

Terrorism Knowledge Discovery Project: A Knowledge Discovery Approach to Addressing the Threats of Terrorism

Edna Reid¹, Jialun Qin¹, Wingyan Chung¹, Jennifer Xu¹, Yilu Zhou¹,
Rob Schumaker¹, Marc Sageman², and Hsinchun Chen¹

¹ Department of Management Information Systems, The University of Arizona,
Tucson, AZ 85721, USA
{ednareid, qin, wchung, jxu, yiluz,
rschumak, hchen}@bpa.arizona.edu

² The Solomon Asch Center For Study of Ethnopolitical Conflict,
University of Pennsylvania, St. Leonard's Court, Suite 305, 3819-33 Chestnut Street,
Philadelphia, PA 19104, USA
sageman@sas.upenn.edu

Abstract. Ever since the 9-11 incident, the multidisciplinary field of terrorism has experienced tremendous growth. As the domain has benefited greatly from recent advances in information technologies, more complex and challenging new issues have emerged from numerous counter-terrorism-related research communities as well as governments of all levels. In this paper, we describe an advanced knowledge discovery approach to addressing terrorism threats. We experimented with our approach in a project called Terrorism Knowledge Discovery Project that consists of several custom-built knowledge portals. The main focus of this project is to provide advanced methodologies for analyzing terrorism research, terrorists, and the terrorized groups (victims). Once completed, the system can also become a major learning resource and tool that the general community can use to heighten their awareness and understanding of global terrorism phenomenon, to learn how best they can respond to terrorism and, eventually, to garner significant grass root support for the government's efforts to keep America safe.

1 Introduction

Ever since the 9-11 incident, the multidisciplinary field of terrorism has experienced tremendous growth. The field has benefited greatly from increased national attention and a rich reservoir of knowledge accumulated through three decades of terrorism studies. Now, information-related issues, such as the communication and sharing of research ideas among counterterrorism researchers and the dissemination of counterterrorism knowledge among the general public, become critical in detecting, preventing, and responding to terrorism threats. The recent advance in information technologies, especially Web technology, has alleviated the problems to some extent. How-

ever, more complex and challenging new issues keep emerging from numerous terrorism-related research communities as well as local, state, and Federal governments.

Terrorism threats have a wide range that spans personal, organizational, and societal levels and have far-reaching economic, psychological, political, and social consequences [14, 21]. A recent report from the National Research Council [12], “*Making the Nation Safer: the Role of Science and Technology in Countering Terrorism*,” has summarized the major issues and challenges as revolving around three different aspects: *Terrorism, Terrorist, and Terrorized (Victims)*.

The first aspect (Terrorism) of the challenges is mainly associated with Information searching and management and knowledge creation and dissemination issues in the terrorism research domain. Currently, there are large and scattered volumes of terrorism-related data from diverse sources available to analyze terrorist threats and system vulnerabilities [12]. However, maximizing the usefulness of the data is a challenge because of: (1) the lack of counterterrorism-related databases that integrate these diverse sources and make them available to researchers, and (2) absence of advanced as well as new methodologies to identify, model, and predict linkages among terrorists, their supporters, and other perpetrators. Furthermore, for researchers new to terrorism, information access and management is a major challenge, especially in reference to identifying where to start, what to focus on, what types of data are available, where to obtain such data, and whether empirical studies are available. Thus, advanced techniques to support intelligent information searching and techniques to analyze and map terrorism knowledge domains are urgently needed.

The second aspect (Terrorist) of the challenges is mainly associated with how to trace dynamic evolution of terrorist groups and how to analyze and predict terrorists’ activities, associations, and threats. While the Web has evolved to be a global platform for anyone to use in disseminating, sharing, and communicating ideas, terrorists are also using the Web to their own advantages. Terrorist-generated online contents and the terrorists’ Web usage patterns could be analyzed to enable better understanding and analysis of the terrorism phenomena. Unfortunately, such terrorist-generated information has seldom been used in traditional terrorism research. On the other hand, since the amount of terrorist-related information has well exceeded the capability of traditional analysis methods, applying advanced techniques such as Social Network Analysis may well provide a significant value-add.

