80$). The procedures were exactly as those described in Study 1 with the exception that the participants always rated themselves first on the 15 bipolar adjective scales followed by the 18 other people, who were presented in a random order which was determined separately for each participant.

The predicted proportion of positive self ratings (0.844) was computed using the model reported in (11.4) above. Given the slight change in procedures, however, a different modeling strategy was employed for computing the predicted proportions of positive other ratings. This strategy was first reported by Lefebvre et al. (1986) in their seminal paper and subsequently used by Grice et al. (2006). Because $a_3$ represented the participant's reflective image of self, and because each participant rated himself or herself prior to rating the other people in the grid, the value for this variable was set to 0 (self was rated negatively) or 1 (self was rated positively). Consequently, when the participant first rated self positively, the frequency of choosing the positive pole when rating others was computed by averaging the results from the cooperative ($\bullet$) and conflictive ($\oplus$) models:

$$
\begin{array}{l}
\qquad\qquad\qquad\quad .5 \\
\qquad\quad 1\cdots\bullet\cdots .5 \qquad\qquad .5\cdots\bullet\cdots .5 \\
\qquad .5\cdots\cdots\cdots\bullet\cdots\cdots\cdots .5 \\
A_1 \equiv .5 \\
+ \\
\qquad\qquad\qquad\quad .5 \\
\qquad\quad 1\cdots\oplus\cdots .5 \qquad\qquad .5\cdots\oplus\cdots .5 \\
\qquad .5\cdots\cdots\cdots\oplus\cdots\cdots\cdots .5 \\
A_1 \equiv .5 \\
= 1.32 \times 1/2 = .660
\end{array}
\tag{11.7}
$$

$$
\begin{array}{l}
\qquad\qquad\qquad\quad .5 \\
\qquad\quad 0\cdots\bullet\cdots .5 \qquad\qquad .5\cdots\bullet\cdots .5 \\
\qquad .5\cdots\cdots\cdots\bullet\cdots\cdots\cdots .5 \\
A_1 \equiv .5 \\
+ \\
\qquad\qquad\qquad\quad .5 \\
\qquad\quad 0\cdots\oplus\cdots .5 \qquad\qquad .5\cdots\oplus\cdots .5 \\
\qquad .5\cdots\cdots\cdots\oplus\cdots\cdots\cdots .5 \\
A_1 \equiv .5 \\
= 1.13 \times 1/2 = .566
\end{array}
\tag{11.8}
$$

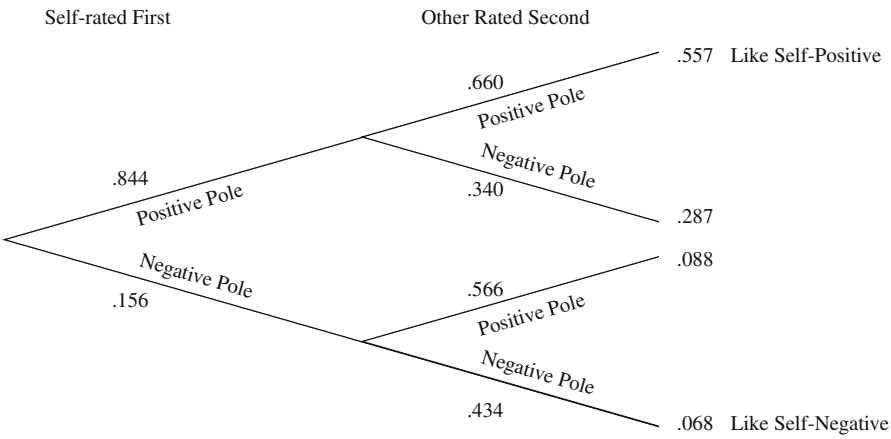



**Fig. 11.2.** Probability graph for computing predicted self-similar proportions.

The conditional proportions which participants will rate the other individuals as similar to themselves can then be computed using the probability graph shown in Fig. 11.2. The proportions on the different branches of the tree of the graph are simply multiplied to obtain the conditional proportions which can then be summed to obtain the predicted proportions of interest. Consistent with previous studies, the following proportions were predicted: (a) positive judgments about self, 0.844; (b) positive judgments about others, 0.645 [0.557 + 0.088]; (c) like-self judgments (i.e., judgments in which the same semantic pole was assigned to the self and others), 0.625 [0.557 + 0.068]; (d) positive judgments of others when self is first rated on the positive pole, 0.660; and (e) negative judgments about others when self is first rated on the negative pole, 0.434.

These five predicted values were compared to the computed results from the grids for each of the three groups (semantic differential, big five, and personal construct), and the differences again tested for statistical significance. As can be seen in Table 11.2, close agreement was again found between the predicted and observed proportions, although most of the differences were statistically significant. For each group and for all 241 participants the predicted proportion of positive self ratings (0.844) was significantly different from the observed proportions, thus indicating the model was not highly accurate. Generally, the observed proportions were greater than 0.844. Similarly, the observed proportions of positive and negative ratings for the other people in the repertory grids were significantly different from their respective predicted proportions (see Table 11.2). The only exception was for the predicted proportion of negative other ratings when the self was first rated negatively; for the big five and personal construct groups, as well as for all 241 partic-

**Table 11.2.** Study 2: Predicted and observed proportions for repertory grid ratings.

| Scales/Proportions | Predicted | Observed | $CI_{.95}$ | Study 1 |
|---|---|---|---|---|
| Semantic Differential | | | | |
| Positive self ratings | 0.844 | 0.873∗ | 0.852, 0.892 | 0.876 |
| Positive other ratings | 0.645 | 0.624∗ | 0.617, 0.631 | $0.607^{++}$ |
| Like Self Overall | 0.625 | 0.602∗ | 0.595, 0.609 | |
| Like Self Positive | 0.660 | 0.629∗ | 0.622, 0.637 | |
| Like Self Negative | 0.434 | 0.414∗ | 0.394, 0.433 | |
| Big Five Markers | | | | |
| Positive self ratings | 0.844 | 0.817∗ | 0.794, 0.838 | 0.810 |
| Positive other ratings | 0.645 | 0.616∗ | 0.609, 0.622 | 0.620 |
| Like Self Overall | 0.625 | 0.596∗ | 0.590, 0.603 | |
| Like Self Positive | 0.660 | 0.629∗ | 0.621, 0.636 | |
| Like Self Negative | 0.434 | 0.445 | 0.429, 0.462 | |
| Personal Constructs | | | | |
| Positive self ratings | 0.844 | 0.885∗ | 0.865, 0.903 | 0.867 |
| Positive other ratings | 0.645 | 0.623∗ | 0.616, 0.630 | 0.616 |
| Like Self Overall | 0.625 | 0.607∗ | 0.600, 0.614 | |
| Like Self Positive | 0.660 | 0.629∗ | 0.622, 0.637 | |
| Like Self Negative | 0.434 | 0.424 | 0.402, 0.445 | |
| All Grids | | | | |
| Positive self ratings | 0.844 | 0.857∗ | 0.845, 0.869 | 0.851 |
| Positive other ratings | 0.645 | 0.621∗ | 0.617, 0.625 | $0.614^{++}$ |
| Like Self Overall | 0.625 | 0.602∗ | 0.598, 0.606 | |
| Like Self Positive | 0.660 | 0.629∗ | 0.625, 0.633 | |
| Like Self Negative | 0.434 | 0.43 | 0.419, 0.441 | |

*Note.* Asterisks indicate proportions that are significantly different from their respective, predicted proportions ($p < 0.05$, two-tailed). The symbol '$^{++}$' indicates the proportion in the first study is significantly different from the proportion in the second study ($p < 0.05$, two-tailed).

ipants, the predicted proportion (0.434) matched the observed proportions (0.445, 0.424, and 0.430, respectively) closely. It is also important to point out that most of the statistically significant differences between the predicted and observed proportions for the various groups and self/other ratings were small in magnitude. In addition, the observed proportion of positive ratings of others for all 241 participants (0.621) was significantly higher than the corresponding observed proportion (0.614) from the 223 participants in Study 1. This result indicates that requiring participants to rate themselves first impacted the observed proportions in a manner consistent with the AMS-R predictions. Moreover, the observed proportions of positive self ratings from all participants in Study 2 (0.857) and Study 1 (0.851) indicated remarkable consistency and suggest the particular AMS-R model used in these studies may be in need of refinement.

## Conclusion

The results of these two repertory grid studies add to the body of literature that offers tentative support for Lefebvre's Algebraic Model of Self-Reflection (AMS-R). On the one hand, many of the point predictions computed from the AMS-R were closely matched by the observed proportions in both studies; and in some instances the differences were not statistically significant. These non-significant findings, which confirmed the model's predictions, are particularly impressive given the abundance of statistical power generated from concatenating the large grids across numerous participants. This wealth of statistical power also explains why even small differences between proportions were judged as statistically significant. The AMS-R also showed sensitivity in detecting the change in procedures from the first study to the second. In the first study, the participants rated themselves and the other people in random order in the repertory grids, and in the second study the participants always rated themselves prior to rating the other people. While the observed proportions in the second study did not match the predicted values, they were nonetheless consistent with the differences predicted by the AMS-R between the two studies.

On the other hand, most of the observed proportions were nonetheless statistically different from the predicted proportions in both studies, particularly, Study 2. As mentioned above these differences were small in magnitude, but the strength of Lefebvre's approach is its ability to generate specific point predictions. Given sufficient control over the testing situation and ample sample size, the observed proportions should show greater agreement with the values predicted from the model than they do in the current two studies. At least two factors must be considered when explaining the negative results. First, the participants were remarkably consistent with regard to the proportion of positive self ratings across the two studies, even though the observed proportions did not match the predicted proportion. This consistency suggests the

model used to predict this particular proportion (see (11.4)) may be incorrect. Given the current literature, however, it is unclear how the model can be reasonably manipulated to yield a slightly higher predicted proportion of positive self ratings. Second, participants in our studies were given the option to select a "does not apply" option when conducting their bipolar ratings. In previous studies, except Grice et al. (2005), participants were required to choose one of the two construct poles. While this adjustment may help to reduce unreliable judgments, it is not clear if it violates the fundamentally Boolean nature of Lefebvre's original modeling approach. It may be that an alternative non-Boolean model is more appropriate (see Lefebvre 2001), or that the current models are again in need of some refinement.

In any event, additional analytical and empirical work regarding the AMS-R is needed, and the current results clearly suggest such work is warranted. Psychological models that yield exact point predictions are extremely rare, and the predictions generated from Lefebvre's models are tantalizingly close to the observed values computed from repertory grid data. Pairing the grid technique with experimental procedures such as subliminal priming (see McDaniel 2005), in future studies could lead to even stronger tests of the models as well as their refinement. Considered more generally, the two studies above show that the repertory grid technique and the theoretical framework of *Personal Construct Psychology* are well suited for investigations concerning Lefebvre's AMS-R (Adams-Webber 1990; Grice and McDaniel 2006). Perhaps through further investigation, the formal processes that underlie bipolar decisions will finally be understood.

## Acknowledgments

The authors would like to thank Kristian Alton, James Barraclough, Bobby Jackson, Laura Kemp, Clint Martin, Shannon Milligan, Heather Orr, Nicole Rosell, Ashley Ryder, and Tiffany Truitt for their help collecting data for studies cited in this manuscript that were conducted in the Personality Research Laboratory at Oklahoma State University.

## References

Adams-Webber, J. 1990. A model of reflexion from the perspective of personal construct theory. In *The Structure of Human Reflexion* 93–112. Ed. H. Wheeler. New York: Lang.

Adams-Webber, J. 1997a. Positive-negative asymmetry in the evaluation of familiar versus unfamiliar persons and objects. *Journal of Constructivist Psychology* 10:139–152.

Adams-Webber, J. 1997b. Positive-negative asymmetry and bipolar contrast. *Journal of Constructivist Psychology* 10:387–393.

Adams-Webber, J. 1998. Some experimental tests of Lefebvre's model of reflexion. In proceedings of *Workshop on Multi-Reflexive Models of Agent Behavior*. Los Alamos, NM.

Batchelder, W. 1990. Some critical issues in Lefebvre's framework for ethical cognition and reflexion. In *The Structure of Human Reflexion* 139–156. Ed. H. Wheeler. New York: Lang.

Bogatyrev, K. K. 1998. Reflexive thoughts about a Medieval Russian epic. In Proceedings of *Workshop on Multi-Reflexive Models of Agent Behavior*. Los Alamos, NM.

Grice, J. W. 2002. Idiogrid: Software for the management and analysis of repertory grids. *Behavior Research Methods, Instruments, & Computers* 34:338–341.

Grice, J. W. 2006. *GammaCalc: A Calculator for Gamma Algebra* [Computer program]. Stillwater: Oklahoma State University.

Grice, J. W., and B. McDaniel. 2006. Reflections on Lefebvre's algebraic model of self-reflexion and Kelly's theory of personal constructs. *Reflexive Processes and Control* 6:88–93.

Grice, J. W., B. McDaniel, and D. Thompsen. 2005. Testing an algebraic model of self-reflexion. *Perceptual and Motor Skills* 100:1036–1048.

Kelly, G. A. 1955. *The Psychology of Personal Constructs*. New York: Norton.

Lefebvre, V. A. 1990. The fundamental structures of human reflexion. In *The Structure of Human Reflexion* 5–70. Ed. H. Wheeler. New York: Lang.

Lefebvre, V. A. 1992. *A Psychological Theory of Bipolarity and Reflexivity*. Lewiston, NY: Edwin Mellen Press.

Lefebvre, V. A. 2001. Algebra of conscience. In *Philosophy and Methodology of the Social Sciences* 30. Ed. W. Leinfellner and G. Eberlein. Boston: Kluwer.

Lefebvre, V. A. 2002. The law of self-reflexion: A possible unified explanation for three different psychological phenomena. *Reflexive Processes and Control* 1:91–99.

Lefebvre, V. A., J. Adams-Webber, and V. D. Lefebvre. 1986. Modeling an experiment on construing self and others. *Journal of Mathematical Psychology* 30:317–330.

McDaniel, B. L. 2005. *Evaluating the Mathematical Structure of the Self via Subliminal Manipulation*. Unpublished Masters Thesis, Oklahoma State University, Oklahoma.

Osgood, C. E., G. J. Suci, and P. H. Tannenbaum. 1957. *The Measurement of Meaning*. Urbana, IL: University of Illinois Press.

Paulhus, D. L. 2002. Socially desirable responding: The evolution of a construct. In *The Role of Constructs in Psychological and Educational Measurement* 49–69. Ed. H. I. Braun and D. N. Jackson. Mahwah, NJ: Lawrence Erlbaum Associates.

Schwartz, R. M. 1997. Consider the simple screw: Cognitive science, quality improvement, and psychotherapy. *Journal of Consulting and Clinical Psychology* 65:970–983.

Schwartz, R. M, A. L. Fasiczka, E. Frank, C. F. Reynolds, and M. E. Thase. 2002. Optimal and normal affect balance in psychotherapy of major depression: Evaluation of the balanced states of mind model. *Behavioral and Cognitive Psychotherapy* 30:439–450.

# Part IV

# Conflict Analysis

# 12

# Anticipating Terrorist Safe Havens from Instability Induced Conflict

Robert Shearer and Brett Marvin

**Summary.** This chapter presents recent methods developed at the Center for Army Analysis to classify patterns of nation-state instability that lead to conflict. The ungoverned areas endemic to failed nation-states provide terrorist organizations with safe havens from which to plan and execute terrorist attacks. Identification of those states at risk for instability induced conflict should help to facilitate effective counter terrorism policy planning efforts. Nation-states that experience instability induced conflict are similar in that they share common instability factors that make them susceptible to experiencing conflict. We utilize standard pattern classification algorithms to identify these patterns. First, we identify features (political, military, economic and social) that capture the instability of a nation-state. Second, we forecast the future levels of these features for each nation-state. Third, we classify each future state's conflict potential based upon the conflict level of those states in the past most similar to the future state.

## Introduction

The pandemic of failed states has left the world littered with countries lacking the power to govern their people. While in recent years this problem has been most acute on the continent of Africa, it is truly a global dilemma, affecting nearly every part of the world since the 1950's. One hundred and thirty five states failed in countries with populations greater than 500,000 between 1955 and 1998. Alarmingly, incidences of state failure increased dramatically over the same period. In 1955, fewer than six percent of nation-states were in failure; by the early 1990's this number had risen to nearly 30 percent, before dropping to about 20 percent in the latter part of that decade (Goldstone et al. 2000).

A state may fail for a variety of causes, ranging from political to military to economic to social. These causes tend to slowly grow over time, destabilizing the nation-state. Typically, failed or failing nation-states share many features in common, including: deteriorating infrastructure; widespread corruption; unregulated borders; stagnant or declining gross domestic product;

S. Argamon, N. Howard (eds.), *Computational Methods for Counterterrorism*, DOI 10.1007/978-3-642-01141-2_12,

rampant crime; and a national currency that is not widely accepted. Additionally, armed groups frequently operate within the boundaries, but outside the control of the government, of destabilized nations. Once failed, a nation-state is susceptible to internal conflict. The chaotic conditions and permissive environments typical of internal conflicts are easily exploited by terrorist organizations in order to further their aims.

Since September 11, 2001, the security challenges posed by failed and collapsed states have garnered increased attention. The chaos and lawlessness state failure engenders can be exploited by terrorist groups who have historically leveraged such conditions to create safe havens for re-grouping, training, and planning their operations. A key component of counter terrorism is to reduce the size and effectiveness of the safe havens that protect terrorist organizations.

While a fool-proof means of promoting effective domestic sovereignty abroad has thus far eluded policy practitioners, it is none-the-less instructive and important to measure the relative stability of global nation-states as a means of prioritizing and marshaling counter terrorism resources in order to counter emerging threats. Towards this end, a reliable, transparent means of evaluating nation-state instability and an effective, accurate, capability to predict which states will be at risk for internal conflict in future years should help to inform counter terrorism policymakers.

Analysts at the Center for Army Analysis (CAA) have studied instability induced conflict since the mid 1990s. Initial efforts identified macro-structural factors at the nation-state level that might contribute to or provide an environment conducive to instability. Subsequent work attempted to forecast conflict potential for nation-states in the future. Multiple methodologies were investigated, to include logistic regression, neural networks, and possibility theory (see O'Brien (2002) for an unclassified discussion of this work). The most recent research conducted at CAA, described in this chapter, sought to improve upon the accuracy of the earlier conflict potential forecasting methodologies in a manner more transparent and understandable to the layman. This research focused on two objectives.

- Determine each nation-state's conflict potential in the year 20xx. We refer to this objective as the Conflict Forecast. We met this objective by classifying each state's future conflict potential as high, uncertain or low.
- Determine the conflict potential trend that each nation-state follows out to 20xx. We refer to this objective as the Path to Conflict. We met this objective by classifying the trajectory (towards high or low conflict potential) and speed (years until low or high conflict potential) along which each state is trending.

## Methodology

Nations that experience instability induced conflict are similar in that they share common instability features that make them susceptible to experiencing conflict. If one could forecast the status of a nation in the future, in terms of

these features, then a reasonable forecast for that nation's conflict potential would be the level of conflict experienced by countries with similar features in the past. We proposed a six-step pattern classification approach based on this logic to meet the research objectives.

- Collect historical conflict (1993–2003) data on select nation-states.
- Select features (political, military, economic and social macro-structural indicators) that represent the instability of a nation-state. Obtain historical data for these features. Create a past feature vector for each nation-state – year (e.g. Albania 1995).
- Forecast future feature vectors for each nation-state from the past feature vectors.
- Select a classifier with which to classify the conflict potential of future feature vectors.
- Train the classifier on the past feature vectors, identifying which combination of feature vectors experienced conflict and which experienced peace.
- Classify the conflict potential of each future feature vector based upon proximity to past feature vectors.

## Conflict data

The research team obtained the historical conflict data from the Heidelberg Institute of International Conflict Research (Heidelberger Institut für Internationale Konfliktforschung or HIIK). The HIIK, co-located with the Department of Political Science at the University of Heidelberg, is a private organization. Their work is dedicated to research, documentation and analysis of national and international political conflicts. They classify wars and conflicts of lesser intensity according to the actual amount of violence observed, not according to the number of fatalities. Conflicts are defined as the clashing of interests (positional differences) on national values and issues (territory, independence, self-determination, autonomy, ideology, power and resources) of some duration and magnitude between at least two parties (states, groups of states, organizations or organized groups) that are determined to pursue their interests and win their case. At least one party is the organized state. The HIIK assigns conflict intensity in four categories: Latent Conflict, Crisis, Severe Crisis and War. Only the highest intensity level of the year is taken into account. Definitions of the four categories of conflict intensity follow.

- *Latent Conflict.* The positional differences and the clashing interests in a latent conflict must be articulated as demands or claims. The other party has to be aware of these demands. A conflict can remain latent for long periods of time and get comparatively little public attention. When one party starts pushing for new negotiations, the conflict can intensify for a certain time before it falls back to its latent state.
- *Crisis.* A crisis includes tensions that are expressed by means that are below the threshold of violence. Tense relations between the parties can reach a turning-point from where the use of force may become more likely.

At this point, many decisions are based on incomplete information and made under time-pressure.

- *Severe Crisis.* A severe crisis is defined as a state of high tension between the parties, who either threaten to resort to the use of force or they actually use physical or military force sporadically.
- *War.* War is a type of violent conflict characterized by (a) the fighting of at least two opponents with organized, regular military forces; (b) the fact that the fighting is not sporadic but last for a considerable period of time; (c) the fact that the fighting is intense, that is, it leads to victims and destruction.

Conflicts of the intensities Latent Conflict and Crisis are predominantly nonviolent, while conflicts of the intensities Severe Crisis and War are mainly violent states of conflict. Historically, the United States has not intervened in instability induced conflicts until casualties are experienced in the failing state. Accordingly, the research team decided to utilize just two conflict intensity categories: Conflict and Peace. We mapped Crisis and Latent Conflict to Peace, War and Severe Crisis to Conflict.

## Features

The research team selected thirteen features (macro-structural indicators) to identify patterns of nation-state instability that lead to conflict. We obtained these features from multiple studies that CAA conducted in the late 1990s and early 2000s that identified the key macro-structural features that affect stability (O'Brien 2002). These thirteen include three political (civil liberties, democracy and political rights), one military (conflict history), three economic (male unemployment, GDP per capita and trade openness) and six social (adult male literacy, caloric intake, ethnic diversity, infant mortality, life expectancy and religious diversity) features.

### Feature data

We obtained the data for the thirteen features from a wide variety of unclassified sources. Definitions and sources for the features follow.

- *Adult Male Literacy.* The percentage of literate males, ages 15 and above. World Bank (2007).
- *Caloric Intake.* An estimate of the average number of calories consumed per person, per day. Food and Agriculture Organization of the United Nations.
- *Civil Liberties.* A measure of the freedom of country's people "to develop views, institutions, and personal autonomy apart from the state." Seven point ordinal scale from 1 (free) to 7 (not free). Freedom House.
- *Conflict History.* The percentage of time (in years) spent in a state of conflict (war or severe crisis). Note: Percentage of time in conflict

spans the years in training data only. Heidelberg Institute of Conflict Research.
- *Democracy.* A measure of degree of democracy. Ordinal scale from −10 (least democratic) to 10 (most democratic). Polity IV Project.
- *Ethnic Diversity.* The population of the largest ethnic group in the country as a percentage of the total population. CIA World Fact Book (Central Intelligence Agency) and Country Indicators of Foreign Policy (Carleton University).
- *GDP per capita.* The annual gross domestic product per person measured in constant 1998 U.S. dollars. World Bank (2007).
- *Infant Mortality.* The number of deaths of children under 1 year of age per 1,000 live births. U.S. Bureau of the Census.
- *Life Expectancy.* The average life expectancy (males and females combined) U.S. Bureau of the Census.
- *Male Unemployment.* The percentage of the male labor force that is unemployed. World Bank (2007).
- *Political Rights.* A measure of rights to participate meaningfully in the political process. Seven point ordinal scale from 1 (free) to 7 (not free). Freedom House.
- *Religious Diversity.* The population of the largest religious group in the country as a percentage of the total population. CIA World Fact Book (Central Intelligence Agency) and Country Indicators of Foreign Policy Project (Carleton University).
- *Trade Openness.* The ratio of a country's total imports and exports to GDP. World Bank (2007).