The last aspect (Terrorized) of the challenges mainly involves how to successfully give educators, students and the public systematic access to system-level thinking about terrorism research. As recommended in the “*Making the Nation Safer*” report, more research needs to be conducted on preparedness for terrorism attacks, human responses to terrorism crises as well as the strategies for providing people with necessary knowledge of terrorism. Thus, how to utilize various information technologies in achieving these goals remains an interesting and challenging problem.

To address the above challenges, the “*Making the Nation Safer*” report recommended establishing a terrorism research infrastructure that uses information fusion techniques because all phases of counterterrorism efforts require that large amounts of information from many sources be acquired, integrated, and interpreted. Information fusion includes data mining, data integration, language technologies for information extraction and multilingual retrieval, and data interpretation techniques such as

visualization. These techniques would also be useful for evidence-based analysis in law enforcement, the intelligence community, emergency-response units, and other organizations involved in counterterrorism.

In the light of the foregoing, the University of Arizona's Artificial Intelligence (AI) Lab is developing Web-based counterterrorism knowledge portals to support the analysis of terrorism research, dynamically model the behavior of terrorists and their social networks, and provide an intelligent, reliable, and interactive communication channel with the terrorized (victims and citizens) groups. Specifically, the portals integrate terrorism-related multilingual datasets and use them to study advanced and new methodologies for predictive modeling, terrorist (social) network analysis, and visualization of terrorists' activities, linkages, and relationships.

The remainder of this paper is organized as follows. Section 2 reviews related research in information searching, analysis, and visualization techniques in relation to the counterterrorism domain. In section 3, we present our research questions. Section 4 describes the knowledge portals we are developing to address the three aspects of challenges in the counterterrorism domain, which include knowledge portals that help resolve the information access and management problems in terrorism research and support the exploration of knowledge creation patterns of contemporary terrorism researchers; knowledge portals that use advanced methodologies to analyze and visualize how the web is used by terrorist groups, how relations are formed and dissolved among terrorists, and detect organizational groups and their tasks; and a knowledge portal using the chatterbot framework to support citizens' and victims' responding to terrorism. The final section provides our concluding remarks in the interim.

2 Existing Approaches

Many information-searching and analytical approaches have been adopted in the academia and industries. The following sections review some of these techniques in relation to the counterterrorism domain.

2.1 General-purpose and Meta-search Engines

Many different search engines are available on the Internet. Each has its own performance characteristics primarily defined by its algorithm for indexing, ranking and visualizing Web documents. For example, AltaVista and Google allow users to submit queries and retrieve Web pages in a ranked order, while Yahoo! groups Web sites into categories, creating a hierarchical directory of a subset of the Internet.

Internet spiders (a.k.a. crawlers), have been used as the main program in the backend of most search engines. These are programs that collect Internet pages and explore outgoing links in each page to continue the process. An example includes the World Wide Web Worm [27]. Most prevailing search engines, such as Google, are keyword-based. Although their search speeds are fast, their results are often overwhelming and imprecise. Low precision and low recall rates make it difficult to obtain specialized, domain-specific information from these search engines.

Selberg and Etzioni [37] have suggested that by relying solely on one search engine, users could miss over 77% of the references they might find most relevant because no single search engine is likely to return more than 45% of relevant results. A study by NEC Research Institute drew similar conclusions and revealed an alarming fact about Internet search engines: they cannot keep up with the net's dynamic growth, and each search engine covers only about 16% of the total Web sites [24].

The emergence of meta-search engines provides a credible resolution of the aforementioned limitations by triangulating outputs from several engines to arrive at relevant results. Several server and client-based meta-search engines, such as Copernic (<http://www.copernic.com>) "search the search engines" [37]. The results from other search engines are combined and presented to users. Although the information returned is comprehensive, the problem of information overload worsens if no post-retrieval analysis is provided.

2.2 Terrorism Research Centers' Portals

Web portal services provide another approach for retrieving information. For terrorism, there exists numerous information portals provided by specialized research centers such as the Center for the Study of Terrorism and Political Violence (CSTPV), located at St. Andrews University, Scotland, and directed by noted terrorism researcher, Professor Paul Wilkinson and formerly co-directed by Dr. Bruce Hoffman, Rand Corporation. These centers conduct terrorism research and provide portals that cater to the needs of academics, journalists, policymakers, students, and the general public. Terrorism research centers' portals are primarily providing information retrieval and dissemination services except for a few organizations such as the Terrorism Research Center (TRC) and the National Memorial Institute for the Prevention of Terrorism (MIPT) that have expanded their functions to include personalization (TRC) and the Emergency Responders Knowledge Base (MIPT). For example, the TRC, founded in 1996, has the highest number of portal features (31/61) including four terrorism databases and is highly recommended with about 5,000 incoming links [7].

The study by Kennedy and Lunn [21] generated lists of organizations conducting terrorism research and 28 different sources of terrorism databases and archives. Using their list of terrorism organizations and Carnegie Mellon University's [5] portal taxonomies, we analyzed terrorism research centers' portals to identify their features including the types of information and applications provided. Based on the analysis of 54 terrorism portals, the most frequently identified features were information retrieval and dissemination services. The information included full-text documents and archives (74%), links to other terrorism resources (54%), news (39%), and emergency preparedness materials (28%). Applications were limited to e-commerce services for selling terrorism books, reports, and multimedia resources (35%), and terrorism databases (22%) of groups, terrorist incidents, and bio-terrorism agents.

2.3 Information Analysis

Terrorism research centers' portals provide access to a diversity of unstructured (e.g., reports, news stories, transcripts) and structured (database) information but offer limited tools for integrating the resources and supporting information fusion (including post-retrieval analysis). After searching the terrorism portals, the user has to manually browse through the list of retrieved documents to locate relevant resources and establish relationships among the documents.

Automatic indexing algorithms have been used widely to extract key concepts from textual data. It has been shown that automatic indexing is as effective as human indexing [36]. Many proven techniques have been developed such as information extraction (IE). Information extraction techniques such as noun phrasing have been applied to perform indexing for phrases rather than just words [40]. These techniques are useful in extracting meaningful terms from text documents not only for document retrieval but also for further analysis.

There is, today, an increased interest in the use of data and web mining and machine learning techniques which focus on identifying terrorism-related patterns in data. These techniques have been applied to the analysis of news articles (such as in the Message Understanding Conference or MUC), online information sources (e.g., the Columbia University's Newsblaster system), and high-speed data streams that are processed and mined in a Distributed Mining and Monitoring System at Cornell University [31].

DARPA, for its part, is also supporting the development and integration of information fusion technologies such as data mining, biometrics, collaborative and knowledge discovery technologies that identify and display links among people, content, and topics [6] to counter "asymmetric threats" such as those found in terrorist attacks. For example, DARPA and other government agencies have solicited counterterrorism proposals such as the Terrorism Knowledge Base project that was eventually approved for \$9 million and is being developed by Cycorp Incorporated based in Austin, Texas, with the assistance of terrorism domain experts from the Terrorism Research Center [15]. The Terrorism Knowledge Base with its inference engine will support analysts in the intelligence community. However, these advanced technological systems are not available in most terrorism research centers or academic libraries that are used by terrorism researchers in the academic community.