**Feature scaling**

We utilized Euclidean distance as the proximity measure between a nation-state's forecasted future (a point in the 13 dimensional feature space) and all other nation-states' pasts (again, points in the 13 dimensional feature space). This measure requires interval data on the same scale if the features are to have equal weight in the classification. The raw data for the 13 features contains both ordinal and interval data with widely ranging scales. The research team formed a 13 dimensional feature vector, $\mathbf{x}$, for each nation-state, for each past year, with the 13 feature scores as components. We denote indicators by the variable $i$, nation states by $j$, and the year by $k$, such that $x_{i,j,k}$ denotes the value of feature $k$ for nation state $i$ at year $j$. We then transformed each feature vector into a scaled feature vector, $\mathbf{y}$, by mapping each feature score into a scaled score between zero and one, where

$$y_{i,j,k} = \frac{x_{i,j,k} - \min_i x_{i,j,k}}{\max_i x_{i,j,k} - \min_i x_{i,j,k}} \tag{12.1}$$

We also identified the upper and lower bounds for each component across the feature space. We limited future forecasts from exceeding these bounds. Wide dispersion of the scaled feature scores across [0,1] would aid the classification process as more similar countries would lie in closer proximity to one another than non-similar countries. All thirteen displayed a wide range of values.

**Feature analysis**

The methodology rests on two major assumptions. First, nations that experience instability induced conflict are similar in that they share common instability features that make them susceptible to experiencing conflict. Second, the distance between the scaled 13 dimensional points serves as a reasonable proxy for the similarity between two nation-states. Prior to forecasting future feature values and applying classifiers, the research team analyzed the scaled feature data to determine if these assumptions appeared valid. Obviously, we could not visually inspect the 13 dimensional feature space within which the points lie. Instead, we attempted to visually test our assumptions by generating 54 three dimensional plots from each of the possible combinations of one political, one social, and one economic factor. Each plot contains 1705 points (11 years x 155 countries), where each point represents a nation-state – year. We then "colored" each point based upon the historical level of conflict observed: black for nation-states that experienced conflict, grey for those that experienced peace. If the assumptions were valid, one would expect that the black and grey points would form distinct groupings or clouds. Nation-states that experienced conflict would lie in close proximity to others that experienced conflict; nation-states that experienced peace would lie in close proximity to others that experienced peace. Figure 12.1 shows one of these plots, the Infant Mortality Rate, Civil Rights and Trade Openness plot.

The grey and the black points form two clouds that have a small intersection – providing support for the validity of the assumptions. One would expect a high level of accuracy in classifying future states based upon their location in this plot, with most classification errors occurring in the intersection. Figure 12.2 shows another plot, the Adult Literacy, Democracy and Unemployment plot.

The grey and the black points form two clouds that have a large intersection – removing support for the validity of the assumptions. One would expect a low level of accuracy in classifying future states based upon their location in this plot, with no distinct groupings. The remaining 52 plots fell somewhere between these two in terms of the size of the intersection. As a result, it remained unclear what was occurring in 13-space, and another approach was needed.

The research team then conducted a Principal Component Analysis (PCA) of the scaled feature data in an attempt to reduce the 13 dimensional clouds into three dimensional clouds. PCA provides a non-parametric method to

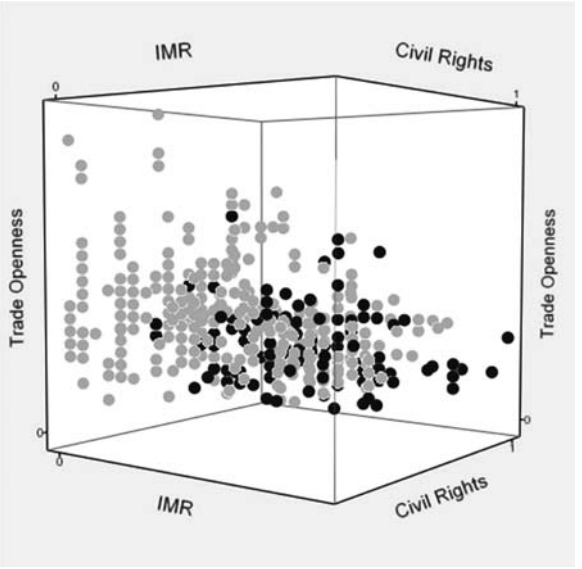



**Fig. 12.1.** Feature Analysis: Infant Mortality, Civil Rights and Trade Openness. The *grey points* denote nation states that experienced peace in a given year. The *black points* denote nation states that experienced conflict in a given year.

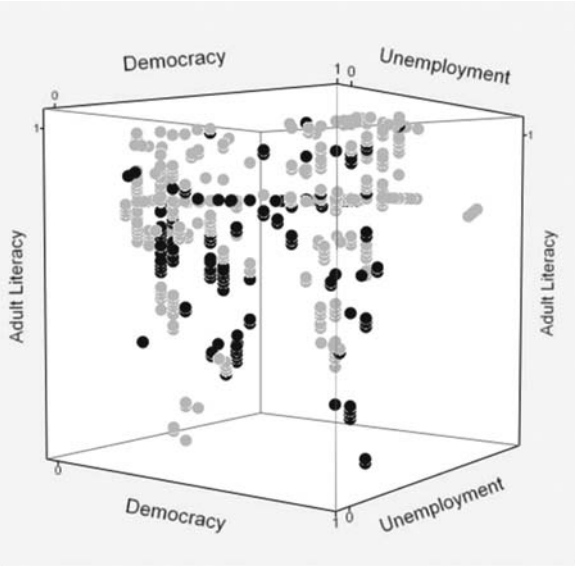



**Fig. 12.2.** Feature Analysis: Adult Literacy, Democracy and Unemployment. The *grey points* denote nation states that experienced peace in a given year. The *black points* denote nation states that experienced conflict in a given year.

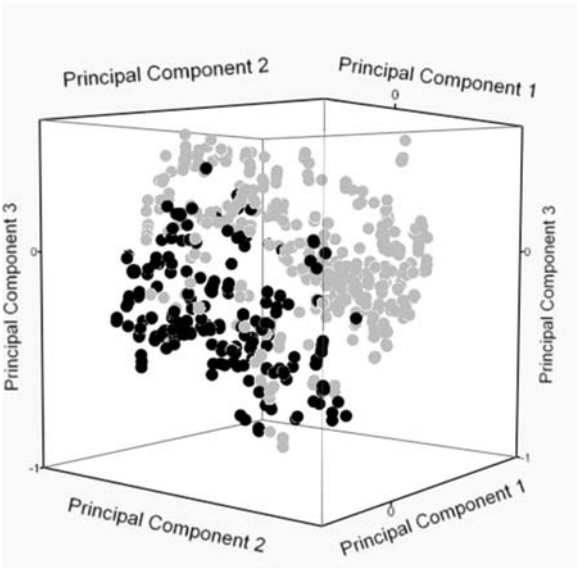



**Fig. 12.3.** Feature Analysis: Principal Component Analysis. The *grey points* denote nation states that experienced peace in a given year. The *black points* denote nation states that experienced conflict in a given year.

reduce the dimensionality of a complex data set. The method searches for linear combinations of the original feature vectors that best express the variance in the data. The transformed vectors, termed principal components, are orthogonal and oriented in a direction that maximizes the variance explained in the data. We then plotted the transformed points in the three dimensional subspace spanned by the first three principal components. As before, we "colored" each point based upon the historical level of conflict observed: black for nation-states that experienced conflict, grey for those that experienced peace. Figure 12.3 shows this plot.

The grey and the black points form two distinct clouds with a small intersection, even smaller than in the Infant Mortality Rate, Civil Rights and Trade Openness plot in Fig. 12.1. We expect a high level of accuracy in classifying future states based upon their location in this plot, with most classification errors occurring in the intersection.

The Principal Component Analysis supports the similarity and distance assumptions that underlie the methodology. The distinct black and grey clouds suggest that states with similar historical levels of conflict share common levels of the 13 features. The separation between the clouds and the small size of the intersection suggest that the distance between points will serve as a reasonable proxy for similarity between states.

**Feature forecast**

The research team utilized a smoothing algorithm, weighted moving average (WMA), to forecast future feature vectors (Abraham and Ledolter 1983). WMA, as applied here, forecasts future feature values as a weighted function of the present value and the changes in the past n values. The weights assigned to past changes, $\alpha_i$, decrease arithmetically, under the assumption that the more recent changes are the most important. The formula for WMA follows.

$$y_{i,j,k+1} = y_{i,j,k} + \alpha_1(y_{i,j,k} - y_{i,j,k-1}) + \alpha_2(y_{i,j,k-1} - y_{i,j,k-2}) + \ldots + \alpha_{n-1}(y_{i,j,k-n-2} - y_{i,j,k-n-1}) \quad (12.2)$$

$$where\ \alpha_t = \frac{n-t}{\sum_{i=1}^{n-1} i}$$

The research team investigated several values for $n$. We divided the feature data into a five year training set and a six year test set. We then forecasted six years of future values from the training set, with $n = 1, 2, \ldots, 5$, and calculated the mean square error against the test set. The $n = 5$ forecast minimized the error. We plan to re-examine this topic again when we expand the feature data set.

## Classifiers

We considered two algorithms to classify the future feature vectors, the k-Nearest Neighbor (kNN) algorithm and the Nearest Centroid (NC) algorithm. Both are non-parametric approaches that make no assumptions as to the underlying distribution of any of the vectors. Both also have simple geometric interpretations.

**k-Nearest Neighbor algorithm**

The kNN algorithm classifies each future feature vector as a function of the n closest past feature vectors. This function, or decision rule, describes the number of neighboring vectors of one class (peace or conflict) required for a similar classification for the new vector. These decision rules call for either a simple majority or a super majority of neighboring vectors to be of the same class for a similar classification. A simple majority requires more than half of the neighbors to be of the same class for a similar classification (e.g. 3 black neighbors out of 5 nearest-neighbors leads to a black (high conflict potential) classification). A super majority requires more than the simple majority for classification (e.g. 4 or more black neighbors out of 5 nearest-neighbors leads to a black (high conflict potential) classification). Several super majority decision

rules may exist for a given k value (e.g. when k = 5 two super majority rules exist: 4 or more neighbors of the same class and 5 neighbors of the same class). Super majority decision rules can lead to no classification (e.g. 3 black neighbors out of 5 nearest-neighbors leads to no classification). An alternative approach is to classify these vectors as having uncertain conflict potential. Figure 12.4 portrays the two options for partitioning the feature vector space into classification regions, creating a Voronoi tessellation of the space. The colored circles represent past state years that experienced conflict (black circles) or peace (grey circles). The black, dark grey and grey regions, or Voronoi cells, identify the classification for future vectors. The white circle represents a future feature vector requiring classification (Duda et al. 2001).

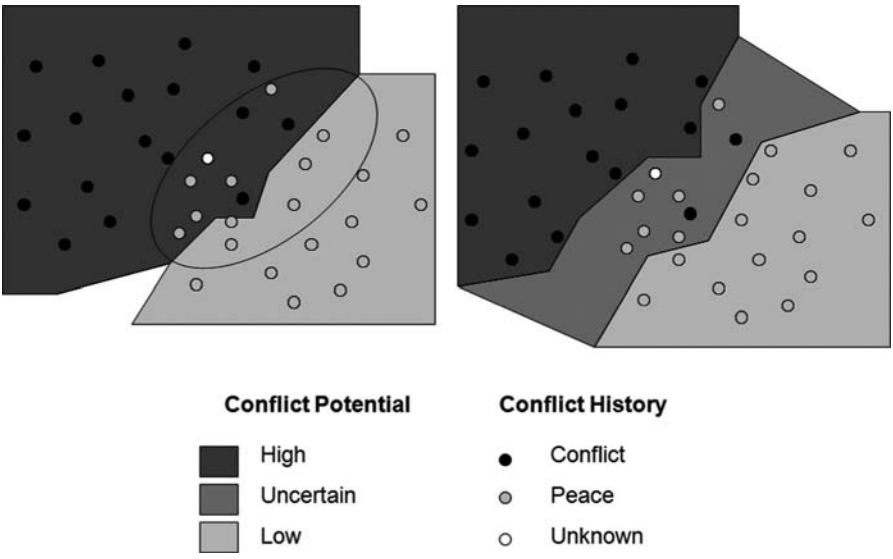



**Fig. 12.4.** k-Nearest Neighbor Algorithm: Feature Space Partition. Future feature vectors will receive a high conflict potential classification if they fall in the black region, an uncertain conflict potential classification if they fall in the *dark grey* region, and a low conflict potential classification if they fall in the *light grey* region.

A simple majority decision rule leads to the partition on the left. This type of rule will always classify a future feature vector. The future feature vector identified by the white circle would receive a black (high conflict potential) classification. The problem with this option is that we know where we will experience the majority of our classification errors, in the intersection discussed in Feature Analysis (identified here with an oval) and yet do nothing to address these errors. A second option would utilize a super majority decision rule, which leads to the partition on the right. This type of decision

rule will have three Voronoi cells. Vectors in the problematic intersection will receive a dark grey (uncertain conflict potential) classification. The future feature vector identified by the white circle would receive a dark grey (uncertain conflict potential) classification. Standard pattern classification techniques do not include classes not found in the original data. Yet this approach provides the capability to identify which nation-states are in the intersection. Classifying a nation-state's conflict potential as uncertain does not mean we have no information about the state. Rather, the classification denotes that the nation-state exhibits patterns of features that have in the past led to conflict in some states, to peace in others. Such states are at tipping points and may prove the most susceptible to stabilizing efforts.

The two objectives require a Conflict Forecast and a Path to Conflict. We provide an artificial two-dimensional Conflict Forecast example of the kNN algorithm in Fig. 12.5.

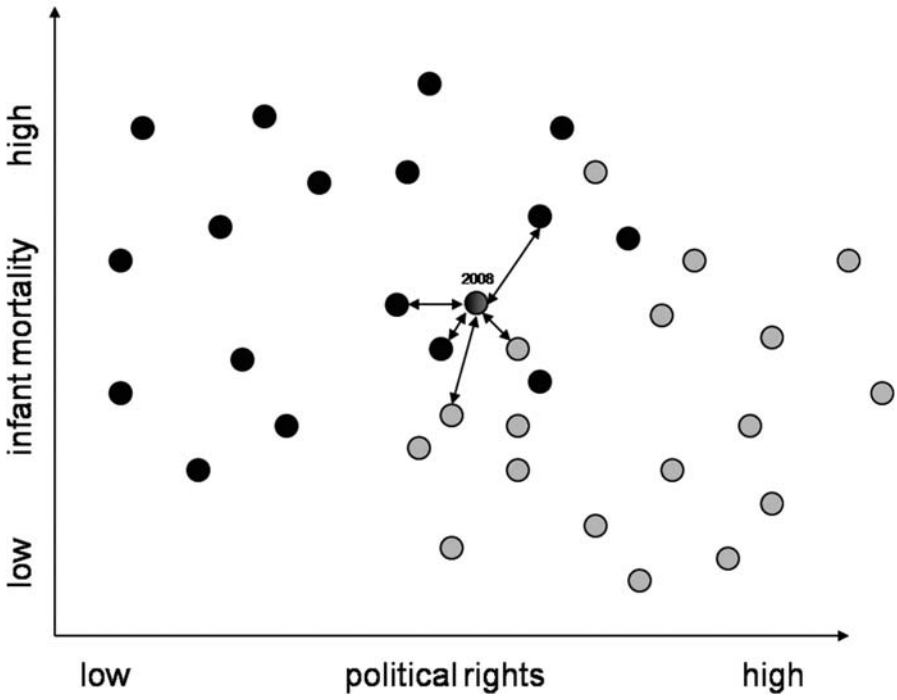



**Fig. 12.5.** k-Nearest Neighbor Algorithm: Conflict Forecast. The forecasted 2008 feature vector had three of five nearest neighbors that had experienced conflict (*black points*)and two of five that had experienced peace (*grey points*).

First, we plot the historical feature vectors (Political Rights, Infant Mortality) and "color" each two-dimensional point based upon the known level of past conflict. Second, we forecast a future feature vector for a given nation-state and plot this two-dimensional point. Third, we identify the k-nearest-neighbors for the future feature vector and classify according to the decision rule. The five nearest-neighbors to the 2008 forecast in the figure include three black and two grey points. A simple majority decision rule leads to a black (high conflict potential) classification. A super majority rule leads to a dark grey (uncertain conflict potential) classification. We provide an artificial two-dimensional Path to Conflict example of the kNN algorithm in Fig. 12.6.

First, we plot the historical feature vectors (Political Rights, Infant Mortality) and "color" each two-dimensional point based upon the known level of past conflict. Second, we forecast future feature vectors for a given nation-state and plot these two-dimensional points. Third, we identify the k-nearest-neighbors for each future feature vector and classify according to the decision rule. The figure portrays the state's last three years as grey, classifies 2008 as

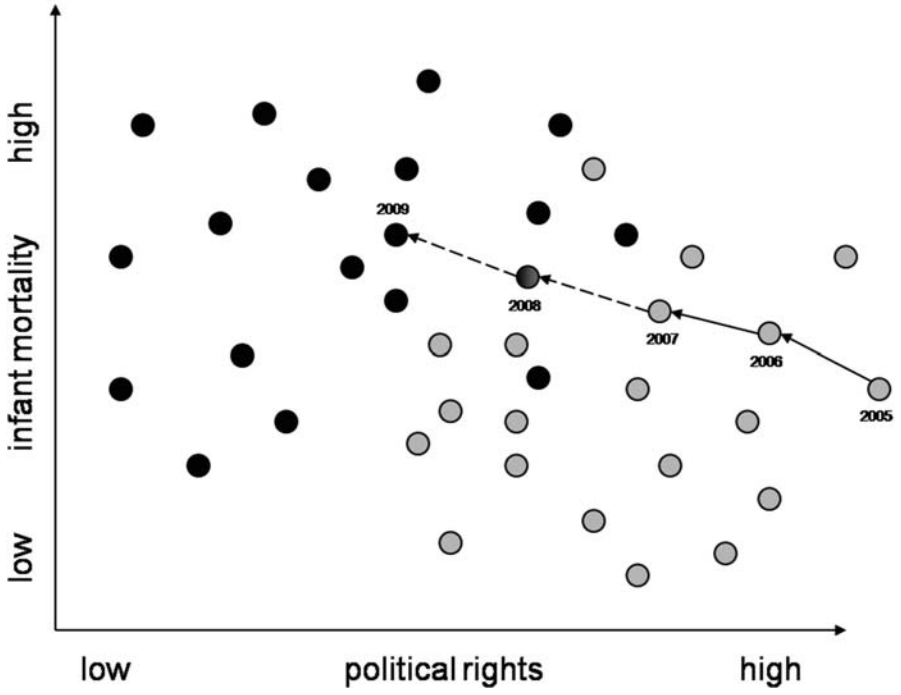



**Fig. 12.6.** k-Nearest Neighbor Algorithm: Path to Conflict. The historical path of the nation state shows periods of peace in 2005, 2006, and 2007 (*grey points*), but also shows the state on a trajectory towards instability induced conflict. The forecasted path classifies the conflict potential as uncertain in 2007 (*black/grey point*) and high (*black point*) in 2008.

either dark grey (uncertain conflict potential) or black (high conflict potential) based upon the decision rule, and classifies 2009 as black (high conflict potential). This nation-state is on a rapid path to the black (high conflict potential) cloud, expected to spend just one year in the intersection before leaving the grey (low conflict potential) cloud.

### Nearest Centroid algorithm

The NC algorithm classifies each future feature vector as a function of it's proximity to the centroid of each cloud. We define the centroid of a cloud as the mean vector of the past feature vectors in that cloud. This function, or decision rule, classifies a future feature vector as the class of the nearest centroid. This decision rule always yields a classification. We provide an artificial two-dimensional Conflict Forecast example of the NC algorithm in Fig. 12.7.

First, we plot the historical feature vectors (Political Rights, Infant Mortality) and "color" each two-dimensional point based upon the known level of past conflict. Second, we forecast a future feature vector for a given nation-state and plot this two-dimensional point. Third, we identify the centroid for each

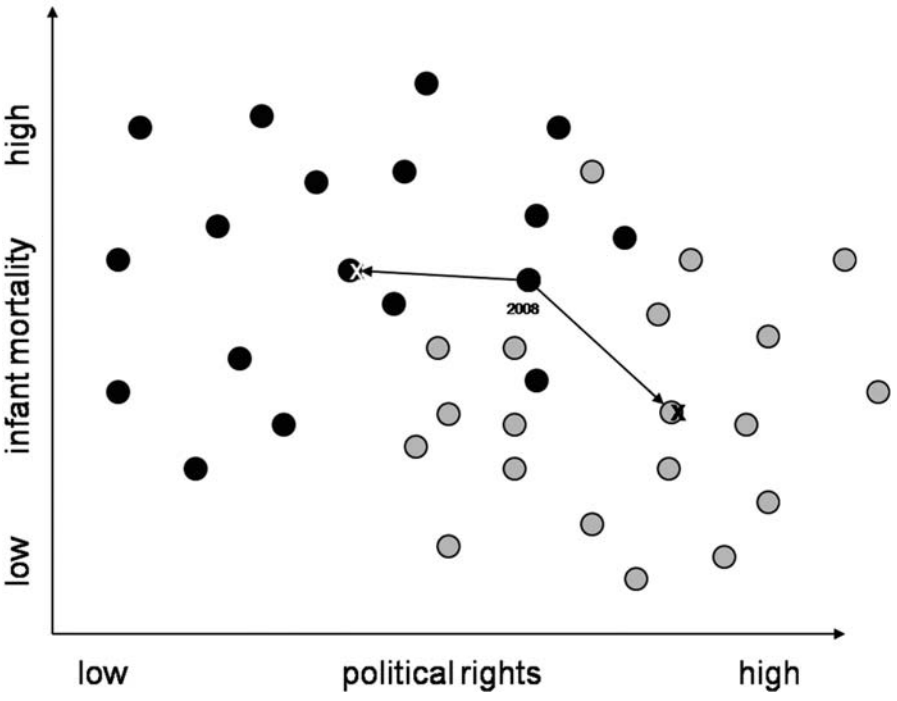



**Fig. 12.7.** Nearest Centroid Algorithm: Conflict Forecast. The forecasted 2008 feature vector lies in closer proximity to the centroid of the conflict cloud (*black points*), resulting in a classification of high conflict potential.

of the two clouds (designated with an "x" in Fig. 12.7). Fourth, we determine the distance from the future feature vector to each of the two centroids. Fifth, we classify the future feature vector the same class as the nearest centroid. The 2008 forecast lies closest to the black centroid leading to a black (high conflict potential) classification. We provide an artificial two-dimensional Path to Conflict example of the NC algorithm in Fig. 12.8.

First, we plot the historical feature vectors (Political Rights, Infant Mortality) and "color" each two-dimensional point based upon the known level of past conflict. Second, we forecast the future feature vectors for a given nation-state and plot these two-dimensional points. Third, we identify the centroid for each cloud. Fourth, we determine the distance from each future feature vector to each of the two centroids. Fifth, we classify each future feature vector the same class as the nearest centroid. This nation-state is on a rapid path to the black (high conflict potential) cloud, expected to leave the grey (low conflict potential) cloud in one year and remain outside the grey cloud for the forecast future.

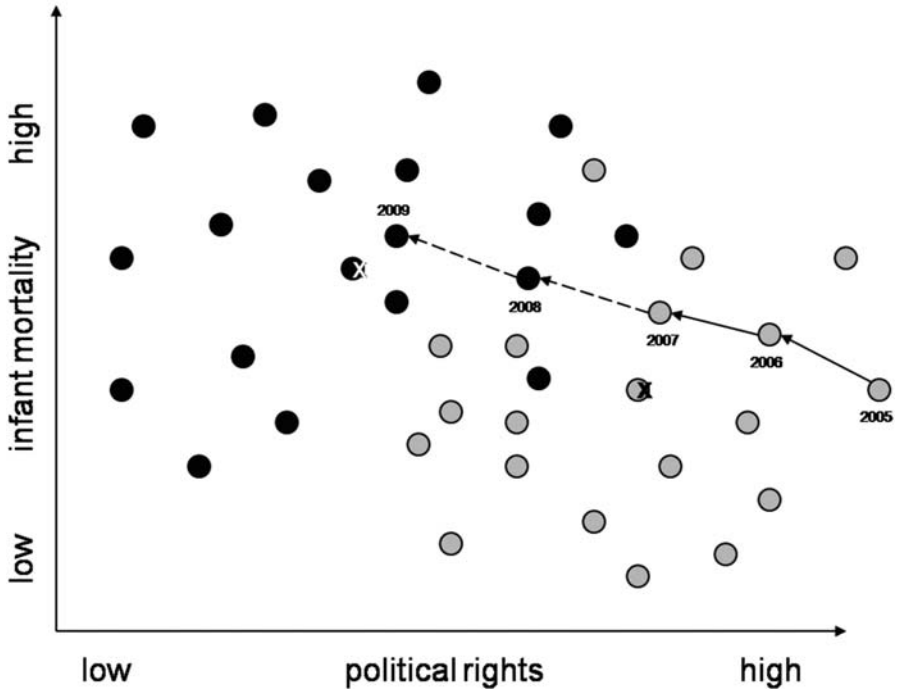



**Fig. 12.8.** Nearest Centroid Algorithm: Path to Conflict. The historical path of the nation state shows periods of peace in 2005, 2006, and 2007 (*grey points*), but also shows the state on a trajectory towards instability induced conflict. The forecasted path classifies the conflict potential as high in 2007 and 2008 (*black points*).

## Validation

The research team conducted two validation efforts. First, we determined the accuracy of the kNN and NC classifiers and selected the most accurate for use in the research. Second, we determined the accuracy of the overall research methodology with both simple and super majority decision rules.

### Classifier accuracy

We determined the accuracy of the two classifiers assuming perfect knowledge of the future, utilizing the known future feature vectors from the test data. This allowed us to establish an upper bound for the accuracy of the classifiers and would later allow us to determine how much of the classification error was due to the future feature forecast and how much was due to the classifier. We divided the data into a training set (1993–1997) and a test set (1998–2003) for validation purposes. The training set contained five years of past feature vectors for all 155 countries, for a total of 775 past feature vectors. The test set contained six years of known future feature vectors for all 155 countries, for a total of 930 known future feature vectors. We trained both classifiers on the training set, forecasted the future feature vectors from the training set with the WMA and evaluated the future feature vectors (both forecasted and known) with both classifiers. We classified with the kNN algorithm at $k = 3$, $k = 5$, $k = 7$ and $k = 9$ utilizing a simple majority decision rule and all variants of the super majority decision rules. We utilized a simple majority decision rule and calculated the accuracy of each algorithm as the percentage of the correct classifications. Table 12.1 contains these results.

The KNN algorithm outperformed the NC algorithm at all four different k values. The KNN algorithm classified with greater than ninety percent accuracy for all four parameter values, with $k = 5$ achieving the highest accuracy of 93%. We adopted the KNN algorithm with $k = 5$ for the rest of the research. Interestingly, the NC algorithm did outperform KNN in accurately forecasting conflicts.

### Forecast accuracy

We then determined the accuracy of the methodology, utilizing the WMA to forecast feature vectors and the KNN algorithm with $k = 5$ to classify, with both simple and super majority decision rules. The classifications made with a simple majority decision rule all yielded conflict or peace classifications. We calculated the accuracy of these classifications as the percentage of the correct classifications The classifications made with super majority decision rules yielded three classifications: peace, conflict, and uncertain. We calculated the accuracy of these classifications as the percentage of the correct peace and conflict classifications. We calculated the uncertainty of these classifications as the percentage of uncertain classifications. Table 12.2 contains these results.

**Table 12.1.** Classifier accuracy: 1998–2003

| k = 3 | Correct | % | Incorrect | % |
|---|---|---|---|---|
| Peace | 1240 | 95 | 63 | 5 |
| Conflict | 339 | 84 | 64 | 16 |
| Total | 1579 | 93 | 127 | 7 |
| k = 5 | Correct | % | Incorrect | % |
| Peace | 1242 | 95 | 61 | 5 |
| Conflict | 338 | 84 | 65 | 16 |
| Total | 1580 | 93 | 126 | 7 |
| k = 7 | Correct | % | Incorrect | % |
| Peace | 1228 | 94 | 75 | 6 |
| Conflict | 330 | 82 | 73 | 18 |
| Total | 1558 | 91 | 148 | 9 |
| k = 9 | Correct | % | Incorrect | % |
| Peace | 1230 | 94 | 73 | 6 |
| Conflict | 331 | 82 | 72 | 18 |
| Total | 1561 | 92 | 145 | 8 |
| NC | Correct | % | Incorrect | % |
| Peace | 1000 | 77 | 303 | 23 |
| Conflict | 351 | 87 | 52 | 13 |
| Total | 1351 | 79 | 355 | 21 |

**Table 12.2.** Forecast and classifier accuracy: 1998–2003, k = 5

| ≥ 3 | Correct | % | Incorrect | % | Uncertain | % |
|---|---|---|---|---|---|---|
| Peace | 662 | 89 | 83 | 11 | n/a | n/a |
| Conflict | 147 | 79 | 38 | 21 | n/a | n/a |
| Total | 809 | 87 | 121 | 13 | 0 | 0 |
| ≥ 4 | Correct | % | Incorrect | % | Uncertain | % |
| Peace | 620 | 91 | 59 | 9 | n/a | n/a |
| Conflict | 155 | 84 | 29 | 16 | n/a | n/a |
| Total | 775 | 90 | 88 | 10 | 67 | 7 |
| = 5 | Correct | % | Incorrect | % | Uncertain | % |
| Peace | 589 | 93 | 47 | 7 | n/a | n/a |
| Conflict | 147 | 86 | 24 | 14 | n/a | n/a |
| Total | 736 | 91 | 71 | 9 | 123 | 13 |

The KNN algorithm with k = 5 and forecasted future feature vectors achieved 87% accuracy with a simple majority decision rule ("greater than or equal to three"), down from the 93% accuracy achieved when the known future features vectors were classified. The forecast and the classifier contribute equally to the overall classification error.

The super majority decision rules ("greater than or equal to four" and "equal to five") yielded higher accuracy rates than the simple majority decision rule. The improved accuracy did come with a cost. Increasing accuracy from 87% to 90% to 91% resulted in an increase in uncertain conflict potential classifications from 0% to 7% to 13%.

### Classification errors

The team was concerned with more than just the overall accuracy of various algorithm/decision rule combinations. We were also concerned with what type of classification errors we made. Two types of errors exist. One can classify peace and have conflict occur and one can classify conflict and have peace occur. The KNN algorithm with k = 5 and an "equal to five" decision rule incorrectly classified 71 nation-states over the six year test set. One third of the classification errors were missed conflicts; two thirds of the errors were missed peaces.

### Conflict forecasts and paths to conflict

The team then developed Conflict Forecasts and Paths to Conflict for all 155 nation-states out to 2020. We forecasted the future feature vectors for each nation-state utilizing WMA and classified these future vectors with a KNN algorithm (k = 5) and a super majority decision rule ("equal to five"). Figures 12.9, 12.10, and 12.11 show the classifications that make up these Conflict Forecasts for all 155 nation-states in 2010, 2015, and 2020.

## Armenia

We close with an in depth analysis of the path to conflict of one nation-state, Armenia, to show how the methodology works, not from a mathematical perspective, but rather from a visual one. The HIIK database has not reported conflict in Armenia since 1994 and yet our Conflict Forecast for the nation-state is, paradoxically, high conflict potential by 2020. What does the methodology "see" that justifies this forecast? Figure 12.12 shows the PCA transformed space through which Armenia moved from the 1998 to 2003. This section of the space is positioned in the grey cloud (peace), just outside the intersection. The spheres represent nation-state – years.

Armenia (1998) is located in the bottom left of Fig. 12.12, colored grey to denote a year of peace. Armenia (1999–2001) also experienced peace, remaining in the grey cloud. Armenia (2002–2003) continued to experience peace, but shifted it's path, turning toward the intersection of the two clouds. The

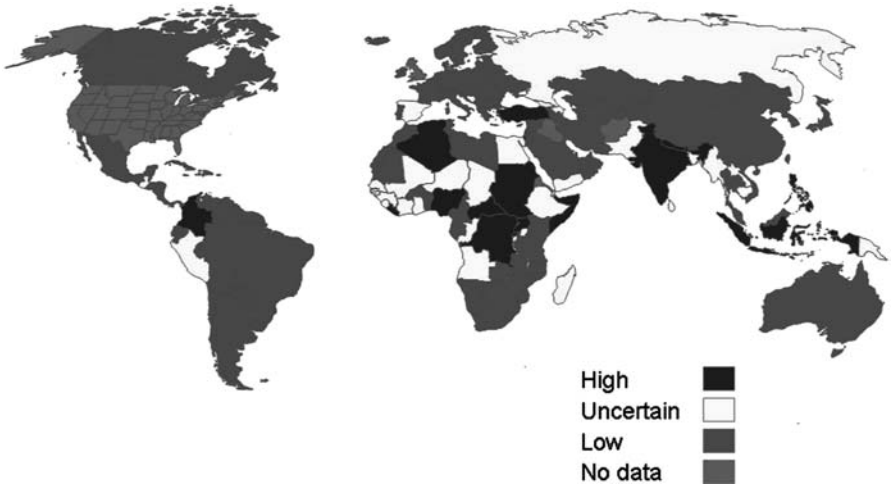



**Fig. 12.9.** Conflict Forecast: 2010. *Black nation* states have high conflict potential, *dark grey* states have uncertain conflict potential, and *light grey* states have low conflict potential.

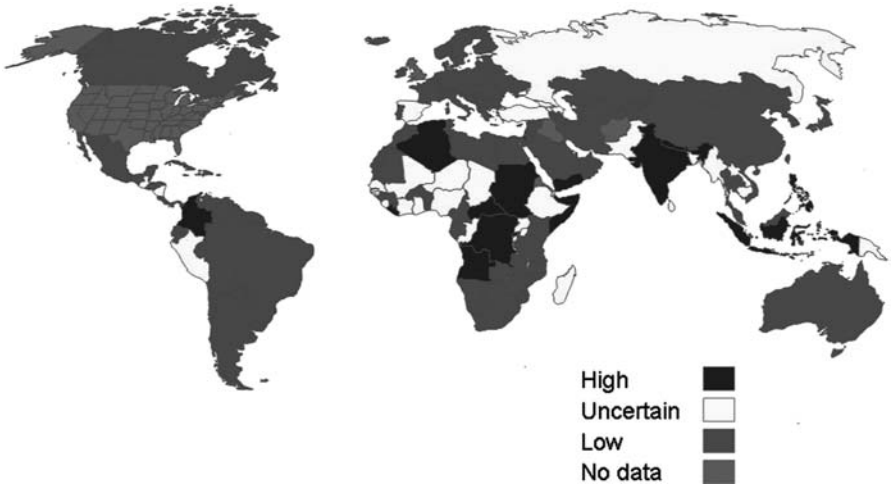



**Fig. 12.10.** Conflict Forecast: 2015. *Black nation* states have high conflict potential, *dark grey* states have uncertain conflict potential, and *light grey* states have low conflict potential.

path forecasted with WMA placed Armenia in the intersection by 2015 with a majority of nearest-neighbors having experienced conflict by 2020. These neighbors include Yugoslavia (1997), Somalia (1997), Haiti (1993) and Iran (1997), all of which experienced conflict. The direction of Armenia's trajec-

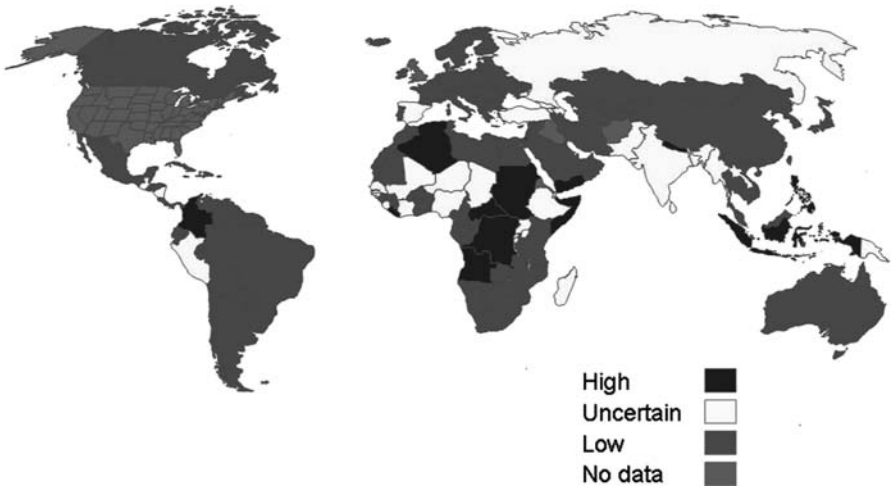



**Fig. 12.11.** Conflict Forecast: 2020. *Black nation* states have high conflict potential, *dark grey* states have uncertain conflict potential, and *light grey* states have low conflict potential.

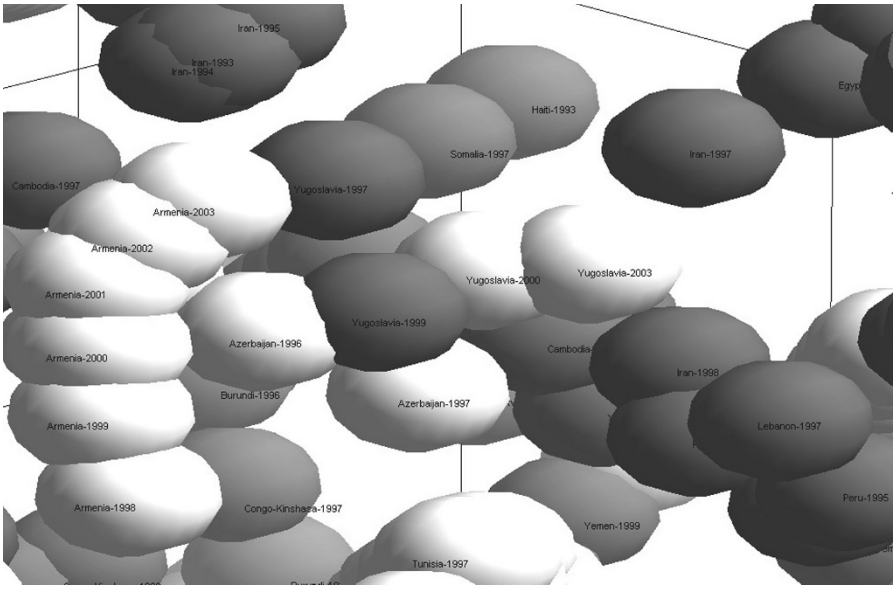



**Fig. 12.12.** Path to Conflict: Armenia. Each ball represents a nation state for a given year. *Dark colored* balls represent nation states that experienced conflict in the given year. *Light colored* balls represent nation states that experienced peace in the given year. The path that Armenia has followed since 1998 begins in the lower left of the figure.

tory increases the potential for the country to experience state failure and instability induced conflict, and might offer safe haven in Europe to terrorist organizations in the next decade.

## Conclusion

While the results of this analysis are predictive, they are not prescriptive. The methodology described here can identify those nation-states susceptible to instability induced conflict with a relatively high degree of accuracy and also help anticipate likely terrorist safe havens. It does not provide a solution to the problem of nation-state failure, nor does it answer the question of how to deal with terrorist safe havens once they have emerged. Additional work is required to answer these more challenging questions, we have here but identified the portion of the world in which to focus attention.

## References

Abraham, B., Ledolter, J.: Statistical Methods for Forecasting. Wiley, New York (1983).

Carleton University. Country Indicators of Foreign Policy. `http://www.carleton.ca/cifp/`. Cited 15 November 2007.

Duda, R., Hart, P., Stork, D.: Pattern Classification. Wiley-Interscience, New York (2001).

Food and Agriculture Organization of the United Nations. Food Security Statistics. `http://www.fao.org`. Cited 15 November 2007.

Freedom House. Freedom in the World. `http://www.freedomhouse.org`. Cited 15 November 2007.

Goldstone, J., Gurr, T., Harff, B., Levy, Marc., Marshall, M., Bates, R., Epstein, D., Kahl, C., Surko, P., Uhlfelder, J., Unger, A.: State Failure Task Force Report: Phase III Findings. Science Application International Corporation, McLean, Virginia (2000).

United States Bureau of the Census. International Database. `http://www.census.gov/ipc/www`. Cited 15 November 2007.

Heidelberg Institute of Conflict Research. KOSIMO. `http://www.hiik.de`. Cited 15 November 2007.

O'Brien, S.: Anticipating the Good, the Bad, and the Ugly: An Early Warning Approach to Conflict and Instability Analysis. *Journal of Conflict Resolution.* **46**(6):791–811 (2002).

Polity IV Project, Center for International Development and Conflict Management. `http://www.systemicpeace.org/polity/polity4.htm`. Cited 15 November 2007.

Central Intelligence Agency. The CIA World Fact Book. `https://www.cia.gov/library/publications/the-world-factbook/index.html`. Cited 15 November 2007.

World Bank. 2007. World Bank Development Indicators 2007. `http://www.worldbank.org`. Cited 15 November 2007.

# 13

# Applied Counterfactual Reasoning

Noel Hendrickson

**Summary.** This chapter addresses two goals: The development of a structured method to aid intelligence and security analysts in assessing counterfactuals, and forming a structured method to educate (future) analysts in counterfactual reasoning. In order to pursue these objectives, I offer here an analysis of the purposes, problems, parts, and principles of applied counterfactual reasoning. In particular, the ways in which antecedent scenarios are selected and the ways in which scenarios are developed constitute essential (albeit often neglected) aspects of counterfactual reasoning. Both must be addressed to apply counterfactual reasoning effectively. Naturally, further issues remain, but these should serve as a useful point of departure. They are the beginning of a path to more rigorous and relevant counterfactual reasoning in intelligence analysis and counterterrorism.

## Introduction

If Iran had nuclear weapons, would they supply this technology to terrorists? If the U.S. has not abolished the Iraqi army in 2003, would that have preempted the growth of the insurgency? Questions like these are sometimes said to be 'counterfactual,' for they ask about an alternate scenario and its consequences.[1] These sorts of questions are essential in intelligence analysis and counterterrorism. In such endeavors, the primary goal is (usually) to understand the actions of human agents, who have many potential actions at their disposal. Thus, there is rarely (if ever) simply one possible way that things might develop. Instead, there are a range of alternatives and consequences.

Counterfactual questions pose an especially difficult challenge. After all, the relevant scenarios have not occurred, and often never will. Thus, it might seem that analysts can only recourse to their own powers of imagination and simple probabilistic extrapolation to answer them. However, it is possible to

[1] While the word 'counterfactual' may seem to imply considering something that is 'counter to fact,' the term has come to refer to all subjunctive conditionals. Hence, the specified alternative could be one that actually does (or will) obtain.

S. Argamon, N. Howard (eds.), *Computational Methods for Counterterrorism*, DOI 10.1007/978-3-642-01141-2_13,

have a thorough and rigorous strategy for evaluating counterfactuals. In the past 30 years, analytic philosophers and logicians have constructed sophisticated accounts of counterfactuals.[2] But, unfortunately, no one has attempted to develop these tools further so that they can address the relevant real-life cases.[3] So, in this chapter, I describe some major points from a broader research program that attempts to do just that. I offer some basic principles of counterfactual reasoning for counterterrorism and intelligence analysis: a structured method for evaluating alternate scenarios and their consequences that can give analysts the confidence of knowing that they have rigorously and thoroughly covered the relevant possibilities. I propose some fundamentals of an account of applied counterfactual reasoning.

## The purposes of applied counterfactual reasoning

While philosophers and logicians may study counterfactual reasoning simply to clarify purely conceptual principles, intelligence analysts do not. They apply counterfactual reasoning to guide real-life decisions. And, there are four major ways that counterfactual reasoning aids that process.

The first purpose of applied counterfactual reasoning is to *improve causal inferences*. In analyzing an event of interest, investigators often face an oversupply of potentially relevant factors. They need a strategy for determining when a factor is actually affecting the outcome of interest. But, inferring causal connections can be extremely difficult as neither statistical nor temporal correlations are sufficient to demonstrate a real connection. Counterfactual reasoning provides a possible solution. For, if event e is counterfactually dependent on factor f (i.e., if f had not obtained, then e would not have occurred), then f is reasonably taken to be a cause of e.[4] For example, if it is true that the Iraqi insurgency would not have reached the strength that it did in 2006 if

[2] Major sources include Stalnaker (1986), Lewis (1973, 1979, 1986), Kvart (1986, 1992, 1994) and Bennett (2003).

[3] A few social scientists have attempted to tackle this challenge from a different direction. Instead of trying to make the rigorous conceptual principles more applicable, they have worked on making the discussion of real-life cases more sophisticated. A sample of major political/social science sources includes Fogel (1964), Elster (1978), Hawthorn (1991), Fearon (1991, 1996), Tetlock and Belkin (1996), Khong (1996), Lebow (2000, 2001) and Lebow and Breslauer (2004). For sources from the 'virtual history movement,' consider Ferguson (1999); as well as (for example) Cowley (2000, 2002, 2004) and Tsouras (2001, 2002).

[4] Note that this does not assume that counterfactual dependence is the same thing as causation. It seems not to be. Since causation is transitive, and counterfactual dependence is not, the two must be distinct. Despite this, counterfactuals can still be useful in improving causal inferences if counterfactual dependence implies causation. The further claim that causation implies counterfactual dependence is not necessary.

the U.S. had not abolished the Iraqi army in 2003, then the abolishing of the army is one of the causes of the strength of the insurgency in 2006. Counterfactual dependence implies causation, and so counterfactual reasoning serves as a useful strategy to separate genuine causes from mere correlations.

The second purpose of applied counterfactual reasoning is to *substantiate after-action reports*. Analysts often have to consider how to respond better to future threats that were not adequately addressed in the past. In such instances, they often consider what could have been done in prior instances in order to avoid the negative outcome. Or, they imagine the absence of a seemingly successful strategy to determine if it was actually necessary for a positive resolution. Both of these employ counterfactual reasoning, for they examine an alternate scenario and its consequences. So, counterfactual reasoning may be applied in order to evaluate the effectiveness of an operation by showing what would have occurred if it had been done differently.

The third purpose of applied counterfactual reasoning is to *guide futures analysis*. Intelligence analysts are not only interested in the alternate ways that past scenarios could have developed differently. There is also substantial concern with the consequences of future alternatives.[5] As a result, counterfactual reasoning may also be employed in order to structure estimates by means of rigorous assessments of what would occur in different future alternate scenarios.

The fourth purpose of applied counterfactual reasoning is to *encourage innovative thinking*. Responding to potential terrorist threats requires an ability to think beyond what simply has been employed in prior security efforts (such as in the Cold War or its immediate aftermath). But, innovative thinking in intelligence analysis faces an important challenge. For, analysts cannot simply resort to open-ended 'brainstorming.' Creative thinking is effectively employed only if it is amenable to rigorous rational procedures. Counterfactual reasoning has a potential solution. It employs both an investigation of possibilities and a structured framework for selecting which of those possibilities are worthy of further consideration. Thus, analysts can use it to develop sensitivity to previously unrecognized possibilities and their consequences, as well as to reason about those possibilities *with precision*.

[5] Strictly speaking, there are some important differences between past possible alternatives and future possible alternatives, since the latter still have the potential to actually obtain (whereas the former do not). As a result, it is sometimes possible to test claims about the latter empirically (i.e., with what actually happens). However, there is still a major *structural* similarity between analyzing past possible alternatives and analyzing future possible alternatives. And, all that I have claimed here is that counterfactual reasoning helps to frame futures estimates properly.

## The problems of applied counterfactual reasoning

Counterfactual reasoning poses many challenges. And, initially, one might think that most will surround the way in which we determine the scenario's final outcome. After all, there are usually many ways in which a particular alternative could develop, and so the selection of scenario outcomes might seem to be the primary difficulty. However, I propose that this is actually not the main problem in counterfactual reasoning. Instead, the major work of applied counterfactual reasoning occurs earlier on in the process. This leads to what I will call the 'problem of antecedent scenarios' and the 'problem of intermediate states.'