2.4 Social Network Analysis

Existing terrorist network research is still at its incipient stage. Although previous research, including a few empirical ones, have sounded the call for new approaches to terrorist network analysis [4, 26, 39], studies have remained mostly small-scale and used manual analysis of a specific terrorist organization. For instance, Krebs [23] manually collected data from public news releases after the 9/11 attacks and studied the network surrounding the 19 hijackers. Sageman [35] analyzed the Global Salafi Jihad network consisting of 171 members using a manual approach and provided an anecdotal explanation of the formation and evolution of this network. None of these studies used advanced data mining technologies that have been applied widely in

other domains such as finance, marketing, and business to discover previously unknown patterns from terrorist networks. Moreover, few studies have been able to systematically capture the dynamics of terrorist networks and predict terrorism trends. What is needed is a set of integrated methods, technologies, models, and tools to automatically mine data and discover valuable knowledge from terrorist networks based on large volumes of data of high complexity.

2.5 Chatterbot Techniques

The idea behind chatterbot techniques is to create an intimate atmosphere where individuals can converse with a natural language program (a chatterbot) and receive meaningful and immediate responses to their queries related to a certain domain without having to search the Internet for the answers themselves. Most chatterbot techniques rely on pattern matching algorithms which takes inputs from the user, parses and matches the input to one of the questions in their question/answer script, then picks out the appropriate response dictated by the script, and displays it to the user [42]. Examples include: ELIZA [42], Parry [11] and ALICE [16]. Previous studies on chatterbot have shown the potential of using chatterbot to provide people with easy access to domain-specific knowledge. We believe that chatterbot techniques can be used to provide the general public with necessary knowledge of the global terrorism phenomena.

3 Research Questions

Building on our research and system development experiences in portal collection building, text mining, information extraction, criminal network analysis and visualization, we propose to use an advanced knowledge discovery approach to design and develop a set of knowledge portals to address the challenges in the counter-terrorism domain. The research questions postulated in our project are:

1. How can intelligent collection building, searching, text mining, and social network analysis techniques be used to help resolve the information access and management problems in terrorism research and support the knowledge creation and discovery patterns among contemporary terrorism researchers?
2. How can intelligent collection building, searching, text mining with multi-lingual and translation capabilities, and social network analysis techniques be used to collect, analyze, and visualize Web contents created by terrorist groups so as to decipher the social milieu, network dynamics, and communication channels among terrorists?
3. How can the chatterbot technology be used to support citizens' and victims' responding to terrorism and provide them with necessary knowledge of terrorism research?

4 An Advanced Knowledge Discovery Approach to Addressing Terrorism Threats

To help us systematically answer our research questions, we propose a project called Terrorism Knowledge Discovery Project. This project consists of several custom-built portals (testbeds): Terrorism Knowledge Portal (a prototype has already been created), Terrorism Expert Finder, Dark Web (consisting of Internet-based terrorist multilingual resources), Terrorist Network Portal, and Chatterbot Portal which will be organized around the three aspects of challenges associated with terrorism research, terrorists, and terrorized victims or target groups. Figure 1, in the next page, provides a summary of the all the portals.