There are usually many ways that a particular alternative could have come to be. For example, there is not simply one way that 'Iran has nuclear weapons' could be true. Iran might develop such weapons completely on their own, *or* they could purchase them on the black market, *or* they could purchase parts on the black market and finish developing them on their own, *or* a foreign country could give them these weapons, and so on. In addition, any one of these could occur in the presence of strong UN sanctions against Iran or in the absence of such sanctions. This is only the beginning of the possible ways in which it might come to be that 'Iran has nuclear weapons.' A similar multiplicity of cases is possible with 'the U.S. does not abolish the Iraqi army in 2003.' This might obtain by keeping all military members in place (apart from Saddam and his immediate underlings), *or* by the U.S. keeping all of the unlisted members in place, *or* by keeping everyone who entered the army after the first Gulf War in place, *or* by ending the army but still paying everyone until they could find new jobs, and so on. In both cases, there are many different possible 'antecedent scenarios'—ways in which the alternative of interest could come to be.

One might think that we can simply arbitrarily select one or more of these scenarios for consideration. However, the selection has the potential to make a substantial difference to the final outcome. For example, the implications of a nuclear Iran in the presence of strong UN sanctions are (likely) different from the consequences of nuclear Iran without such sanctions. As well, a nuclear Iran that acquired the technology from another country is (likely) different from one that developed them completely on their own, and so on. In a similar way, the implications of abolishing only the top members of the Iraqi army are (likely) different from the implications of only abolishing those positions filled by people present during the first Gulf War. Now, there is nothing wrong with asking specifically about the consequences of a particular antecedent scenario. However, in most cases, when we are simply wondering about the consequences of a nuclear Iran or not abolishing the Iraqi army (without a particular scenario in mind), there has to be a way to select among the possible options; time does not usually permit us to consider them all (especially since there are a potentially infinite number of options). Therefore, there needs to be a rigorous way to choose. This is the *problem of antecedent scenarios.*

Once one has selected a specific antecedent scenario (or a limited set of scenarios), it may be tempting to immediately begin considering the possible outcomes of that scenario. But, that also would be premature. For, there is (usually) a substantial time gap between obtaining the antecedent scenario and obtaining the timeframe of interest for the scenario outcomes. For example, one might be interested in the five year consequences of a nuclear Iran (or of an Iraq without the U.S. having abolished their army). There are many events that are already known (or projected) to (likely) occur during that timeframe. Some of these events (such as those estimated, say, for Iraq) are likely to be affected by our alternate scenario with Iran (and *vice versa* with the Iraq case). Other events (such as those extrapolated for, say, Mexico) are likely not to be (in general) influenced. And, there is a large range of events that could easily go either way (such as those with other potential nuclear powers like North Korea, and new potential allies for Iran like Venezuela). All of these 'intermediate states' will (or at least could) likely make a substantial difference to the ultimate resolution of the scenario. For instance, the extent to which one has to also deal with new (or greater) difficulties with North Korea and Venezuela, it will likely follow that there is a more limited range of options available with Iran. By contrast, if other countries are likely not to present difficulties, then there would be a much greater range of possible responses to a nuclear Iran. Therefore, there needs to be a rigorous way to choose which of these states should also be present in our alternate scenario. This is the *problem of intermediate states*.

## The parts of applied counterfactual reasoning

There are three stages to applied counterfactual reasoning. My characterization of these stages follows directly from the centrality of the two proposed problems in counterfactual reasoning. These two challenges are so critical to thinking effectively about alternative scenarios that they serve to divide counterfactual reasoning into three stages. Now, there is always some arbitrariness in any division of the thinking process. However, that does not mean that such separations are without merit or reason. Rather, they are done in order to help address major challenges. In this case, the division is made to help analysts recognize the importance of the problem of antecedent scenarios and the problem of intermediate states.

Analyzing an alternate scenario begins with a possible event (antecedent) and time-frame for which one wishes to know its outcome (consequent). The first part of counterfactual reasoning is to *select antecedent scenarios*. One does not begin evaluating the nuclear Iran scenario by imagining the possible outcomes and weighing them. Instead, one first explores the ways in which that antecedent could have come to be. And, most importantly, one determines which constitute the relevant antecedent scenario. One must begin with a principled construction of the scenario backstory.

The second part of counterfactual reasoning is to *develop scenarios*. Once a particular antecedent scenario has been selected, one will again be tempted to transition immediately to examining which of the potential scenario outcomes are most likely. However, one must first develop the scenario further in order to fill in the events during the time-gap between the antecedent scenario and the time of interest for the outcome. Since many attempts at counterfactual reasoning will involve projecting over months, years, or even decades, the events that are included in this period are critical to the final outcome. Thus, in this stage, one examines the potential intermediate events that hold during this period and establish those that hold in the scenario.

It is only in the third part of counterfactual reasoning that one directly addresses the thing that (initially) seems to be the primary object of our attention: to *establish scenario outcomes*. There are often many potential ways that a scenario can end, and there will have to be a way to generate and select among those possibilities. So, in this stage, one evaluates the possible outcomes to select those that are most likely given the antecedent scenario and intermediate states.

## The principles of applied counterfactual reasoning

Each stage of counterfactual reasoning has its own set of principles that govern the most effective thinking in that stage. This is a further reason to separate the stages as I have. One has to keep the different principles confined to their appropriate stage. What is done at one stage will not (in most cases) be appropriate at a later stage. Naturally, there are many other important principles in counterfactual reasoning that space does not permit me to discuss. These are simply some of the most basic tools needed to form a useful account of applied counterfactual reasoning.

### Principles for selecting antecedent scenarios

To begin selecting antecedent scenarios, one should ($P_1$) *maximize the range of possible scenarios*. In describing the possible ways that the antecedent could come to be, one should try to cover as many possibilities as time permits. It is essential not to assume (implicitly) that any scenario is not the right one by failing to consider it altogether. Instead, scenarios are to be excluded only if one has an explicit reason for so doing. In particular, one must never limit the options simply to what has happened historically. A certain amount of imagination is always useful at this point. However, it is important to note that an excess of creativity is not dangerous here, since rigorous standards will be applied to make the ultimate selection from these possible options. Hence, any unjustified scenario will eventually be excluded. The thing to avoid at this point is not a failure to exclude but a failure to include. While there will

be numerous opportunities to dispense with a scenario later, there will not be room to add any new ones to the process.

Once one has a wide range of possible ways that the antecedent could have come to be, one has to *weigh the plausibility of those scenarios with multiple criteria*.[6] I offer a series of different standards (in order of both importance and application) that should be employed in weighing the possible options. First, ($P_2$) *exclude antecedent scenarios that contain highly improbable events*. While it is obviously theoretically possible that Iran have nuclear weapons because Israel gives them the technology, it is so incredibly improbable (at present) that there would be little point to evaluating such a scenario. It is simply not what we are referring to when we ask about the consequences of a nuclear Iran. Now, it is important to be clear that this principle does not exclude all improbable events, but only those that are *highly* improbable. And, I have specifically chosen not to define that notion further in order to ensure that it is clear to anyone applying these principles that they must choose carefully and never with excessive overconfidence.[7]

Second, one should ($P_3$) *prefer antecedent scenarios that reach the antecedent faster*. One way to compare antecedent scenarios is by using the

[6] At this point, I markedly part company with the dominant Lewis-Stalnaker 'similarity' theory of counterfactuals (at least in selecting antecedent scenarios). In addition, I do not embrace the rival 'probability' theory of Igal Kvart. On the former view, we are to select as the antecedent scenario(s) what obtains in the possible world most similar to ours (or, the states that obtains in the worlds such that there are no more similar antecedent worlds that do not contain those states). On the latter view, we are to select the antecedent scenario(s) that offer the shortest path to the antecedent that maximizes its probability (or, the antecedent scenario that makes the antecedent most probable from among those that have the highest probability for the least probable link in the sequence that leads to the antecedent). My account will make use of some of the concepts that are in each of these accounts, but in a different way. However, I also add further notions that are not present in either view.

[7] For a historical example of why this principle is necessary (and, for instance, what is wrong with the Lewis-Stalnaker approach of the most similar world), consider Chamberlain's appeasement of Hitler at the Munich conference in the summer of 1938. Which of the many antecedent scenarios are we to select if we imagine that Chamberlain fails to appease Hitler? If we prefer the scenario that maximizes similarity to the actual world (or want the shortest run-up to the antecedent), then we would imagine that Chamberlain simply makes a radically out of character break with his policy and issues an ultimatum to Hitler. However, it seems clear that this is not the scenario we should use. Instead, we should imagine a longer scenario in which Chamberlain has a different Cabinet (urging him to adopt a different policy), or perhaps is not even prime minister. But, in any case, we should not accept the scenario in which he makes the highly improbable (for him) choice to fail to appease Hitler (Although this principle is my own, for a counterfactual evaluation of this scenario that is consistent with it, see Khong (1996).).

amount of time that it takes to reach the ultimate antecedent (e.g., a nuclear Iran) from the original point of departure. The shorter scenarios are preferable. We could imagine a nuclear Iran by means of a completely different origin for the nuclear bomb in which Iran is the first to develop it. But, that is obviously not the sort of thing one should be evaluating. A shorter path seems preferable. In addition, the shorter the scenario, the greater the amount of prior history that is persevered, and hence the more (sure) information that can be used in making projections. For with every alteration, there is greater potential for mistakes. Therefore, a shorter scenario has the added benefit of less potential for error.

While substantial differences in scenario length seem reasonably resolved in favor of the shorter option, more modest differences in length are much more complex. In these cases, there seem to be other factors that can contribute to the plausibility of an antecedent scenario. Furthermore, these sometimes are worth having at the expense of a longer scenario. So, third, one should ($P_4$) *prefer antecedent scenarios that have fewer independent sequences of events leading to the antecedent (i.e., greater scenario unity).* In generating the antecedent of interest, there is (usually) more than one sequence of events that will converge to generate that possibility. However, in selecting among different possible ways to lead to an antecedent, the fewer such sequences, the better. For example, in evaluating the consequences of a leader's absence, it is often better to imagine their assassination than their being voted out of office. It often requires only a narrow range of alterations to generate a successful assassination, whereas a large number of factors need to be altered to change an election outcome (especially if there actually were attempts to assassinate that leader and the election was a landslide). In such a scenario, one can say that the changes trace back to a *single* event (an assassination) rather than to a series of different events (economic or social factors needed to change a landslide election). In other words, two sequences are independent of each other if and only if they begin with different events (i.e., all later events would then be dependent upon that initial deviation.).[8] Now, less independent event sequences are not always preferable. Occasionally, fewer independent sequences will actually require a much longer period of time than would more independent sequences. And, in those cases, the greater *unity* to the scenario is likely not worth the tradeoff. Much like shorter scenarios, choosing a more unified

[8] For a historical example of why this principle is necessary, consider the consequences of no Reagan or Gorbachev in the middle 1980s. Would we prefer Reagan being successfully assassinated in 1981, or his being voted out of office in 1984? Suppose that we could intervene after the assassination attempt and construct a (shorter) path to the antecedent without any improbable events (leading to an election loss in 1984), would that be the right antecedent scenario? It would seem that such a scenario is not to be chosen since it would require (likely) a series of different converging sequences of events instead of merely one (Although this principle is also my own, for a counterfactual evaluation of this scenario that is consistent with it, see Lebow and Breslauer (2004).).

scenario also has the benefit of having less potential for error, since the more independent sequences of events that converge to generate one's antecedent, the greater potential there is to project the wrong outcome.

Every antecedent scenario begins with a deviation away from what we know (or project) will be the case. We have already seen that it is (generally) better to have a more recent deviation. But, in addition, it is also sometimes better to prefer a more probable deviation. That is, fourth, one should ($P_5$) *prefer antecedent scenarios with a higher probability for the first event in the sequence leading to the antecedent.* We have already excluded scenarios that involve highly improbable events, so this involves something beyond that. Suppose that one can generate a nuclear Iran by means of two scenarios of (roughly) the same length and unity. If one can also select one whose triggering event has a higher probability, then that is surely worthwhile. Now, like scenario unity, this is not typically worthwhile if it requires a much longer scenario. But it is probably worth tolerating a bit more length.[9]

### Principles for developing scenarios

In developing scenarios, the primary challenge is to determine which intermediate states should be added to the antecedent scenario (to create the basis from which the scenario outcome will be projected). A major source of potential intermediate states will be the set of events that are already known (or projected) to (likely) occur between the antecedent scenario and time of interest for the outcome. Which of these events should we include in our scenario? The first task is to ($P_6$) *include independent events as intermediate states.* An event is 'independent' if and only if: (1) it is known to (likely) occur during the timeframe of interest and (2) its probability is unaffected by the antecedent scenario.[10] That is, the event's probability is no different than it was when

[9] For a historical example of why this principle is necessary, consider Archduke Ferdinand's not being assassinated in the Summer of 1914. As the famous story goes, an initial attempt to assassinate Ferdinand is unsuccessful, and the conspirators scatter. Ferdinand begins his return trip but decides to visit those who were injured in the initial attack. His driver is unaware of the change and makes a wrong turn, and as he turns back in the right direction, the car is noticed by one of the remaining conspirators, who then successfully assassinates the Archduke. Now, there is a short path to the antecedent that is no less unified and has no highly improbable events: postulate that the conspirator's bullet misses the Archduke. Ultimately, though, this seems wrong. It is better to reverse an improbable event (postulate no motorcade driver's mistake) rather than to reverse a more probable one (postulate that the assassin's point-blank shot misses Ferdinand) *even though the latter would involve a shorter scenario* (Although this principle is also my own, for a counterfactual examination of this scenario that is consistent with it, see Lebow (2001).).

[10] For an event to be independent, it is probably not enough that (for antecedent $A$ and event $E$) $Prob(E/A) = Prob(E/\neg A)$. This equivalency might come about

one does not imagine any alternate possibility. For example, if some sequence of events is projected in Mexico, and their probability is same regardless of whether we postulate a nuclear Iran or not (or abolishing the Iraqi army or not), then such events should be included in the scenario.

There is an important potential mistake that should be noted at this point. Events should be considered for inclusion or exclusion from the scenario by *time* and not *topic*. That is, all events at a particular time should be considered together, rather than considering together events at different times (but, say, involving the same country). For even if the antecedent scenario *alone* does not affect the probability of an event, the antecedent scenario *combined with intermediate states* might do so. In addition, whether an event should be included is a matter of three different principles (not simply the rule that independent events should be included). Hence, one should consider each event at a particular time in terms of each of the three principles for developing scenarios. Then, one moves to a later time and considers those events in terms of the three principles *having added the previously included events (that are intermediate states) to the antecedent scenario.*

Even if an event is not independent, it may still merit inclusion. So, the second task in developing scenarios is to ($P_7$) *include positively affected events as intermediate states.* An event is 'positively affected' if and only if: (1) it is known to (likely) occur during the timeframe of interest and (2) its probability is increased by the antecedent scenario.[11] If an event has a different probability because of the antecedent scenario, but only because its probability is increased, then it is reasonable to include it. If we project a strong Iranian influence in Iraq, and the probability of that is only increased by imagining a nuclear Iran, then it is reasonable to develop the nuclear Iran scenario in a way that includes a strong influence over Iraq. Now, once again, this principle has to be applied in the proper temporal order, with all events at a particular time being assessed simultaneously, and events worthy of inclusion being used to assess later events.

---

because $A$ prevents two events $E_1$ and $E_2$, where $E_1$ lowered $E$'s probability, and $E_2$ increased $E$'s probability (both by the same amount). In that case, where $A$ affects an event that affects the probability of $E$, it would seem that the independence is not satisfied. A fully adequate definition of independence would have to exclude cases like this. For more on this (in relation to counterfactuals) see Kvart (1986).

[11] As with the previous principle, there are some important technical issues here that are beyond the scope of this chapter. It seems that (for antecedent $A$ and event $E$) $Prob(E/A) > Prob(E/\neg A)$ is not sufficient for $E$ to be a positively affected event. For perhaps $A$ affects two events $E_3$ and $E_4$, where $E_3$ lowers $E$'s probability and $E_4$ increases $E$'s probability (in such a way that $E$ comes out ahead in the final calculation). A complete account of being positively affected would have to exclude cases like this (or explain why they should be included). Again, for more on this, consider Kvart (1986).

Independent events and positively affected events are unique in that they were already known (or projected) to occur between the antecedent scenario and time of the consequent. We simply need to show that (in effect) these projections should not change because of the alterations made to generate the antecedent scenario. However, there are presumably also going to be events that obtain in the scenario that were *not* already known (or projected) to occur. In fact, it would seem that (apart from very short counterfactuals) most of the intermediate events are likely not to be already something we know will occur. But, how do we determine when to include this sort of event? This creates the third task, which is that we should ($P_8$) *include events from the consequents of supporting counterfactuals as intermediate states.* For example, among the events that should be included as intermediate states in determining the five-year outcome of a nuclear Iran will be the outcomes of the four-year nuclear Iran, which will be determined using (as intermediate states) those that are the outcome of the three-year nuclear Iran, and so on. Therefore, once one has included the independent and positively affected events for a particular time period, one should then continue on to the final stage of counterfactual evaluation and determine the outcome of the antecedent scenario and the included events for that time period. That result is then fed back into this stage of counterfactual reasoning in order to select later intermediate states. Here again, it is essential to recognize the progressive nature of proper scenario development. All the events at a particular time should be assessed before moving to a later period in order not to overlook a potential relationship that would require an event to be included (or not).

**Principles for establishing scenarios outcomes**

Throughout this chapter, I have urged that one not consider possible scenario outcomes too early in the counterfactual reasoning process. Since our primary interest in counterfactual reasoning is to determine the scenario outcomes, it is only natural that we will want to begin by thinking about those possibilities. However, we do this effectively only if we have first thoroughly and rigorously selected both antecedent scenarios and intermediate states. After all, determining scenario outcomes is (in the end) a matter of probabilistic extrapolation. But what is the basis for making such extrapolations? One cannot simply use well-known regularities and the minimal statement that defines the counterfactual's antecedent. One must have a fully developed backstory for the antecedent, as well as for the period between that antecedent and the time of interest for the consequent. Then, and only then, it is possible to project an outcome effectively.

Selecting scenario outcomes begins by generating the possible consequences of the antecedent scenario and intermediate states. Just as in generating the antecedent scenarios, it is essential not to exclude any possibility without a clear definitive reason for so doing. Any scenario that is consistent with the antecedent scenario and intermediate states should be considered.

One must be especially careful to avoid only offering outcomes that mirror what has happened in similar prior historical cases. Thus, the first task is to ($P_9$) *maximize the range of possible outcomes.*

Now, if one has rigorously and thoroughly selected one's antecedent scenario and intermediate states, selecting from these possible outcomes should be (relatively) straightforward. One should ($P_{10}$) *weigh the probabilities of the possible outcomes given the antecedent scenario and intermediate states.* Naturally, one will develop some sort of a ranking of the scenario outcomes in terms of their probability. But, it is important to realize that not all aspects of the scenario outcomes will have the same probability. It is possible that all outcomes consistent with the antecedent scenario and intermediate states will have some common elements. Those need to be separated from others that are simply more or less probably part of the scenario outcome.

Once one has selected the final scenario outcome, one forms the final counterfactual. For scenario outcome O, and antecedent A, the final counterfactual will be 'if it were the case that A, then it would be the case that O.' Now, it is important to note, again, that much of what really matters in counterfactual reasoning *is not part of this counterfactual.* That is, the counterfactual makes no explicit statement of the way in which the antecedent was generated. Furthermore, it does not specify what occurs between A and O. However, both of those are essential to justifying the counterfactual. In fact, it would seem that they are (arguably) *the most important factor(s).*

There is one final principle that should be employed in establishing scenario outcomes. One should always be sure to ($P_{11}$) *reevaluate the original counterfactual question.* Consider the final counterfactual. Does it really answer what was at issue in the original question? What would have happened if one had asked a slightly different original question? Look at the way in which the reasoning proceeded in generating the final outcome (through all the stages). Is there reason to think that the outcome would be different if the original question had been slightly different? This is particularly important if the outcome turns out to be exactly what one had expected. If that is the case, then it may be worthwhile to do a further counterfactual analysis with a slightly different antecedent. Obviously, time will determine whether this can be done. If nothing else, a few moments of reflection will bring clarity about what has *and has not* been established. Additionally, a precise qualification of the scope of one's conclusion is often just as important as the conclusion itself.

## Conclusion

This brief discussion of counterfactual reasoning is grounded in a much larger research project: an attempt to integrate different approaches to counterfactuals from analytic philosophy, logic, and the social sciences in order to create both a counterfactual reasoning research tool as well as a counterfactual

reasoning *teaching* tool. As a research tool, these principles help to form a structured method to aid intelligence and security analysts in assessing counterfactuals. And, as a teaching tool, the principles help to form a structured method to educate (future) analysts who are in the process of learning and honing the skill of counterfactual reasoning. Any attempt at understanding applied counterfactual reasoning should serve both of these goals. In order to pursue these objectives, I have offered some basic purposes, problems, parts, and principles of applied counterfactual reasoning. In particular, I urged that the ways in which antecedent scenarios are selected and the ways in which scenarios are developed constitute essential (albeit often neglected) aspects of counterfactual reasoning. Both must be addressed if we are to apply counterfactual reasoning effectively. Naturally, further issues remain, but these should serve as a useful point of departure. They are the beginning of a path to more rigorous and relevant counterfactual reasoning in intelligence analysis and counterterrorism.[12]

## References

Bennett, J. 2003. *A Philosophical Guide to Conditionals*. New York: Oxford University Press.

Cowley, R., ed. 2000. *What If? The World's Foremost Military Historians Imagine What Might Have Been*. New York: Berkley Books.

Cowley, R., ed. 2002. *What If? 2: Eminent Historians Imagine What Might Have Been*. New York: Berkley Books.

Cowley, R., ed. 2004. *What Ifs of American History*. New York: Berkley Books.

Elster, J. 1978. *Logic and Society: Contradictions and Possible Worlds*. UK: Wiley & Sons.

Fearon, J. 1991. Counterfactuals and hypothesis testing in political science. *World Politics* 43:169–195.

Fearon, J. 1996. Causes and counterfactuals in social science. In *Counterfactual Thought Experiments in Global Politics: Logical, Methodological, and Psychological Perspectives* 39–68. Ed. P. E. Tetlock and A. Belkin. Princeton, NJ: Princeton University Press.

Ferguson, N., ed. 1999. *Virtual History: Alternatives and Counterfactuals*. New York: Basic Books.

Fogel, R. W. 1964. *Railroads and Economic Growth*. Baltimore, MD: Johns Hopkins Press.

[12] A previous version of this chapter was presented at the Descartes Conference on Mathematical Methods in Counterterrorism on September 28, 2006. I thank the Center for Advanced Defense Studies for the opportunity to present at this conference, as well as those who participated in it. In addition, the research that underlies this chapter was sponsored by a grant from the Institute for Information and Infrastructure Assurance at James Madison University (as part of their Critical Infrastructure Protection Program). I also wish to thank Ken Newbold, Steve Stewart, John Noftsinger, and Mike Hutton for their comments and/or support regarding this project.

Hawthorn, G. 1991. *Plausible Worlds: Possibility and Understanding in History and the Social Sciences*. New York: Cambridge University Press.

Khong, Y. F. 1996. Confronting Hitler and its consequences. In *Counterfactual Thought Experiments in Global Politics: Logical, Methodological, and Psychological Perspectives* 95–118. Ed. P. E. Tetlock and A. Belkin. Princeton, NJ: Princeton University Press.

Kvart, I. 1986. *A Theory of Counterfactuals*. Indianapolis, IN: Hackett Publishing Company.

Kvart, I. 1992. Counterfactuals. *Erkenntnis* 36:139–179.

Kvart, I. 1994. Counterfactuals: Ambiguities, true premises, and knowledge. *Synthese* 100:133–164.

Lebow, R. N. 2000. What's so different about a counterfactual? *World Politics* 52:550–585.

Lebow, R. N. 2001. Contingency, catalysts, and international systems change. *Political Science Quarterly* 115(4):591–616.

Lebow, R. N., and G. Breslauer. 2004. Leadership and the end of the Cold War. In *Ending the Cold War: Interpretations, Causation and the Study of International Relations* 161–188. Ed. R. K. Herrmann and R. N. Lebow. New York: Palgrave Macmillan.

Lewis, D. 1973. *Counterfactuals*. Malden, MA: Basil Blackwell.

Lewis, D. 1979. Counterfactual dependence and time's arrow. *Nous* 13:455–476.

Lewis, D. 1986. Postscripts to counterfactual dependence and time's arrow. In *Philosophical Papers*, vol. 2, 52–66. Oxford: Oxford University Press.

Stalnaker, R. 1968. A theory of conditionals. In *Studies in Logical Theory*. Ed. N. Rescher. New York: Oxford University Press.

Tetlock, P. E., and A. Belkin. 1996. Counterfactual thought experiments in global politics: Logical, methodological, and psychological perspectives. In *Counterfactual Thought Experiments in Global Politics: Logical, Methodological, and Psychological Perspectives* 3–38. Ed. P. E. Tetlock and A. Belkin. Princeton, NJ: Princeton University Press.

Tsouras, P. G., ed. 2001. *Rising Sun Victorious: The Alternate History of How the Japanese Won the Pacific War*. London, UK: Greenhill Books.

Tsouras, P. G., ed. 2002. *Third Reich Victorious: Alternate Decisions of WWII*. London, UK: Greenhill Books.

# 14

# Adversarial Planning in Networks

Sviatoslav Braynov

**Summary.** In this chapter, we discuss adversarial planning in networks and propose methods for detecting it. We show that the traces attackers leave follow a logical order that can be used to recognize attackers' goals. The chapter studies three methods used by attackers to obfuscate their traces in order to make them difficult for analysis: dummy actions, decoy goals, and buffers. The chapter also presents a model for detecting cells in malicious networks. In order to detect cells, we search for coordination links between attackers and perform link analysis. The link analysis produces a coordination graph that includes all members of a cell. In addition, the chapter proposes a formal metric on coordination graphs that help identify the roles attackers play, and differentiate central from peripheral attackers.

## Introduction

Adversarial plan recognition is important for predicting intentions and future actions of attackers, recognizing unknown attacks, and planning appropriate responses. Because of its ability to detect attacks in the early stages of preparation, adversarial plan recognition plays a constantly increasing role in systems for early detection and prevention of attacks against critical infrastructures.

Adversarial plan recognition was first suggested by Geib and Goldman (2001a) as an addition to the traditional models of keyhole and intended recognition (Cohen et al. 1982). It has also been OK independently proposed by Jensen et al. (2001) for predicting the opponent's moves in robotic games. In adversarial recognition, the observed agent is hostile to the observation of his actions and attempts to thwart the recognition.

Although there has been significant recent work in adversarial plan recognition (Bui et al. 2002; Geib and Goldman 2001a, b, 2002, 2005; Qin and Lee 2004), little thought has been given to the question of how to plan attacks in order to avoid detection. Instead, current research on plan recognition assumes that hostile agents are using planning methods similar to those used by cooperative agents.

S. Argamon, N. Howard (eds.), *Computational Methods for Counterterrorism*, DOI 10.1007/978-3-642-01141-2_14,

In this chapter, we discuss adversarial planning in networks and propose methods for detecting it. The detection methods are based on the observation that attackers' traces follow a logical order that can be used to recognize attackers' goals. For example, an attacker's earlier action usually prepares a later action to be taken by the same or another attacker. By analyzing the logic of the attacker's trace, we can determine whether the attacker is working on his own or he is a member of a larger cell.

It is often the case that attackers try to confuse the plan recognition process by leaving ambiguous or incoherent traces. We identify different methods for obfuscating attackers' traces and show how to detect them.

In the chapter, we also propose a method for detecting cells in malicious networks. The model is based on the observation that members of a malicious cell need to cooperate in order to prepare and launch an attack. Therefore, in order to detect whether certain agents are involved in a malicious coordinated activity, we search for coordination links between them and build a coordination graph that includes all attackers who, in some way or another, cooperate with each other, i.e., the maximal cell of attackers.

The chapter is organized as follows. In the following section, we discuss adversarial planning by a single agent and present a model for detecting it. We then describe a model for detecting malicious cells in networks. The chapter concludes with a brief summary of results.

## Adversarial planning by a single agent

Adversarial planning by a single agent is based on the assumption that the planning agent is self sufficient and does not require the help of other agents to attack the target. In this case, the attacker's objective is to find a plan, a sequence of actions, that leads to the target and could withstand the defender's counteractions. To understand the attacker's plan, we need to consider all of its actions together and not in isolation. Single actions, taken out of context, typically have numerous meanings. When all actions are put in a causal sequence, the specific meaning of any action can be derived from its relationships to other actions.

The basic observation is that actions in the attack plan follow a logical order and form a sequence in which former actions prepare later actions. Each action in the sequence can be described by a set of preconditions and a set of postconditions. The preconditions specify the prerequisites for taking the action. If any of the preconditions are not met, the action either cannot be taken or will fail to produce the intended result. The postconditions specify the results of the action and they are usually used to satisfy the preconditions of a later action. In other words, an attacker's early actions prepare his later actions.

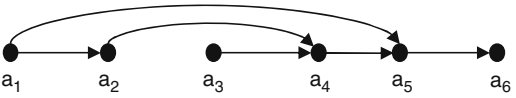



**Fig. 14.1.** Coherent action trace.

**Definition 1 Coherent trace**
The attacker's trace is coherent if the preconditions of every action taken by the attacker are logical consequences from the initial state and the postconditions of all preceding actions.

Formally, the attacker's trace is a sequence of actions:

$$tr = a_0(pre_0, post_0), a_1(pre_1, post_1), ..., a_k(pre_k, post_k)...$$

where action $a_k(pre_k, post_k)$, executed at moment $k$, has preconditions and postconditions $pre_k$ and $post_k$, respectively. The trace is coherent if:

$$\{S_0, post_0, post_1, ..., post_{k-1}\} \models pre_k \tag{14.1}$$

where $S_0$ is the initial state. In other words, the attacker's trace is coherent if the preconditions of the action executed at moment $k$ can be logically derived from the initial state of the system, $S_0$, and the results of all preceding actions. In a sense, coherence could be viewed as self-sufficiency: the attacker cannot rely on somebody else's assistance and needs to prepare his future actions on his own.

Although Eq. (14.1) requires that an action be prepared by all previous actions, it does not specify which previous actions actually prepared the given action. For example, it could be the case that a trace is coherent, yet not all previous actions contributed to the preparation of the given action. The following definition refines the meaning of "preparation."

**Definition 2 Action dependence**
An action $a_n(pre_n, post_n)$ depends on another action $a_k(pre_k, post_k)$ executed earlier, $k < n$, iff:

$$\{S_0, post_0, post_1, ..., post_{k-1}, post_{k+1}, ..., post_{n-1}\} \not\models pre_k$$

In other words, an action $a_n$ depends on a previous action $a_k$ if the removal of $a_k$ from the trace invalidates the preconditions of $a_n$, thereby making it impossible for the attacker to take $a_n$ at moment $n$.

Figure 14.1 shows an attacker's trace where action dependencies are depicted by arrows. For example, action $a_5$ depends on actions $a_4$ and $a_1$; action $a_4$ depends on actions $a_2$ and $a_3$. What is remarkable about this trace is that the last action executed, $a_6$, depends directly or indirectly on all previous actions. In a sense, every action has its role in the sequence of actions and there are no superfluous or unnecessary actions. More formally:

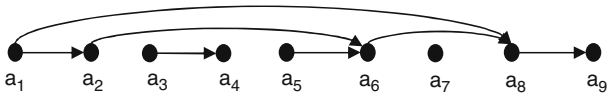



**Fig. 14.2.** Coherent but ambiguous action trace.

**Definition 3 Unambiguous trace**
The attacker's trace is unambiguous iff it is coherent and every action depends directly or indirectly on all previous actions.

It is easy to see that an attacker's trace is unambiguous if and only if every action has an incoming path from every other previous action in the trace. For example, in Fig. 14.1, there is a path from every action to the last action $a_6$, meaning that $a_6$ was prepared directly or indirectly by all previous actions.

Unambiguous traces are relatively easy to analyze because the attacker has a single goal (or target) and all actions directly or indirectly lead to the goal. The more actions the trace has, the closer to the goal the attacker is, and the clearer his intentions are. Obviously, unambiguous traces easily reveal the intentions of the attacker, exposing him to detection and investigative analysis. An attacker will avoid leaving unambiguous traces except in situations where the attack is relatively short and there is no time to analyze the trace.

There are several ways in which an attacker could obfuscate his trace, making it ambiguous, and therefore more difficult for analysis and interpretation. In this section, we study two basic obfuscation techniques: inserting *dummy actions* and inserting *decoy goals*. Both techniques make the trace ambiguous while still keeping it coherent.

Inserting dummy actions is a technique where the attacker repeatedly takes actions irrelevant to his goal in an attempt to make the goal recognition more difficult. Dummy actions increase the complexity of the trace by interfering with other actions and adding superfluous action dependencies. An example of an ambiguous action trace is shown in Fig. 14.2. Here, actions $a_3$, $a_4$, and $a_7$ could be dummy and irrelevant to the mainstream of action. It is easy to see that the trace is coherent because every action is fully prepared by the actions preceding it in the trace.

The following algorithm helps identify dummy actions in an action trace.

A. Divide the original trace into unambiguous subtraces using standard graph-theoretic algorithms. For example, the trace in Fig. 14.2 can be divided into three unambiguous subtraces: $(a_1, a_2, a_5, a_6, a_8, a_9)$, $(a_3, a_4)$, and $(a_7)$.
B. If the trace is complete (the attacker is not expected to take further actions), identify the goal (the target) of the attack and the subtrace that achieves the goal. All other traces constitute dummy actions.
C. If the trace is incomplete (the attack started but has not yet finished), find possible attack targets associated with each subtrace. This requires

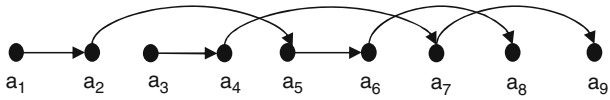



**Fig. 14.3.** Action trace with two goals.

that each subtrace be extended into the future by applying a planning algorithm on the trace (Weld 1994; Boutilier and Brafman 2001; Braynov and Jadiwala 2003). If a subtrace cannot be extended to a complete attack, it could be dummy.

Inserting decoy goals is another technique for obfuscating an attacker's trace. In this case, instead of inserting dummy actions into the trace, the attacker inserts several subtraces, each subtrace pursuing a different goal. In other words, the attacker acts as if he is trying to simultaneously accomplish several goals. Usually, all goals but one are decoy, i.e., they are used for the sole purpose of confusing the goal recognition process. In practice, it could be difficult to tell apart a decoy goal from a real goal because the decoy goal could be a real target that the attacker is trying to reach. The attacker could spend time and resources moving closer to the decoy, making it look as a legitimate goal. The difference between a decoy goal and a real goal is that the attacker does not intend to reach a decoy goal. In most cases, the attacker would start moving towards the decoy goal and later, would abandon it in favor of the real goal. In the extreme case, the decoy goal could be reached simultaneously with the real goal in order to further confuse the defender and waste his response time and resources.

Figure 14.3 shows an attacker's trace with two subtraces: $(a_1, a_2, a_5, a_6, a_8)$ and $(a_3, a_4, a_7, a_9)$. Each subtrace is unambiguous and leads to a different goal. Let $G_1$ and $G_2$ denote the goals, respectively for the first and the second subtrace. Without additional information, it is impossible to differentiate between the following cases:

- The attacker's intention is to achieve both $G_1$ and $G_2$.
- The attacker wants to achieve only $G_1$. $G_2$ is used as a decoy in order to confuse the detection system.
- $G_2$ is intended and $G_1$ is a decoy.
- Both $G_1$ and $G_2$ are decoy goals.

## Adversarial planning in networks

In this section we present a model for detecting cells in malicious networks. The model is based on the observation that members of a malicious cell need to cooperate in order to prepare and launch an attack. Therefore, in order to detect whether certain agents are involved in a malicious coordinated activity, we search for coordination links between them and perform link analysis.

In general, there are two main reasons for attackers to cooperate:

- Cooperation allows attackers to perform actions that are beyond the power or capabilities of a single attacker. That is, cooperation allows a group of attackers acting together to achieve what single attackers cannot achieve by acting alone.
- Even when an attacker is capable of executing his tasks alone, cooperation allows attackers to achieve their goals better, i.e., it could decrease costs, increase concealment, increase speed, etc.

In general, one could identify two types of cooperation:

Cooperation through action correlation in which agents' actions interfere with one another. A particular type of interference occurs when an agent can perform an action that enables a future action to be performed by another agent who is not willing or not able to perform it.

Cooperation through task correlation in which agents' actions do not interfere with one another. Instead, cooperation is achieved by dividing a large task into a set of independent subtasks:

$$T_1 \wedge T_2 \wedge ... \wedge T_k \rightarrow T$$

In this case, the execution of tasks $T_1, T_2, ..., T_k$ implies task $T$. Apparently, each subtask can be assigned to a different attacker who can execute it independently of other attackers. Because tasks are independent, agents' actions do not interfere or correlate with one another.

In this chapter, we study how to detect action correlation. The problem of detecting task correlation is very difficult and there is no indication that a general solution exists. First, the problem has a non-trivial complexity. For example, for a given set of tasks one needs to check whether any subset of tasks correlate. Determining whether a set of tasks correlate requires complex logical inference that is often computationally untractable. Second, cooperation through task correlation requires information about agents' intentions that are usually difficult to quantify, identify, and detect. Only the attacker knows what his real intentions are, and he can easily manipulate the intention recognition process.

In cooperation through action correlation, agents' actions interfere with one another. By interference we mean the fact that an action can affect the outcome of another action. In other words, the intended effect of a single action may depend on other action(s) taken previously or concurrently with the given action. An action performed by a single agent or a group of agents could modify the outcome of an action performed by another group, thereby invalidating the outcome, or improving it.

In general, we have two types of interference: positive and negative. Positive interference occurs when one action enables another action, or improves its results. Negative interference takes place when an action invalidates the

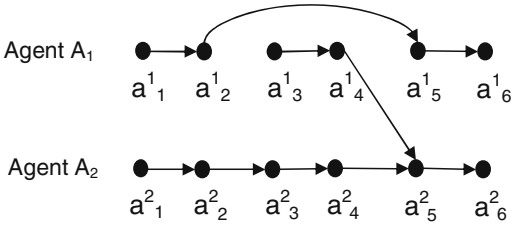



**Fig. 14.4.** Dependence between action traces.

result of another action or merely disables it. Two dimensions of interference can be identified: interference between the actions of the same agent, and interference between actions of different agents.

We view cooperation between attackers as an interference of their actions. The main objective of cooperation is to avoid negative interference and take advantage of positive interference between attacker's actions. This observation serves as a starting point for cooperation detection. To detect cooperation between attackers we look for patterns of assistance.

**Definition 4 Agent dependence**
An agent $A_2$ depends on another agent $A_1$ if the action trace of agent $A_2$ has an action that depends on an earlier action performed by agent $A_1$. In other words, agent $A_1$ performs an action that brings about a precondition required by a subsequent action performed by agent $A_2$.

According to this definition, Agent $A_1$ prepares an action taken later by agent $A_2$. Therefore, there is at least one action in $A_2$'s trace that is not prepared by himself. As a result, the action trace of $A_2$ must be incoherent. To determine whether an action trace is incoherent one needs to check whether Eq. (14.1) holds for every action of the trace. This can be done efficiently using standard theorem proving algorithms.

An example of action dependence is shown in Fig. 14.4 where agent $A_2$ depends on agent $A_1$ for executing action $a_4^1$. In this case, agent $A_2$ needs to execute action $a_5^2$ that must be prepared by actions $a_4^1$ and $a_4^2$. Agent $A_2$ does half of the preparation by executing $a_4^2$ and leaving the execution of $a_4^1$ to agent $A_1$.

It has to be pointed out that the dependence between the action traces of agents $A_1$ and $A_2$ is not always strong evidence of cooperation between them. For example, it could be a coincidence that agent $A_1$, by executing his action, unintentionally and accidentally assisted agent $A_2$. To rule out such cases we look for two types of additional evidence:

- The agent being assisted is not able to assist himself. That is, if an action taken by agent $A_2$ requires a precondition satisfied earlier by an action taken by agent $A_1$, then agent $A_2$ is not capable, not allowed, or not

willing to satisfy the precondition by himself. To find such evidence we need to keep a profile of every agent that specifies the agent's abilities, resources, and constraints. For example, in Fig. 14.4, agent $A_2$ must not be able to execute action $a_4^1$. In another scenario, executing the action could be illegal for agent $A_2$ or he might not be authorized to execute it. Obviously, agent $A_2$ needs the assistance of another agent (agent $A_1$ in this case) who is capable or authorized to execute the action. Finally, agent $A_2$ could be able and authorized, yet not willing to execute $a_4^1$ for reasons of avoiding being detected and investigated. In this case, agent $A_2$ might prefer to delegate the action to agent $A_1$.

- The action that agent $A_1$ takes to assist agent $A_2$ is irrelevant to agent $A_1$. If the action is irrelevant to agent $A_1$, then, obviously, the sole purpose of taking it is to assist agent $A_2$. An action would be irrelevant to $A_1$ if it does not prepare any future actions of the same agent. That is, there is no subsequent action taken by agent $A_1$ that depends on the given action. Therefore, we can remove the given action from agent $A_1$'s trace, without violating its coherence. In Fig. 14.4, agent $A_1$'s trace is ambiguous and the subtrace $(a_3^1, a_4^1)$ does not contribute to further actions. The reason is that actions $a_3^1$ and $a_4^1$ are irrelevant and they have been taken solely to assist agent $A_2$. We have to admit, however, that a shrewd attacker can manipulate this type of evidence by pretending to use a irrelevant action in its future activity, thereby making it seem relevant. For example, in Fig. 14.4, agent $A_1$ might extend subtrace $(a_3^1, a_4^1)$ to lead to a dummy goal.

To detect a malicious network we build a coordination graph. The nodes in the graph represent agents, and the arcs represent dependence. There is an arc from agent $A_1$ to agent $A_2$ iff agent $A_2$ depends on agent $A_1$. The arcs are labeled with frequencies showing how many different dependencies were detected between two agents. Once the coordination graph is built, it can be used to detect malicious coordination. Depending on the application domain and the accuracy of detection, the system analyst may set a threshold to eliminate those arcs whose frequency is below threshold. The coordination graph can quickly be searched for cells of coordinating agents. A cell is defined as a connected component (a subgraph in which there is a path between every pair of users). Maximal cells can be found, for example, using a depth-first search. In searching for cells, we treat the coordination graph as an undirected graph. The intuition is that it does not matter who assists whom. Once a coordination link between two agents is discovered, both agents belong to the same cell. An example of a coordination graph is represented in Fig. 14.5. It shows the coordination links between five agents belonging to the same cell.

The set of the *actual attack executers* consists of all agents that directly harm, break, block, or destroy a target. The attack executers achieve the final objective of a malicious cell by taking actions directly on the tar-

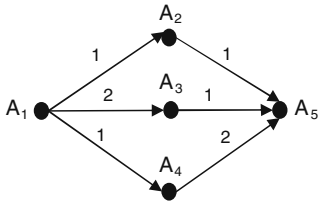



**Fig. 14.5.** An example of coordination graph.

get. In Fig. 14.5, the attack executer is agent $A_5$. He does not assist any other agents, although he is assisted directly or indirectly by all agents from the cell.

One problem with detecting malicious cells is that they might include many other agents in addition to the actual attack executers. There are often agents who organize, prepare, and make an attack possible, without taking active part in it. We call such agents *shadow agents*, because they usually stay in the shadow of the attack by performing perfectly legitimate actions that often are difficult to detect. Shadow agents present a real threat because if they remain unrecognized after the attack, they can prepare and launch new attacks.

Shadow agents are those members of a malicious cell who are located away from the main stream of attack, i.e., away from the attack executers. For example, in Fig. 14.5, agent $A_1$ is a shadow agent.

**Definition 5 Shadow agents**
The set of shadow agents consists of all agents $\boldsymbol{s}$ whose distance, $\boldsymbol{d}$, from the actual attack executers, $\boldsymbol{A}$, exceeds a certain domain dependent threshold $d_0$:

$$\min_{a \in A} d(s, a) > d_0$$

where the minimum is taken over the set of the actual attack executers.

Shadow agents are the exact opposite of the attack executers who are directly exposed to detection and prosecution. An agent could be a shadow agent for several reasons. One reason is to avoid detection. If an attack is detected and investigated, staying away from the main attack stream offers protection by making it difficult to unravel the conspiracy. Second, an agent may not be able to contribute to the attack preparation and execution due to limited resources, knowledge, and capabilities. In this case, staying in the "shadow" is not a deliberate choice, but the result of one's own limited ability to participate actively in the attack.

In order to become a shadow agent, an attacker can deploy a series of "buffers" in which a single task is intentionally and artificially divided into smaller subtasks assigned to different attackers. The "buffers" decrease the amount of direct involvement in an attack by replacing a direct coordination link with a chain of indirect links.

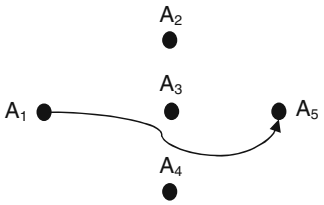



**Fig. 14.6.** Direct coordination link without using buffers.

Figure 14.6 shows a direct coordination link without using buffers. In this case, agent $A_1$ directly assists the actual attack executer, agent $A_5$. In other words, agent $A_1$ takes action(s) that prepare future actions of agent $A_5$. Such an arrangement, however, does not provide enough protection for agent $A_1$ if he wants to stay away from the attack executer. Agent $A_1$ could decrease his direct involvement in the attack by using intermediate agents as "buffers" between him and agent $A_5$. Figure 14.5 shows a different arrangement of the situation that uses three "buffers": agents $A_2$, $A_3$, and $A_4$. Here, instead of assisting directly agent $A_5$, agent $A_1$ assists agents $A_2$, $A_3$, and $A_4$ who in turn assist the attack executer, agent $A_5$.

We also apply graph-theoretic measures to analyze the structure of a malicious cell, and to identify the *central attackers* and the *peripheral attackers*. The concept of centrality reflects an attacker's position in the coordination graph and his involvement in the coordination. A central attacker might be or might not be a shadow agent depending on how deeply he wants to conceal his activity.

In our model, we use the following centrality measure:

$$\text{Centrality}(A_i) = \sum_{k \in N} (link(A_i, A_k) + \text{Centrality}(A_k)) \qquad (14.2)$$

where $N$ is the set of all agents assisted by agent $A_i$, $link(A_i, A_k)$ is the frequency of how many times agent $A_i$ assisted agent $A_k$. According to Eq. (14.2), the centrality of an attacker depends on how many times the attacker assisted other attackers. In addition, Eq. (14.2) takes into account the centrality of the attackers who were assisted. An attacker who assists central attackers is also a central attacker. For example, in the coordination graph in Fig. 14.5, agent $A_1$ has centrality 8, and plays the most central coordination role in the cell.

It has to be pointed out that the notion of centrality is complimentary to the notion of the actual attack executers. An attacker may be a central player who does not take part in the actual attack execution, as is the case in the coordination graph in Fig. 14.5. In this case, agent $A_1$ is both a central attacker and a shadow agent. Effective detection and response require quick localization and neutralization of both immediate executers and central attackers. In our example, the first two agents that have to be neutralized are agent $A_1$ and

agent $A_5$. To discover agent $A_1$, however, a coordination graph needs to be built and analyzed.

## Conclusion

We have studied the problem of detecting malicious cells in networks. We showed that attackers' traces follow a certain logical order in which earlier actions prepare later actions. We also demonstrated how the structure of attackers' traces could help identify attackers' goals and differentiate between real and decoy goals.

This chapter also presented a model for detecting cells in malicious networks. In order to detect cells, we perform link analysis on attackers' traces and search for coordination links. The output of the link analysis is a coordination graph that identifies all members of a malicious cell, including not only the attack executers but also their assistants.

We also analyze the structure of a malicious cell, and proposes a graph-theoretic metric that helps identify central, peripheral, and shadow attackers.

The methods presented in this chapter are generic and domain independent which makes them applicable to a wide range of applications, including antiterrorist research and computer security.

## References

Boutilier, C., and R. Brafman. 2001. Partial-order planning with concurrent interacting actions. *Journal of Artificial Intelligence Research* 14:105–136.

Braynov, S., and M. Jadiwala. 2003. Representation and analysis of coordinated attacks. In *FMSE '03: Proceedings of the 2003 ACM workshop on Formal methods in security engineering* 43–51. ACM Press.

Bui, H., S. Venkatesh, and G. West. 2002. Policy recognition in the abstract hidden Markov models. *Journal of Artificial Intelligence Research* 17:451–499.

Cohen, P., J. Allen, and C. Perrault. 1982. *Beyond Question Answering. Strategies for Natural Language Processing.* Lawrence Erlbaum Associates.

Geib, C., and R. Goldman. 2001a. Plan recognition in intrusion detection systems. In *DARPA Information Survivability Conference and Exposition (DISCEX II'01)* 46–55.

Geib, C., and R. Goldman. 2001b. Probabilistic plan recognition for hostile agents. In *Proceedings of the FLAIRS 2001 Conference.*

Geib, C., and R. Goldman. 2002. Requirements for plan recognition in network security systems. In *Fifth International Symposium on Recent Advances in Intrusion Detection (RAID 2002).*

Geib, C., and R. Goldman. 2005. Observability and probabilistic plan/goal recognition. In *IJCAI-05 Workshop on Modeling Others from Observations.*

Jensen, R., M. Bowling, and M. Veloso. 2001. Obdd-based optimistic and strong cyclic adversarial planning. In *Proceedings of ECP.*

Qin, X., and W. Lee. 2004. Attack plan recognition and prediction using causal networks. In *Proceedings of The 20th Annual Computer Security Applications Conference (ACSAC 2004)*.

Weld, D. S. 1994. An introduction to least commitment planning. *AI Magazine* 15(4):27–61.

# 15

# Gaming and Simulating Ethno-Political Conflicts

Barry G. Silverman, Gnana K. Bharathy, and Benjamin D. Nye

**Summary.** This chapter begins by describing a universally recurring socio-cultural "game" of inter-group competition for control of resources. It next describes efforts to author software agents able to play the game as real humans would – which suggests the ability to study alternative ways to influence them, observe PMESII effects, and potentially understand how best to alter the outcomes of potential conflict situations. These agents are unscripted, but use their decision making to react to events as they unfold and to plan out responses. For each agent, a software called PMFserv operates its perception and runs its physiology and personality/value system to determine fatigue and hunger, injuries and related stressors, grievances, tension buildup, impact of rumors and speech acts, emotions, and various collective and individual action decisions. The chapter wraps up with a correspondence test from a SE Asian ethnic conflict, the results of which indicate significant correlation between real and agent-based outcomes.

## Introduction and purpose

Gaming and simulation of socio-cultural groups is a newly evolving field, motivated by the need to better understand how leaders and followers behave, what motivates them, how dangerous ideas spread, and how they might be influenced to cooperate, mitigate conflicts, and benefit the overall good (Pew and Mavor 1988). Green (2002) studies the array of methods for forecasting conflict and show that predictions are significantly improved when subjects first participate in role playing games about the issues at stake. Hence, *one aim* of this research is to isolate the components needed for a generic role playing game to be used to rapidly mock up a class of conflicts commonly encountered in today's world. In other words, create a widely applicable game generator. Since it is often impossible to find humans to play all the roles of such games, or to play out all the possible scenarios, a *second aim* is to create plausible models of leaders and followers based first principles about what makes them tick and so they may play some of the roles in the game. If these cognitive agents are realistic, they can help trainees and analysts explore the

S. Argamon, N. Howard (eds.), *Computational Methods for Counterterrorism*, DOI 10.1007/978-3-642-01141-2_15,

range of their possible actions under varieties of conditions, thereby helping others to see more clearly how to influence them and elicit their cooperation.

It is a human tendency to project our own value systems upon others and presume they want the same things we want (the mirror bias). Once we form such hypotheses, we tend to look only for confirming evidence and ignore disconfirming facts (the confirmation bias). Heuer (1999) points out that it is vital to break through these and related biases, and that methodical approaches such as realistic simulations, if well done, might help to elucidate and explore alternative competing hypotheses of agents' motivations, intentions and consequent behavior. Thus generation and testing of new hypotheses is a *third aim*, and another potential benefit of simulations.

## Socio-cultural game theory

How can an analyst or trainee devise policies that will influence groups for the collective good? And what must a socio-cultural game generator encompass? Fig. 15.1, explained below, attempts to portray a fairly universal class of leader-follower game that groups often find themselves in and that are worthy of simulation studies. This could be for competing groups in a crowd, in an organization, in a region or nation, or even between nations. Analysts would need an appropriate suite of editors and a generator, to help them rapidly mock up such conflict scenarios and analyze what outcomes arise from different courses of action/policies. We describe this game intuitively here and more formally in Appendix I.

Specifically, the socio-cultural game centers on agents who belong to one or more groups and their affinities to the norms, sacred values, and interrelational practices (e.g., language, gestures, social rituals) of those groups. Specifically, let us suppose there are N groups in the region of interest, where each group has a leader archetype and two follower archetypes (loyalists & fringe members). We will say more about archetypes shortly, and there can certainly be multiple leaders and followers, but we stick in this discussion to the smallest subset that still allows one to consider beliefs and affinities of members and their migration to more or less radical positions. There is an editable list of norms/value systems from which each group's identity is drawn. The range across the base of Fig. 15.1 shows an example of a political spectrum for such a list, but these could just as easily be different parties in a common political system, diverse clans of a tribe, different groups at a crowd event, and so on. Each entry on this list contains a set of properties and conditions that define the group, its practices, and entry/egress stipulations. The authority of the leader in each group is also indicated by a similarly edited list depicted illustratively across the top of Fig. 15.1.

The vast majority of conflicts throughout history ultimately center around the control of resources available to a group and its members. Before delving into our model, we invite the reader to take a look at a stylized resource control game through game theoretic framework, as set up in the

Appendix 1. The inspiration for this stylized game comes from Wood's (2003) civil war settlements paper. This game theoretic exercise not only gives you a 5,000 ft view, but also illustrates why and how deep models such as ours will be useful. While a number of assumptions made by the game theoretic frameworks are defensible (well-ordered preferences, transitivity), others are meant for mathematical elegance. Without assumptions doing most of the "heavy lifting," it is impossible develop mathematically tractable models (De Marchi, 2005).

Many of these stylized game models are unable to encode domain information, particularly the depth of the social system. For example, human value systems are almost always assumed, hidden, or at the best, shrunk for the purpose of mathematical elegance. Yet, human behavior is vital to the conflict-cooperative game behavior.

While mathematical convenience is one explanation, there is more involved. Many modeling platforms would simply not allow value systems to be made explicit, and there is no modeling process that would allow one to revisit the values. As computational power increases to accommodate more complex models, social system modelers are beginning to address this curse of simplicity.

Even though such models can not be solved mathematically, we can find solutions through validated simulation models with deep agents. If one could find clusters of parameters that pertain to a corresponding game model, we can also start talking about correspondence between game theoretic models and cognitively deep simulation models. There is room for a lot of synergy.

Now, let us return to the cognitively detailed game. The resources of each group are illustrated along the left side of Fig. 15.1 and are summarized for brevity into three tanks that serve as barometers of the health of that aspect of the group's assets – (1) political goods available to the members (jobs, money, foodstuffs, training, healthcare etc.); (2) rule of law applied in the group as well as level and type of security available to impose will on other groups; and (3) popularity and support for the leadership as voted by its members. Querying a tank in a culture game will return current tank level and the history of transactions or flows of resources (in/out), who committed that transaction, when, and why (purpose of transactional event).

To start a game, there are initial alignments coded manually, though these will evolve dynamically as play unfolds. Specifically, each group leader, in turn, examines the group alignments and notices Loyal Ingroup (A), Resistant Outgroup (C), and those "undecideds" in middle (B) who might be turned into allies. Also, if there are other groups, they are examined to determine how they might be enlisted to help influence or defend against the out-group and whatever alliance it may have formed. Followers' actions are to support their leader's choices or to migrate toward another group they believe better serves their personal value system. Actions available to Leader of A are listed in the table on the right side of Fig. 15.1 as either speech acts (spin/motivate, threaten, form pact, brag) or more physical/political acts. Of

the latter, there are 6 categories of strategic actions. The middle two tend to be used most heavily by stable, peaceful groups for internal growth and development. The upper two are economic and militaristic enterprises and campaigns taken against other groups, while the lower two categories of actions are defensive ones intended to barricade, block, stymie the inroads of would be attackers. The right hand column of the action table lists examples of specific actions under each of these categories – the exact list will shift depending on whether the game is for a population, organizational, or small group scenario. In any case, these actions require the spending of resources in the tanks, with proceeds going to fill other tanks. Thus the culture game is also a resource allocation problem. Leaders who choose successful policies will remain in power, provide benefits for their followers, and ward off attackers. Analysts and trainees interacting with this game will have similar constraints to their policies and action choices.The lead author spent much of 2004 assembling a paper-based version of Fig. 15.1 as a role playing diplomacy game and play-testing it with analysts: Silverman et al. (2005). The goal of the game is to help players to experience what the actual leaders are going through, and thereby to broaden and deepen their understanding, help with idea generation, and sensitize them to nuances of influencing leaders in a given scenario. The mechanics of the game place the player at the center of the action and play involves setting objectives, figuring out campaigns, forming alliances when convenient, backstabbing when necessary. This is in the genre

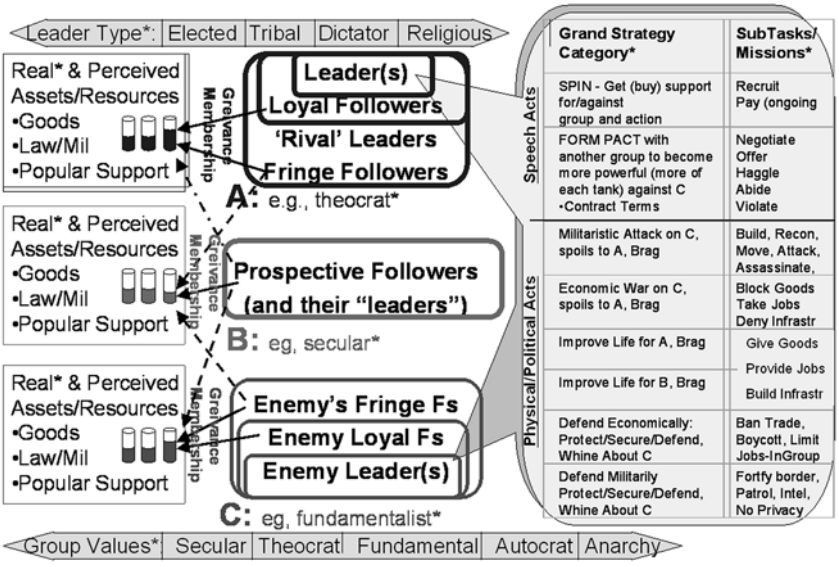



**Fig. 15.1.** Overview of the basic leader-follower game within cultureSim (* - editable list).

of the Diplomacy or Risk board games, though unlike Diplomacy, it is rapidly reconfigurable to any world conflict scenario.

After completing the mechanics and play-testing, three implementations of the game were created: (1) a software prototype called LeaderSim (or Lsim) that keeps world scenarios and action sets to the simplest possible so that we can easily build and test all of the core ideas of the theory; (2) a scaled up version called Athena's Prism that has been delivered as a fully functioning computer game in mid 2005, though AI opponent features are continually being added; and (3) a streamlined version of the paper-based game has been turned into a boardgame called BigWig© aimed at being played to conclusion within an hour (it is thus intended to serve as an intro to the diplomatic strategy genre for new players).

In general, when humans play the game, they rapidly evolve a portfolio of strategies that they tend to pursue asynchronously and in parallel, where a strategy is a high level goal that might be implemented by any of a number of alternative actions. An "action" is defined as a sequence of low level moves governed by the rules of the game. There are only a few moves (e.g., tap/untap tokens, re-assign tokens to resources, etc.). This portfolio or strategy-action-move hierarchies tend to reflect the culture and personality of the leader in a scenario as they attempt to navigate the "game" against the other players.

For the AI to be able to replace a human player and to assemble and manage a portfolio in a way as to reasonably emulate a world leader, a number of components are required in the mind of the agent as shown as the next few subsections amplify. In particular, Performance Moderator Function Server (PMFserv) is a human behavior modeling framework that manages an agent's perceptions, stress and coping style, personality and culture, social relationships, and emotional reactions and affective reasoning about the world: Silverman et al. (2002a,b, 2006a,b; Silverman 2005).

## Agent personality, emotions, culture, and reactions

In LeaderSim, each leader is modeled within a framework known as PMFserv (Silverman, 2005) where the leader's cultural values and personality traits represented through a Goals, Standards and Preferences (GSP) tree. These are multi-attribute value structures where each tree node is weighted with Bayesian importance weights. A Preference Tree is one's long term desires for world situations and relations (e.g., no weapons of mass destruction, stop global warming, etc.) that may or may not be achieved in the scope of a scenario. In Lsim agents this translates into a weighted hierarchy of territories and constituencies (e.g., no tokens of leader X in resource Y of territory Z). When faced with complex decision spaces, different individuals will pursue different long-term strategies which, mathematically, would be very difficult to compare objectively. Chess players, athletes, and scientists develop their own styles for solving the types of problems they encounter. We make use of

the *preference* structure of an agent to account for much of this. For example, one can say that a particular chess player *likes* or is comfortable with certain configurations of the pieces on the board. This allows for the expression of long-term strategic choices that are simply a question of style or preference as to how the world should be.

The Standards Tree defines the methods a leader is willing to take to attain his/her preferences. As described in the authors' other writings, the Standard tree implements a personality profiling tool that is mostly Hermann traits governing personal and cultural norms (Hermann, 1999), plus the additions of protocol vs. substance, and top level guidelines related to Economic and Military Doctrine. Also, we add two standards from the GLOBE study (House et al., 2004) on scope of doing and sensitivity to life (humanitarianism). Personal, cultural, and social conventions render inappropriate the purely Machiavellian action choices (one shouldn't destroy a weak ally simply because they are currently useless). It is within these sets of guidelines where many of the pitfalls associated with shortsighted AI can be sidestepped. Standards (and preferences) allow for the expression of strategic mindsets. When a mother tells her son that he shouldn't hit people, he may not see the immediate tactical payoff of obeying. However, this bit of maternal wisdom exists and has been passed down as a standard for behavior precisely because it is a nonintuitive strategic choice whose payoff tends to derive from what *doesn't* happen far into the future as a result. Thus, our framework allows our agents to be saved from their shortsighted instincts in much the same way as humans often are.

Finally, the Goal Tree covers short-term needs and motivations that implement progress toward preferences. In the Machiavellian and Hermann-profiled world of leaders, the goal tree reduces to a duality of growing vs. protecting the resources in one's constituency. Expressing goals in terms of power and vulnerability provide a high-fidelity means of evaluating the short-term consequences of actions. To this, Athena also adds 3 options for managing reputation (switch from none, to mirroring, to bounded rational) instead of just mirroring in Lsim.

With GSP Trees thus structured, we believe it is possible to Bayesian weight them so that they will reflect the portfolio and strategy choices that a given leader will tend to find attractive, a topic we return to below. As a precursor to that demonstration and to further illustrate how GSP trees represent the modified Hermann profiles, consider the right side of Fig. 15.2. There we see the weighted GSP tree of a leader of a SE Asian nation (name withheld at request of our sponsor) who will be called BlueLeader. Other papers discuss how the weights may be derived so as to increase credibility: e.g., see Bharathy (2006), Silverman et al (2002a,b, 2006b). Here it is more pertinent to discuss how the G-tree implements the Hermann power vs. protect trait. Beneath each subnode that has a + sign, there are further subnodes, but under the G-tree (and P-tree) these are just long sets of constituency resources with importance valuated weights and hence they aren't show here.

The standards or S-tree holds most of the other Hermann traits. Likewise, there are subnodes for the intersection of In Group Bias vs. Degree of Distrust. Openness, as mentioned earlier, is a direct replacement for two other traits, while task vs. relationship focus is also supported. The modifications to Hermann show up as the protocol vs. substance subnodes and the key resource specific doctrines of importance to that leader. In BlueLeader's case, he leans heavily toward power and growth which is also consistent with his P-tree weights on his own resources. His standards reveal him to be dishonest, narrow in scope (self-interested), and task-oriented. While the figure does not expand the lower level detail of all nodes, he also is insensitive to life and outgroups (he did order the harsh treatment which lead to the slaughter of unarmed villagers).

Just to the left of the weight value on each node of the GSP trees of Fig. 15.2 are two "reservoirs" that reflect the current activation of success and failure of this node, respectively. These reservoirs are activated and filled by events and states of the game world as observed by the agent. Figure 15.2 shows early in a game where BlueLeader, a Buddhist, has a lot of conflicted emotions about the outgroup of Muslim villagers in the Southern provinces and some of their demonstrations. In general, we propose that any of a number of $k$ diverse activations could arise with intensity, $\xi$, and that this intensity would be somehow correlated to importance of one's GSP values or node set (GSP) and whether those concerns succeed or fail for the state in question. We express this as

$$\xi_k(b \in B) = \sum_{j \in J} \sum_{v \in V} [W_{ij}(v \in V) * \phi(r_j) * \zeta(v) * \Psi] \tag{15.1}$$

Where,
$\xi_k \rightarrow \xi_k(b \in B)$ = Intensity of activation, k, due to the b'th state of the world.
$J_k$ = The set of all agents and objects relevant to k. J1 is the set consisting only of the self, and J2 is the set consisting of everyone but the self, and J is the union of J1 and J2.
$W(v \in V)$ = Weighted importance of value set V to the agent.
$V$ = The set of goals, standards, and preferences held by the agent.
$\phi(r_j)$ = A function that captures the strength of positive and negative relationships one has with agent or object j that are effected or spared in state b.
$\zeta(v)$ = degree of activation for a goal, standard, or preference.
$\Psi$ = A function that captures temporal factors of the state and how to discount (decay) and merge one's GSP activations from the past (history vector), in the present, and for the future.
It is important to note that the weights adhere to principles of probability; e.g., all child node insights add to unity beneath a given parent, activations and weights are multiplied up a branch, and no child has multiple parents (independence). Although we use fixed weights on the GSP trees, the reservoirs serve to render them dynamic and adaptive to the agent's current needs. Thus,

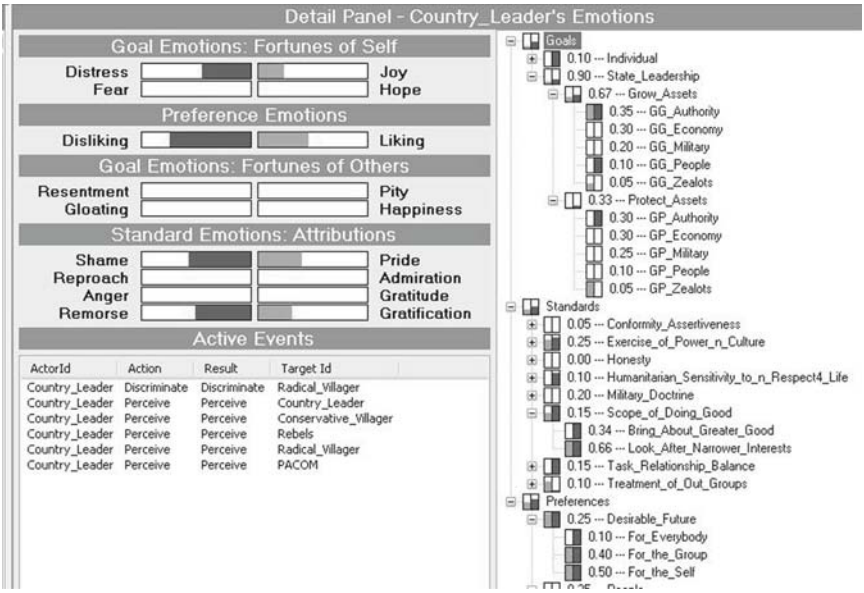



**Fig. 15.2.** GSP tree structure, weights and emotional activations for BlueLeader.

when a given success reservoir is filled, that tends to nullify the importance of the weight on that node (or amplify it if the failure reservoir is filled). In this fashion, one can think of a form of spreading activation (and deactivation) across the GSP structure as the scenario proceeds.

According to other best-of-breed models (Damasio, 1994; Ortony et al., 1988, etc), our emotions are arousals on a set of values (modeled as trees) activated by situational stimuli as well as any internally-recalled stimuli – e.g., see full descriptions of these models in Silverman et al. (2002a,b, 2006a,b). These stimuli and their effects act as releasers of alternative emotional construals and intensity levels, and they assist the agent in recognizing problems, potential decisions, and actions. According to the theory, the activations may variously be thought of as emotions or subjective (moralistic) utility values, the difference being a matter of semantic labeling. Within such a framework, simply by authoring alternative value trees, one should be able to capture the behavior of alternative "types" of people and organizations and predict how differently they might assess the same events, actions, and artifacts in the world around them.

### Agent decision making

What is missing in the previous section is how an agent notices the game world, moves of others, and sense of situation. This discussion will illustrate

how this happens using one Hermann factor (power and vulnerability) as an example. Hopefully, it is fairly straightforward for the reader to extend that to how the other factors are also deployed. Full details exist in Johns (2006).

Central to a given leader's G-Tree reasoning is its perceptions of who threatens it and/or whom it's vulnerable to. Likewise a given leader may be equally interested to estimate who can it influence to best increase its resource assets and thereby its power in the world. Obviously, GSP tree weights will govern how aggressively a given leader pursues each of these vulnerability vs. power concerns, however, we assume that all leader agents need to be able to compute how vulnerable and/or powerful they are at each turn of a game. Since the game rules define precisely which resources can be used to take hostile actions against which other resources, one can derive a measure of a player's vulnerability directly from the state of the game world and the rule set. Intuitively, by factoring *vulnerability* into the world utility calculation, an agent can avoid world configurations in which another is poised to conduct a devastating attack. Adding border defenses, stocking up on supplies, and pulling money out of the economy can all be viewed as behaviors motivated primarily by vulnerability management.

The vulnerability formula ($\beta$) works by generating the percentage of a given player's tokens that can be expected to be lost to a given player in the coming round of attack actions ($a_i$). For each hostile action ($a_i \in A$) that can be initiated by another player ($g$), the number of tokens available to attack and defend is tallied. From this the probability of victory is determined, and then multiplied by the percentage of tokens vulnerable to this attack versus the total number owned by the vulnerable player in each resource category. This is the expected percentage of tokens to be lost if this attack occurs in the next round. The maximum over all attacks, then, gives this player $l$'s vulnerability score $\beta$ to player y.

$$\beta_{xy} = \max_{a \in A} \{Pr(a) * \sigma(x, a)/C(x)\}] \tag{15.2}$$

Agents who purely manage vulnerability, while interesting in their behavior, are not entirely realistic. Human players tend to balance vulnerability against its inverse, *power*. Where vulnerability measures the expected number of tokens a player can lose to other players in the coming round, power measures the expected number of tokens a player can take from others. The calculation of the power heuristic is exactly the opposite as for vulnerability. Player A's vulnerability to Player B is the same as Player B's power over Player A.

Taking the leader's perceived difference between power and vulnerability provides a surrogate for the leader's overall sense of utility of the current state of the world, G, when divorced from his value system and other factors:

$$U_l(G_x) = \alpha x - \beta x \tag{15.3}$$

Recall, however, that a given leader agent (1) tracks who is aligned with whom, tallying things like trust, (2) monitors all resource levels and who used

what actions upon them, and (3) its own actions to achieve its long term preferences or P-tree, as modulated by its standards. Thus $\alpha$ and $\beta$ serve primarily as activations on the leaf nodes of some of the GSP tree branches. PMFserv uses a wide assortment of similar activation mechanics for other factors and computes the Expected Utility (EU) of the world and of new action possibilities when projecting next steps. That is, PMFserv serves as the point where diverse GSP personality and cultural value sets, stressors, coping style, memories, and perceptions are all integrated into a decision for action (or inaction) to transition to a new state (or remain in the same state) and to determine the portfolio of strategies-moves-actions that best maximize that agent's GSP Tree values as follows.

$$\max_{a \in A}[EU(a)] = \sum_{b \in B_a} U_l(b) * pr(b) \tag{15.4}$$

Where,
$A$ = action set available after GSP and stress-constrained perception
$a \in A$
$pr(b)$ = probability of action $a$ leading to state $b$
$U_l(b) = \sum_{k \in K} \zeta_k(b)/11$
Utilities for next actions, $a_k$, are derived from the activations on the GSP trees in the usual manner as in Silverman et al. (2002a,b) and as Silverman et al. (2002a,b) and as briefly summarized for power and vulnerability here. That is, utility is the simple summation of all positive and negative activations for an action leading to a state. Since there will be 11 pairs of oppositely valenced activations in PMFserv's emotion model, we normalize the sum as follows so that utility varies between $-1$ and $+1$.

## Modeling follower value systems

We introduce three refinements in order to also be able to model the values and motivations of followers - (1) additions to the GSP trees, (2) a group-affinity profiling instrument, and (3) group transfer dynamics (exit, voice, and loyalty). In keeping with Aim 2, each of these refinements is an implementation of a well-respected model drawn from the social sciences. Details are omitted, but may be found in Silverman, Bharathy et al. (2006c).

Mathematically, the reader may recall $\phi(r_{ij})$ from earlier Eq. (15.1). Here we examine the case where j is a group (or leader) and the term refers to the membership, relationship, or strength of affinity of agent i to group j. An agent i can belong to multiple groups at varying strength according to:

$$\Phi(r_{iA}) = Superiority_A \times GSPcongruence/VID_{Ai} \tag{15.5}$$

where Superiority and VID are from DI instruments if available, else derived by GSP trees of agent i in reacting to leader or group A.

Groups are characterized by GSP weights for the average of all members as well as by property lists defined a priori (religion, political system, etc.), and GroupPorosity factors. GSP congruence is estimated using the sum of the means square differences in the GSP nodes. $GSPcongruence = 1 - \sqrt{\sum(w_{i1} - w_{i2})^2}$, which is the correlation of the weights between two GSP trees. If an agent is in Group B, it will not be drawn to a Group C whose GSP archetype is substantially incongruent to its own. If an agent is in a group (or under control of a leader) whose average GSP is greatly different from its own, the agents tend to use Voice to resist the leader or attempt to Exit to another group, depending on porosity.

As the leaders did with Fig. 15.1, the followers similarly take each set of opposing groups and place them along a scale as shown below. The decision that the villagers make is expressed as grievance, where the grievance is in the scale of $-4$ to $+4$ are given below (also shown are the Grievance State IDs of the simulation above):

| ⇐ | | | | Villager Decision | | | | ⇒ |
|---|---|---|---|---|---|---|---|---|
| Sacrifice, Go on Attacks | Support, Vote for GroupA | Join Authority Group A | Agree | Neutral (undecideds in Group B) | Disagree, Vote against A | Join, Opposition Group C | Oppose, Non-Violent | Fight Rebel, Exit A |
| -4.0 | -3.0 | -2.0 | -1.0 | 0.0 | +1.0 | +2.0 | +3.0 | +4.0 |
| | | | | GS0 | GS1 | GS2 | GS3 | GS4 |

These actions are on abstract scale, which ranges from total support of the majority that is oppressing you (if you can't lick them, join them), to being undecided and/or helpless in the middle, to the other extreme of supporting and ultimately exiting A and joining the insurgency. At the extremes on either end, the agent will submit to militaristic commands of the leader of that group, while at the next level two lower levels they will be only willing to go to protests, and verbally and economically support the activities of that group's leaders. Thus, every state all the way through GS4 represent Voice. We only permit Exit from A and joining of C after occupying GS4 for a significant interval.

## Socio-cultural game results to date: Turing, correspondence, and senstivity testing

The previous section delineated the human behavior model of leaders and followers in terms of value trees and activation mechanics in PMFserv. Here we turn to an examination of how these archetypes work in a game. In this Section, we exam a scenario where a larger group of one religion and its leader (Blue) discriminate against two smaller groups (the moderate Villagers, Yellow, and the rebels, Red) of a different religion, both in terms of values as well as overt policies. The main policy concern here will be to find out: how should Blue leader address this problem so as to attain his own goals

yet also be more tolerant and prevent a full blown insurgency from being spawned? Why is violence rising? What are the consequences for domestic politics? What would be the best targets and times to intervene? This case will draw training and test data from events during the past three years in SE Asia. Only publicly available data sources have been used, but the sponsor requested this land remain unnamed, and hence we will refer to them solely as Blue, Yellow, and Red (see Fig. 15.3A).

**Correspondence test**

There were three types of data/ empirical information employed in this model:

- Numerical data as well as empirical materials on BlueLand, particularly the violent incidents occurring in the rural Yellowland provinces under the control of Blue,

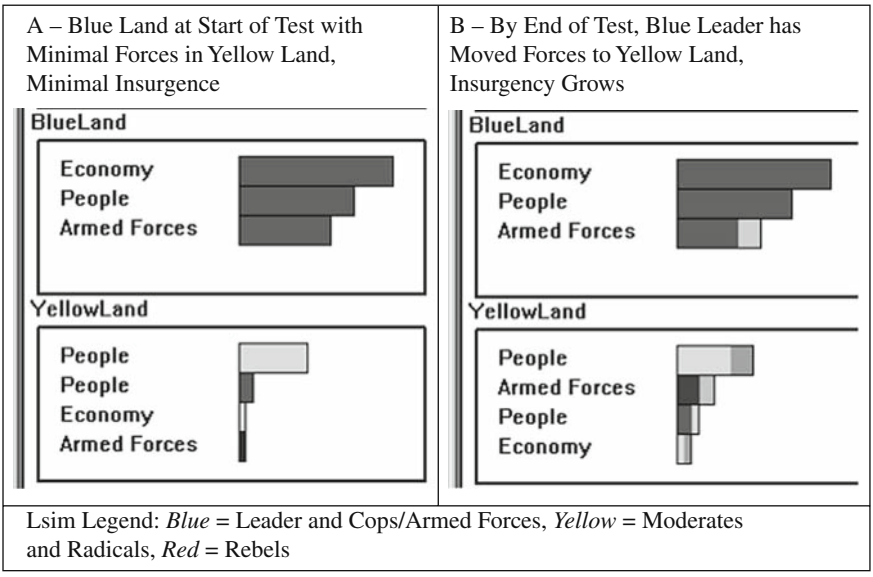



**Fig. 15.3.** Start and end states during the correspondence test: LeaderSim summary view. **Civil violence view of population membership before and after correspondence test.**

| C – Starting State (Avg of Weeks 1 & 2) Muslim Population at Start Is Neutral with Few Grievances Registering | | D – End State (Avg of Weeks 103, 104) Muslim Population Reflects Radicalization and Spread of NonViolent and Violent Protest | |
|---|---|---|---|
| GrievanceState0 - Neutral | 30% | 6% | GrievanceState0 - Neutral |
| GrievanceState1 - Disagree | 55% | 1% | GrievanceState1 - Disagree |
| GrievanceState2 - Join Oppost | 15% | 37% | GrievanceState2 - Join Oppost |
| GrievanceState3 - Nonviolent | 0% | 39% | GrievanceState3 - Nonviolent |
| GrievanceState4 - Fight-Rebel | 0% | 17% | GrievanceState4 - Fight-Rebel |
| TOTAL | 100% | 100% | TOTAL |

- Empirical information about the decisions made, along with the contexts of these decisions, by the specific personnel being modeled (the Leader of the Blue), and
- Culture specific information for the Blue and Yellow from such studies as GLOBE (House et al., 2004), as well as religious doctrines affecting the people of concern.

During the 1990s, the country was relatively stable, however, in the last few years, the rural provinces (YellowLand) have seen a rise of Muslim anger against the central Blue government, and the internal security situation in these provinces has rapidly decayed. Certain factions in YellowLand are seeking independence from BlueLand. During 2004, a small group of people, indicated as Red in Fig. 15.3A has committed an increasing number of violent acts against Buddhists (Blue people). The level and sophistication of the attacks has been increasing to the point where people are questioning whether there may be outsiders assisting this group. The reaction of the Blue Leader to these violent incidents has been generally viewed as heavy-handed, and even inappropriate. The Blue Leader has branded the separatists as bandits, and has sent the worst behaving police from the north (BlueLand) to handle all protesters in the YellowLand. There are many accounts of police brutality

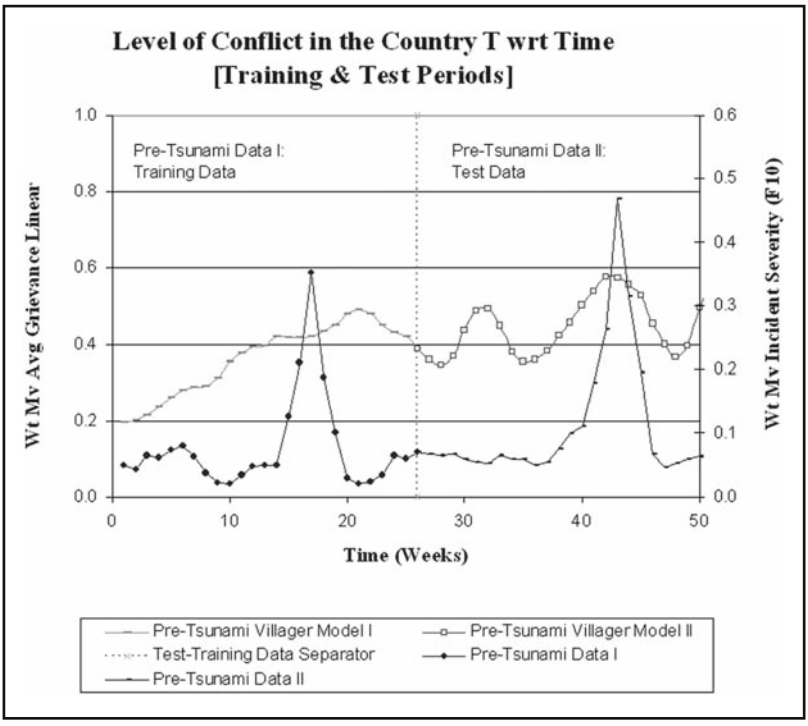



**Fig. 15.4.** Grievance from simulation and observed data (incident severity).

and civilian deaths. In December 2004, the Tsunami hit and ravaged portions of YellowLand. The massive arrival of relief workers lead to an interruption of hostilities, but these resumed in mid-2005, and Blue Leader declared martial law over YellowLand in the summer of 2005.

The violent incidents in the country were classified based on the size and intensity of the incident. The incidents were then aggregated and plotted against time to obtain a longitudinal plot of incidents (Fig. 15.4). The data was then longitudinally separated into "independent sets" with training set consisting of January–June 2004 while test set beginning in July 2004 and running till December 2004. We curtail the test data to end before the tsunami.

**Setting up the testbed and tuning it with the training dataset**

Training data and evidence were used to calibrate three types of agents in PMFserv:

- *Blue Leader* (structure of his GSP trees are in Fig. 15.3) – data indicates harsh, cruel, task, corrupt, wealthy, successful. Sends worst behaving cops down to YellowLand, never discourages brutality.
- *Moderate villagers* – Lack of cultural freedom, schools, etc. Want own land and autonomy.
- *Radical villagers* – Wahhabi and college-trained, unemployed, running religious schools in family homes.

In order to adequately test these PMFserv agents' ability to interact at the population level, the PMFserv run groups are connected to a cellular automata that is known as the Civil Violence model (Epstein et al., 2001), though Leader Legitimacy is replaced with PMFserv agents' view of membership. The Civil Violence model involves two categories of actors, namely villagers (or simply agents) and cops. "Agents" are members of the general population of YellowLand and may be actively rebellious or not, depending on their grievances. "Cops" are the forces of the BlueLand authority, who seek out and arrest actively rebellious agents. The main purpose of introducing the Civil Violence model is to provide a social network for the cognitively detailed PMFserv villagers to interact with. The social network consists of one layer of the normal arena or neighborhoods as well as a second layer of secret meeting places, simply represented as a school. Civil Violence agents can exist in more than one layer (namely in the normal as well as school layers), however, the PMFserv agents that show up in the school layer are only the young Wahhabi- and college-trained males.

The training data set also was used to fit the between-the-models parameters, especially between the PMFServ and CV model bridge and to tune up the Civil Violence villagers. Specifically, three types of cellular automata villagers were added:

- Neutral Villagers (these are modeled as simple agent automata in the CV model) – 1,360 of them exist. The simple villagers are uniformly distributed in terms of risk aversion, but derive their grievance from witnessing cop activities in their neighborhood, from polling neighbors for opinions, and from hearing about hardships and news from PMFserv agents they may be in contact within their own neighborhood or school.
- Moderate Villagers – there are 80 of these in Civil Violence who are controlled by 80 PMFserv agents. They influence neutrals via small world theory in different neighborhoods of the Civil Violence cellular automata.
- Radical Villagers – there are 80 of these in Civil Violence who are controlled by 80 PMFserv agents. They influence neutrals via small world theory in different neighborhoods of the civil violence cellular automata and in the school layer.

The bridge between PMFserv and Civil Violence includes Blue Leader and 160 villagers, and works as follows. Blue Leader examines the state of the world and makes action decisions to assist or suppress Red or Yellow (e.g., pay for Buddhist schools, add more cops, reduce cop brutality, etc.). The 160 PMFserv agents then assess their view of the world, react to how cops handle protester events, how their GSPs are being satisfied or not by leader actions, and to their emotional construals. The grievance level and group membership decisions by 160 archetypical villagers in PMFServ are passed via an XML bridge to 160 agents they control in the cellular automata based population model. These agents influence the neutrals of the population who spread news and form their own view of the situation. The number of Civil Violence villagers in each of the five states of the Grievance Scale (neutral through Fight Back) are added up and this information is passed back to PMFserv to help determine its starting level of grievance for the next cycle of reactions to Blue Leader actions. For the purposes of this writeup, the Red Group has no active agents, but is marked up as rebels that afford activations as mentioned above.

**Running the simulation**

The correspondence test is whether the overall parameterization for the GSP tree-guided PMFserv agents in the bridge with the Civil Violence population will faithfully mimic the test data set. That is, by tuning the GSP trees of 1 leader and 160 villagers, and by connecting all that to the Civil Violence mode of spreading news and grievances, do we wind up with a simulation that seems to correspond to what happened in the real world test dataset? Specifically, we are interested in testing the null hypothesis that there is no statistically significant correlation between real decisions and the simulated decisions. That is to say that real incidents and simulated base case are mutually independent.

The simulation starts on the left side of Figs. 15.3A for Lsim and 15.3C for Civil Violence. When the simulation is run, one observes Blue Leader trying some assistance measures initially (usually offering to set up Buddhist

school and institutions) but maintaining a high police presence, and turning increasingly suppressive as the run proceeds – Suppressing by Increasing Militarization and by Increasing Violence Unleashed. The end state is reflected in Fig. 15.3B and D for each view, respectively. We can also examine what happened as the run proceeded. Figure 15.5A shows the average PMFserv villager perceptions of the Blue Leader actions in terms of the Dangerous Ideas model's terms – Vulnerability, Injustice, Distrust. Initially, Moderate Villagers respond positively to needed assistances given by the Blue Leader (negative VID and grievance is positive support). However, once they are suppressed violently and lose faith in the government, they tend to disagree with even positive government decisions. Radical Villagers start out disagreeing with Blue Leader and shift to "fight back", an action that might continue for a long time before they realize the helplessness of the situation and abandon membership in the moderate side, and join the opposition.

Figure 15.5B shows the output of the Civil Violence model being sent back to the PMFserv villagers. Specifically, it shows what percent of the population has been shifted from Neutral Grievance to higher states (recall the scale earlier): GS0 (neutral) through GS4 (fight back). From the first graph, it can be seen that at the start, most villagers are neutral and occupy GS0 while a small percent start in GS1. Many of them rapidly shift to GS1 (disagree), then abandon that and shift to GS2 and higher states. The occupancy in lower grievance states fall with time, while that in higher grievance states climb. From about week 50 onwards, there is a fairly stable, though regularly punctuated equilibrium in which the highest occupied states are GS3 and GS4. This is an indication of progressive escalation of violence in the society.

In order to compare this simulated grievance to that of the real world, we need some reliable measures of the population's grievance during actual events. Unfortunately, there are no survey or attitude results available. In the real world (test) dataset, the incident data was available, however, with a record of fatalities and injuries. There are a number of schemes for weighting those (e.g., depression and morale loss, lost income, utility metrics, others), however, here we take the simple approach of just computing a weighted incident severity. We computed incident severity scores using weighted average of fatalities and injuries, where injuries are simply counted, but the weight on fatalities is 100. $IncidentSeverity = w_f \times fatalities + w_i \times injuries$. The result serves to indicate how severe these incidents were. While severity is only an indirect measure of how the population might have felt, it is a measure that can be tested for correlation to the rise and fall of grievance expression due to leader actions in our simulated world.

To conduct the comparison, we apply the non-parametric Kendall's Tau measure of correlation. This statistic estimates the excess of concordant over discordant pairs of data, adjusted for tied pairs. With a two sided test, considering the possibility of concordance or discordance (akin to positive or negative correlation), we can conclude that there is a statistically significant lack of independence between base case simulation and observed grievances

rankings at a confidence interval of 88%. Since there is a probabilistic outcome determining if a simulated leader's action choice will result in injury and fatality incidents (and how the news of these events are propagated through the cellular automata is probabilistic as well), we repeated the simulation runs thirty times and the confidence interval mentioned above is the mean across those 30 correlations. In sum, the null hypothesis is rejected and real (test interval) incident data and simulation results are related.

### Validation

In order to assess the validity, we examine what transpires inside the heads of the various types of agents in the simulated world. In the test dataset, the real world leader made 52 decisions affecting the population and that we sorted into positive, neutral, and negative actions. In the simulated world, Blue leader made 56 action decisions in this same interval. At this level of classification (positive, neutral, negative), we were able to calculate a mutual information or mutual entropy (M) statistic between the real and simulated base cases. M ranges from 0 to 1.0, with the latter indicating no correlation between two event sets X and Y. $M(X : Y) = H(X) - H(X|Y)$ where X and Y are the simulation and historic sources, respectively, and $H(.)$ is the entropy function, defined by:

$$H(X) = -\sum p(x)_i \log p(x).$$

Applying this metric, the mutual entropy values were found to be less than 0.05, indicating correlation between real and simulated data. Details of this validation as well as villager validation are given in the Appendix II. With an M metric, one cannot make statements about the confidence interval of the correlation, however, the Blue Leader in the current scenario seems faithful

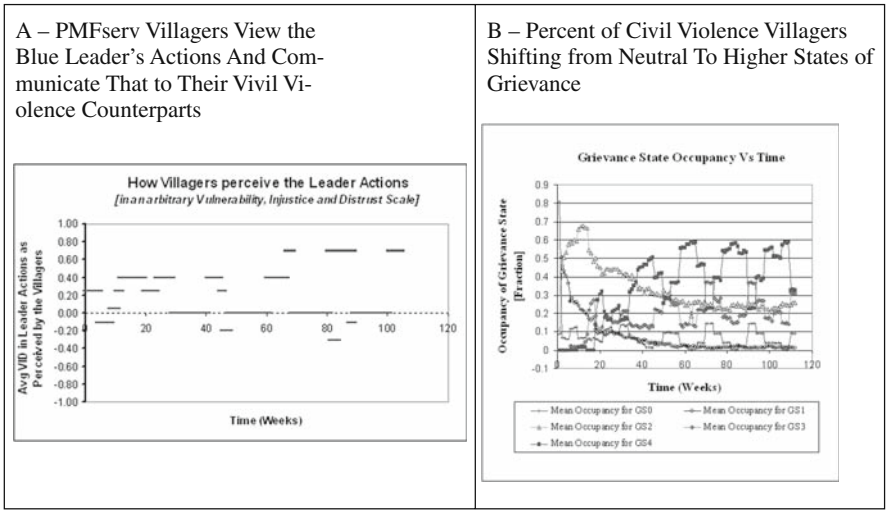



**Fig. 15.5.** Villager grievance states.

to his real world counterpart. This gives us reason to suspect the Hermann- and GLOBE-based GSP tree structure works equally well across time periods, locations, and cultures.

Not all aspects of the agent mindset work as well as this, however. As an example, one would expect agents to waiver somewhat in their resolve, yet our mechanism for that waivering seems like it may be too heavy-handed. The reader will recall that at the end of each cycle, the XML bridge feeds average Civil Violence neighborhood grievance back to the PMFserv agents in that neighborhood as a dampening of their reactions. This feedback is in the form of a replacement of the PMFserv agent's memory of past grievance level. Resetting of past grievance leads to flip-flopping behavior of PMFserv agents. This causes moderate villagers to flip between supporting opposition (GS2) and joining government (GS-2) depending on what Blue Leader does, particularly during the early stages. Hence they seem overly fickle. Since radicals are more grieved, their flip-flops are between neutral (GS0) and fight back (GS4), and this seems less troublesome semantically, though a confirmed Jihadist probably has fewer of these issues. We are not sure if this amount of flip-flopping is warranted, and this could only be resolved by further behavioral studies of such individuals (studies of which may be underway by Atran 2006). If research shows them to be less fickle in their inner beliefs, this is a relatively easy process to dampen in future versions.

**Sensitivity analysis**

Once again, it is interesting to explore "what-ifs" and whether alternative decisions by Blue Leader will yield different outcomes amongst the YellowLand population. The reader will recall, however, that our model's output is conflict parameters (action decisions, grievances, group membership), whereas the model's inputs are characteristics of the leader and the followers. To change the outputs implies shifting the weights on the GSP trees of various archetypes of the population. Here we shift those weights for the Blue Leader, since we are interested to see if his personal decision style and choices are key to driving the villagers toward insurgency.

- By altering Blue Leader's InGroup Bias we should be able to alter his decisions to provide more or less needed assistance (economic goods, non-Wahabbi schools), and then we can observe if that alters the outcome. Specifically, we perturb InGroup Bias on his Standards Tree by 15% in either direction. Figure 15.6A shows the result.
- By altering Blue Leader's Sensitivity to Life (Humanitarianism) we should be able to alter his decisions to provide more or less violent cops, and then we can observe if that alters the outcome. Specifically, we perturb SensitivityToLife on his Standards Tree by 15% in either direction. Figure 15.6B shows the result.

- By altering Blue Leader's Openness we should be able to alter the immediacy of his response to opposition and protest. Thus he would send fewer cops down to YellowLand if he were more open, and more cops if he were less open, and then we can observe if that alters the outcome. Specifically, we perturb Openness on his Standards Tree by 15% in either direction. Figure 15.6C shows the result.

In examining these three sets of what-ifs in Fig. 15.6A, B, and C, as expected, a larger fraction of population occupies higher grievance states of 4 and 3, when the leader exhibits lesser degree of sensitivity-to-life and/or more InGroup Bias. Conversely, the population remains at lower grievance states when Blue Leader is more sensitive-to-life and less InGroup Biased. However, the trend is not the same with respect to the openness trait of the leader. It appears that more open leadership does not necessarily result in lower grievances in the community, but in less sustained (shorter) expressions of grievance. These types of results help us begin to calibrate the population's demand curve mentioned earlier for exit, loyalty, and voice. Specifically, we presented several equations that help to determine the members' decisions about expressing their grievance (voice) and/or exiting the legitimate authority and joining the separatist movement. In Fig. 15.6D, we see the graphical expression of several of those equations. We plot loss of civil rights and growing vulnerability, injustice, and distrust (VID) up the vertical, with group members along the horizontal. The demand curve is negatively sloped indicating people tend to increase their strength of membership in group A (BlueLand in this instance) as VID drops. The Transfer Rate helps to define the elasticity or slope of the demand curve. As the policies of the Blue Leader are altered for more Sensitivity to Life and Less InGroup Bias (Fig. 15.6B and A, respectively), this is equivalent to shifting from point 2 on the demand curve toward point 1. On the other hand, more InGroup Bias and less Sensitivity to Life shift the YellowLand to point 3. In Fig. 15.6D, we had the means to compute the desire to exit, and this is plotted along the horizontal axis. To the right of point 1, $\delta\Phi$ is below zero, and few members favor the separatism. This is how YellowLand in fact was in the 1990s. In the interval between 1 and 3, the YellowLand people are indifferent since separatism cost is so high. There, they use their Voice and we observe the protests that occurred in the simulation runs of this paper, and that in fact occurred in the real world test dataset. Finally, to the left of point 3, Blue Leader's treatment becomes so intolerable, that he loses the hearts of the Yellow People and even the moderate followers are now on the side of the Red separatists or insurgents. In fact, in late summer of 2005, the Leader of this land had to declare Martial Law complete with curfews and movement checkpoints. The sign of our simulation results thus correspond well to the real world, and give us an ability to suggest outcome possibilities that are realistic for the Leader's policy choices.

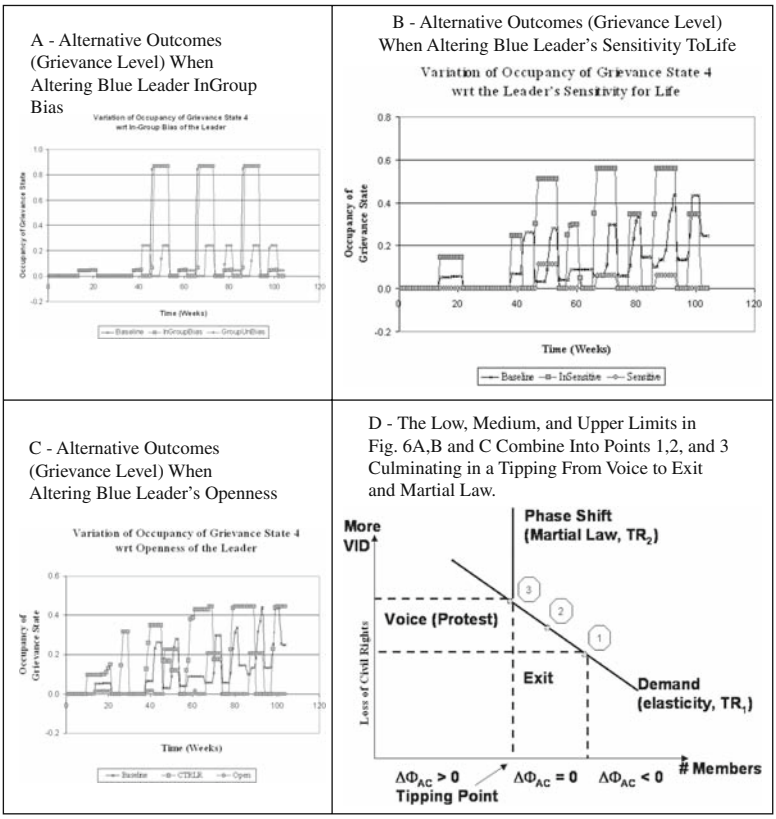



**Fig. 15.6.** Alternative outcomes.

## Lessons learned and next steps

In concluding, it is useful to revisit the three aims of the introduction, and to see what has been learned in each of them and to point out some items seen as priorities for further development.

Aim 1 was to create a role-playing game generator where one could rapidly set up and play out numerous conflict scenarios from around the world. Conflicts arise when groups vie over the control and allocation of resources (land, economy, markets, militias, media outlets, followers, etc.). Socio-cultural aspects concern any perceived injustices that have arisen historically with respect to these allocations, where perception is a matter of the value systems, norms/standards, and emotional utility of the perceivers. An example of its usage can be seen in SE Asian scenario described in this paper. The game generator was shown to reduce conflicts to the bare essentials that allow one to explore the intertwined issues affecting welfare (economy, in-group standards,

health services), security (freedoms/liberties, military), and political support for leaders (popularity of positions).

In zero sum games, what one spends on actions affecting one area of welfare, security, or populace effects what one has to allocate to other areas. Borrowing from diplomatic video games the idea here is to make the game immersive and engaging, and to date hundreds of players have participated in multi-hour sessions that they were unwilling to terminate. All this gameplay also gave us a rich source of data to help guide the construction of agents who can serve as synthetic opponents, allies, followers, and the like. Also, we have learned that our game state representations are intuitive and that domain experts can readily use them to express conflict scenarios that are hard to verbalize. As with anything done in software, there are always next levels of sophistication and detail that one can add, and we identified many new features we would like to add such as, to mention a few examples, (1) scale up of all features shown here for the larger game generator we call Athena's Prism; (2) resources and assets (e.g., economy and black markets) that are supported by institutions that grow more self-sustaining and resilient, the larger they are; and (3) logging services and explanation functions that help users to generate reports on model outcomes, agent decision choices, and effects. These are some of the laundry list of next steps for the game generator.

Modeling leaders and followers is a complex enterprise and one would like to use only first principles of social science, yet the field has not matured sufficiently. Still, that is no excuse for modelers to "make up" their own rules and algorithm for how groups behave, nor is it justification to just create entertaining agents. The alterative we explored here is to try and adopt best-of-breed and well-respected social science models for leadership, group dynamics, and the hearts and minds of the populace (Aim 2). These models are implemented atop a unified architecture of cognition, call PMFserv that manages six modules of an agent's mind: memory, perception, physiology/ stress/ coping level, value system, and emotional construal, relationships and models of other, and (stress and emotion-constrained) decision making processes. PMFServ exposes many parameters in each of these modules and permits analysts/developers to visually "program" best-of-breed social science models that govern how the modules work, and in turn, how that agent tends to behave. This framework supported the ready implementation of leader models from Hermann (style), Hofstede and Globe (cultural factors), and Heuer (biases) atop pre-existing models in the PMFServ modules. These synthetic leaders passed the Turning and Correspondence tests, where the leader and minority villagers attempted to maximize his respective economic welfare, security, and populace resources in accord with his GSP trees of goals, standards, and preferences in the game scenarios. It was no surprise that leaders' biased models of others often proved to be self-fulfilling prophecies. The PMFServ modules allowed group followers to be readily modeled via their personal motivations (Maslow-style), group member factors (injustices, vulnerabilities, etc.), and loyalty decisions

(follow happily, helplessly, vocally, separate, etc.). Again the followers' behavior passed Turing and Correspondence tests of Muslim moderates and radicals as the outgroup leader's policies shifted: a real world case study was used. Our population model involved a cellular automata with 1,360 agents influenced in their neighborhoods and schools by 160 PMFserv agents. Inside the PMFserv agents, one can readily observe and track their GSP tree implementation of Maslow, Hofstede, and GLOBE factors, and preference functions. One can follow how they update the Eidelson model factors of group and leader achievement. One gains confidence that these agents are realistic, particularly when one can calibrate them with validated instruments such as Hermann's profiling method or Eidelson's IGBI instrument, just as is done for real world human participants: e.g., see Maoz and Eidelson (2007).

## References

Bharathy, G. K. (2006). *Agent Based Human Behavior Modeling: A Knowledge Engineering Based Systems Methodology for Integrating of Social Science Frameworks for Modeling Agents with Cognition, Personality & Culture.* (Doctoral dissertation, University of Pennsylvania, July 2006)

Damasio, A. R. (1994). *Descartes' Error: Emotion, Reason, and the Human Brain.* New York: Avon.

De Marchi, S. (2005). *Computational and Mathematical Modeling in the Social Sciences.* Cambridge: Cambridge University Press.

Epstein, J., Steinbruner, J. D., Parker, M. T. (2001), Modeling civil violence: An agent-based computational approach. *Proceedings of the National Academy of Sciences.* Washington DC: Brookings (CSED WP#20).

Green, K. C. (2002). Forecasting decisions in conflict situations: a comparison of game theory, role-paying, and unaided judgement. *International Journal of Forecasting,* 18, 321–344.

Hermann, M. G. (1999). *Assessing Leadership Style: A Trait Analysis.* Hilliard, OH: Social Sci. Automation, Inc.

Heuer, R. J., Jr. (1999). *Psychology of Intelligence Analysis.* Washington, DC: Center for the Study of Intelligence, Central Intelligence Agency.

House, R. J. Hanges, P. J., Javidan M, et al. (2004), *Culture, Leadership, and Organizations: The GLOBE Study of 62 Societies.* Thousand Oaks, CA: Sage Publications.

Johns, M., Silverman, B. G. (2001). How emotion and personality effect the utility of alternative decisions: A terrorist target selection case study. *Proceedings of the 10th CGF & BR, SISO.* Norfolk, Virginia.

Johns, M. (2006). *Deception and Trust in Complex Semi-Competitive Environments.* (Doctoral dissertation, University of Pennsylvania.

Maoz, I., Eidelson, R. J. (2007). Psychological bases of extreme policy preferences: How the personal beliefs of Israeli Jews predict their support for population transfer in the Israeli-Palestinian conflict. *American Behavioral Scientist*, 50, 1476–1497.

Ortony, A., Clore, G. L., Collins, A. (1988). *The Cognitive Structure of Emotions.* Cambridge: Cambridge University Press.

Pew, R. W., Mavor, A. S. (1998). *Modeling Human and Organizational Behavior: Application to Military Simulation.* Washington, DC: National Academy Press.

Silverman, B. G., Johns, M., Weaver, R., O'Brien, K., Silverman, R. (2002a). Human behavior models for game-theoretic agents. *Cognitive Science Quarterly,* 2(3/4), 273–301.

Silverman, B. G., Johns, M., O'Brien, K., Weaver, R., Cornwell, J. (2002b). Constructing virtual asymmetric opponents from data and models in the literature: Case of crowd rioting. *Proceedings of the 11th Conference on Computer Generated Forces and Behavioral Representation,* Orlando, Florida, 97–106.

Silverman, B. G. (2005). Human performance simulation. In J. Ness, D. Ritzer, V. Tepe (Eds.), *The Science and Simulation of Human Performance (Chapter 9).* New York: Elsevier.

Silverman, B. G., Rees, R, et al. (2005, May), Athena's prism: A diplomatic strategy role playing game for generating ideas and exploring alternatives. *Proceedings of the Internat'l Conference on Intelligence Analysis*, MacLean, VA: Mitre.

Silverman, B. G., Johns, M., Cornwell, J., O'Brien, K. (2006a). Human behavior models for agents in simulators and games: Part I – enabling science with PMFserv. *Presence* , v. 15: 2 April.

Silverman, B. G., Bharathy, G. K., O'Brien, K., Cornwell, J. (2006b). Human behavior models for agents in simulators and games: Part II – gamebot engineering with PMFserv. *Presence,* v. 15: 2, April.

Silverman, B. G., Bharathy, G. K., Johns, M., Nye, B., Eidelson, R., Smith, T., *Socio-Cultural Games for Training and Analysis.* Tech Report avail from the authors (submitted for publication 2006c).

Wood, E. J. (2003). *Modeling Robust Settlements to Civil War: Indivisible Stakes and Distributional Compromises.* Santa Fe, Unpublished Paper. Available from the author: wood@santafe.edu

## APPENDIX I: Stylized game description

The agents in our model participate in a multi-stage, hierarchical, n-player game in which each class or type of agent ($D_x$) observes and interacts with some limited subset of y agents (human or artificial) via one or more communication modalities. We make three empirically plausible assumptions about multiple hierarchies of agents: (1) play multiple distinct games, (2) are cognitively detailed, (3) agents are self-serving, and attempt to maximize its utility (u) within this iteration of the game, as follows:

$$GAME = (a \in A, U_x, D_x) \forall x \in X \tag{15.6}$$

In this case, we set it up as a game (similar to that of Wood, 2003) between the Leader of the CountryT (Leader or L) and the minority villagers (V), who have conflicting-interests. The leader wants to divert his resources to his constituency, thereby discriminating against this minority. If they rebel, the leader would not hesitate to use violence (fight or f). If the leader were to compromise, he would not help his constituency in the short term (in turn

may lose some support or votes especially from extremists), but would manage to settle the conflict (compromise or c). The outcomes are as in Fig. 15.7 for single shot and iterative games.

Ranges of probabilities of belief required of opponent's likelihood of compromise, over which a party will compromise, could be obtained as shown in the figure (robust equilibriums with symmetry). While such stylized games are useful to elegantly illustrate the division of tangible and divisible resources as a zero sum game, they do not take into account the intangible resources such as emotional and cultural pay off (as well as indivisible resources). On the other hand, our simulation models do handle these added factors.

Single Shot:

| | | Villager | |
|---|---|---|---|
| | | Fight | Compromise |
| Leader | Fight | $f_{1L}, f_{1V}$ | $f_{2L}, 0$ |
| | Comp | $0, f_{2V}$ | $c_{3L}, c_{3V}$ |

Mutual conflict or fight-fight are Nash equilibriums, as compromising while the other player fights have a low pay off. $c_{3V} + c_{3L} = c_3 and c_{3L} = y.c_3$ and $c_{3V} = (1-y).c_3$ If $c_{3V} > f_{2V}$ and $c_{3L} > f_{2L}$, then mutual compromise is also a Nash equilibrium. This zero sum game assumes, fraction y of the total compromised divisible pay off ($c_3$) is allocated to the leader.

Repeated (as negotiated solutions typically are):

| | | Villager | |
|---|---|---|---|
| | | Fight | Compromise |
| Leader | Fight | $f_{1L}.(1+i)/i$, $f_{1V}.(1+i)/i$ | $f_{2L}+f_{1L}.[1/i]$, $f_{1V}.[1/i]$ |
| | Comp | $f_{1L}.[1/i]$ $f_{2V}+f_{1V}.[1/i]$ | $c_{3L}.(1+i)/i$, $c_{3V}.(1+i)/i$ |

If future discount rate (rate of time preference) is i for both parties, then over infinite horizons: Initially negotiate or fight as per single shot game. Simultaneous, repeated game with information input from historic events. Mutual compromise or conflict, once established, tends to persists in the absence of exogenous shocks. For mutual compromise,
$c_{3L}.(1+i)/i > f_{2L} + f_{1L}.[1/i]$ and
$c_{3V}.(1+i)/i > f_{2V} + f_{1V}.[1/i]$

Let $p_{LVC}$ and $p_{VLC}$ be the probabilities estimated by the leader and villager for compromising by villagers and leader respectively. Using these, one may estimate the expected payoff for:

- the leader to compromise as:
  $E(LC) = p_{LVC} * c_{3L}.(1+i)/i + (1-p_{LVC}) * \{f_{1L}.[1/i]\}$
- the villagers to compromise as:
  $E(VC) = p_{VLC} * c_{3V}.(1+i)/i + (1-p_{VLC}) * \{f_{1V}.[1/i]\}$

Similarly, the estimated pay off for:

- the leader to fight is:
  $E(LF) = (1-p_{LVC}) * \{f1L.(1+i)/i\} + p_{LVC} * \{f_{2L} + f_{1L}.[1/i]\}$
- the villagers to fight is:
  $E(VF) = (1-p_{VLC}) * \{f_{1V}.(1+i)/i\} + p_{VLC} * \{f_{2V} + f_{1V}.[1/i]\}$

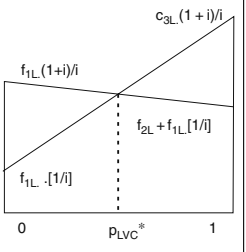



**Fig. 15.7.** Outcomes for single shot and iterative games.

## APPENDIX II: Validation

This appendix provides a more detailed look at the runs and results between simulated vs. actual leaders and followers.

### Correspondence between simulated vs. actual leader decisions

This correspondence test has been attempted with the use of the mutual entropy statistic. Our estimation of mutual entropy for CountryT is much less than 1, indicating that the real and simulated base cases might correspond well. Currently, there are no benchmarks that could indicate what would be an acceptable limit of mutual entropy for establishing correspondence. If we accepted an arbitrary limit of correspondence of mutual entropy less than or equal to 0.1 [an order of magnitude less than the mutual entropy associated with no correspondence], then we would be able to reject the null hypothesis of no-correspondence between real and simulated outputs. The following figure shows a summary of decisions of the simulated and actual leaders through the bar charts on the left and right respectively. The entropy calculations are shown below the charts.

### Villager behavior correspondence

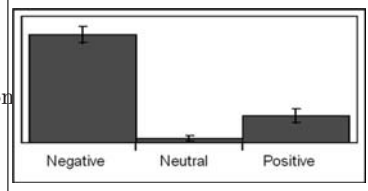
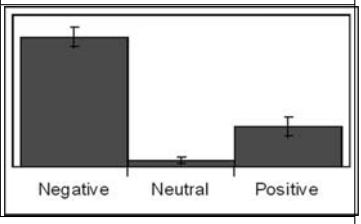
We have employed Kendall's Tau to relate the real vs. simulated population's grievance data. Kendall's Tau is a measure of correlation, and so measures the strength of the relationship between two variables. It employs paired observations, and is scale-free. It is computed as the excess of concordant over discordant (nd) pairs, divided by a term representing the geometric mean between the number of pairs not tied on variable 1 and the number not tied on variable 2. There is no well-defined intuitive meaning for Tau -b, which is the surplus of concordant over discordant pairs as a percentage of concordant, discordant, and approximately one-half of tied pairs. We obtained two sets of Kendall's Tau (KT) values by comparing:

- simulated agent grievances from each base case simulation against every other base case simulation, and
- real population incident severities against the simulated grievances from every base case simulations.

It can be seen that KT values among base cases with moving averages are distributed in bins 3 (0.0–0.25), 4 (0.25–0.50), and −6 (0.75–1.0), with proportions increasing in that order. This implies that a large majority of the pairwise comparisons (a little over half) among base cases result in high Kendall's Taus, while a smaller fractions have limited correlation. Comparisons between real data (injury-fatality aggregated and smoothened, of course)

and base cases (moving averages) are found in bins 3 (0.0–0.25) and 4 (0.25–0.50). This shows that all base cases show positive correlation with real data. A Kendall's Tau of 0.25 or 0.5 might appear to be a small correlation compared to a KT value of, say, 1.0, but in reality, these numbers indicate a fairly good degree of correlation, especially considering this is a time series and any mismatch would be counted as discordance resulting in negative correlation. This is illustrated by converting Kendall's Tau to p values, as seen below. Although the real data is an outlier, it is no more outlier than about half the simulation base cases themselves.

While we recognize that the p values are considered weak in the case of Kendall's Tau, and therefore exercise caution in the interpretations, it is hard to not notice p values for the same Kendall's Tau. The range of p values from the models run is less than 5%. If one were to take the approach of significance level, this would mean that one would be able to reject the hypothesis of mutual independence between the base case simulation grievance and observed incidents (villager or follower decisions) with a significance level of 0.05. With a two-sided test, we are considering the possibility of concordance or discordance (akin to positive or negative correlation). In our example, we can conclude with reservation (weak indicator, and hence the caution) that there is a statistically significant lack of independence between base case simulation and observed grievances rankings. That is, real incident data and simulation results are related (Figs. 15.8, 15.9 and 15.10). (From these range of results, and based on the alpha value of 5%, it seems reasonable to reject the null hypothesis).

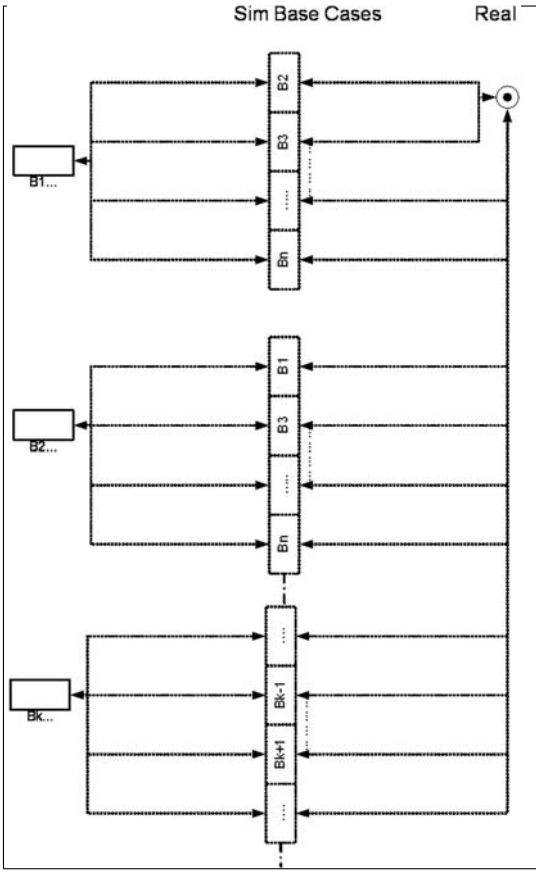
| | PMFserv-Simulated Prime Minister's Actions | Real Leader's Chosen Actions |
|---|---|---|
| Distribution |   |   |

| Mutual Entropy Calculations | | | |
|---|---|---|---|
| | Joint Entropy of Sim & Real | 1.396 | $H(SIM, REAL) = -\Sigma\, p(sim_i, real) \log p\,(sim_i, real_i)$ |
| | Entropy of Sim | 0.681 | $H(SIM) = -\Sigma\, p(sim)_i \log p(sim)$ |
| | Entropy of Real | 0.760 | $H(REAL) = -\Sigma\, p(real)_i \log p(real)$ |
| | Mutual Entropy of Sim & Real | 0.045 | $M(SIM, REAL) = H(SIM) - H(SIM\ REAL)$ |

Legend of Leader Actions

| Negative Actions: | Neutral Actions: | Positive Actions: |
|---|---|---|
| Discriminate | Perceive (Observe Events) | Give Culturally Sensitive Assistance |
| Suppress - Increase Number of Cops | | Give Essential Assistance |
| Suppress - Increase Violence of Cops | | Reduce Suppress by Number |
| | | Reduce Suppress by Violence |

**Fig. 15.8.** Correlation of simulated leader vs. real action decisions.




We took averages of Kendall's Tau for EACH base case grievance, against EVERY OTHER base case grievance. As has been illustrated in the left of the Fig. 15.9, each base case simulation output was pitted against every other base case output. For each base case simulation, the Kendall's Tau obtained with every other base case is averaged to obtain a value of KT for that given base case. Then, we plotted the average frequency of occurrence of KT values, obtained by binning the KT values into one of the bins shown in the figure. We compared visually the two sets of Kendall's Taus obtained as above. This comparison gave structural independence, and was similar to the "jack-knifing" technique.

**Fig. 15.9.** Pairwise comparison of base case simulations and real data.

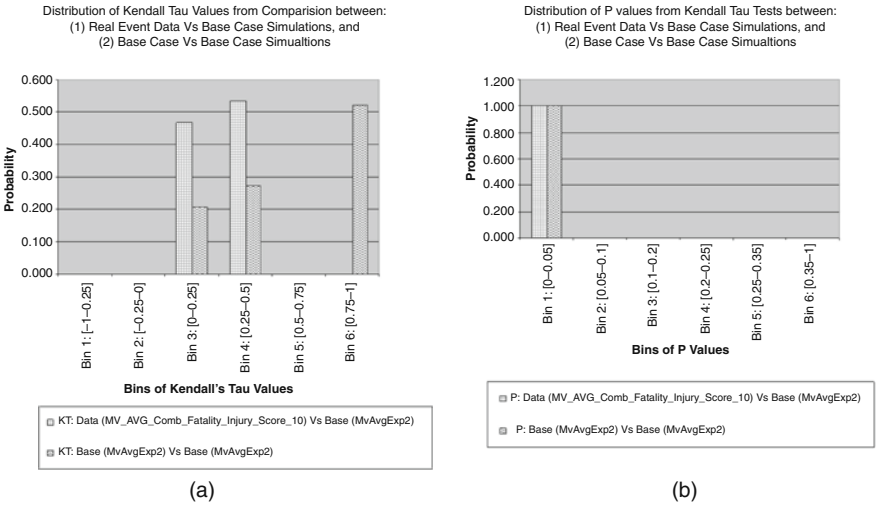



**Fig. 15.10.** Statistical distributions for data and base simulation. (**a**) Kendall's Tau; (**b**) $p$-values.

# Index