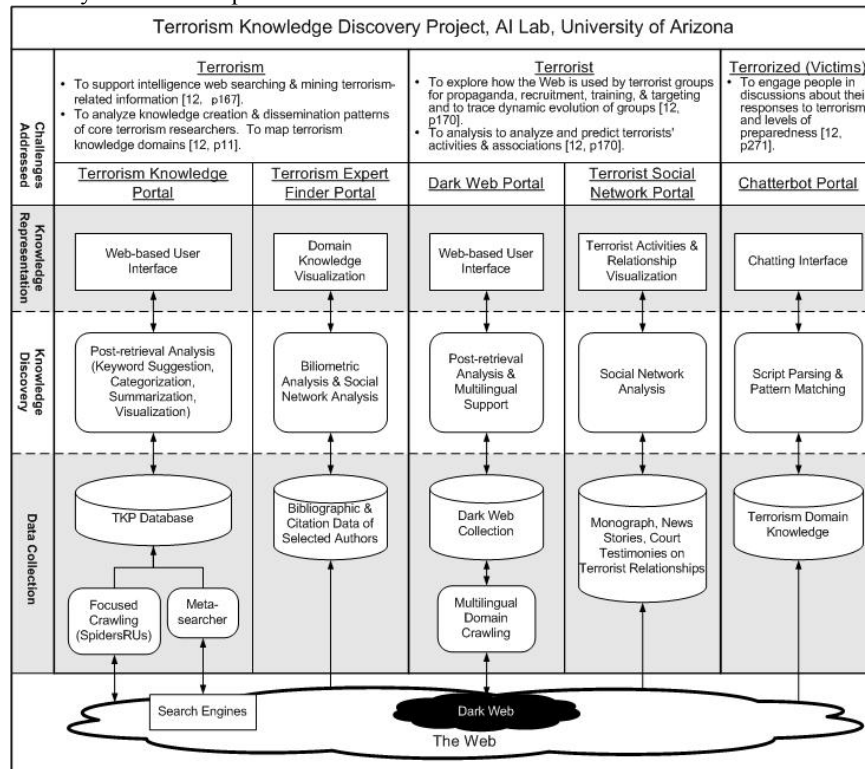


Fig. 1. The Knowledge Portals in the Terrorism Knowledge Discovery Project

In the following sections, we will describe the knowledge portals that address each of the three aspects of challenges in terms of the key components and techniques used in the portals.

4.1 Addressing Challenges Associated with Terrorism Research: Terrorism Knowledge Portal and Terrorism Expert Finder Portal

4.1.1 Terrorism Knowledge Portal

To address information access and management problems in the terrorism research domain, we developed the Terrorism Knowledge Portal, or TKP (<http://ai.bpa.arizona.edu/COPLINK/tkp.html>). The goal is to facilitate the searching and browsing of terrorism information on the Web. Developed based on an integrated knowledge portal approach [10], TKP consists of various components: searching of its own database, meta-searching other terrorism information sources, keyword suggestion, Web page summarization, categorization and visualization. Each sub-component is described below.

TKP Database. TKP supports searching of a customized terrorism research database with more than 360,000 quality pages obtained from automatic spidering various terrorism-related Web sites and multiple search engines. Based on an extensive research, we identified 38 high-quality terrorism Web sites as seed URLs for spidering Web pages. A breadth-first search spidering program called Offline Explorer produced by MetaProducts (<http://www.metaproducts.com/>) was used to automatically download 1 million Web pages using the seeds and after filtering out pages that contain too little text or are irrelevant, we obtained 200,000 Web pages in our local database.

Apart from domain spidering, we used meta-spidering to collect Web pages from 11 major search engines and news Web sites to provide rich terrorism information. From terrorism research publications produced by top-rated researchers (e.g., Paul Wilkinson, Bruce Hoffman and Brian Jenkins), we identified 795 terrorism-related keywords and used them as queries to search the 11 sites. After filtering duplicated pages from different sites, we collected more than 160,000 distinct quality pages. Together with domain spidering, we obtained 360,000 high-quality Web pages in total in our local search index.

TKP Meta-Searchers. In addition to searching its own database, TKP supports meta-searching of various terrorism information sources. This ensures comprehensiveness and recency in covering the domain and reduces information biases. Based on an extensive study on terrorism information sources, we identified four categories of meta-searchers: terrorism databases, research institutes, government Web sites, and news and general search engines. Users can search with one query all the sources selected by them, thus saving their time in accessing different sites. Results are listed according to the meta-searchers' names. Figures 1a and 1b show the search page and result page of TKP.

Keyword Suggestion. To facilitate searching with different keywords, TKP provides keyword suggestion function, developed using the concept space approach [9]. It identified pairs of keywords co-occurring on the same pages and extracts them for use as thesaurus terms in our database. Each query to the TKP concept space thesaurus elicits a ranked list of keywords highly related to the input keyword. In addition, we used Scirus (<http://www.scirus.com>), a major Web site providing keyword suggestion of scientific terms, to retrieve more related terms.

Document Summarization. The TKP summarizer, a modified version of a text summarizer called TXTRACTOR, uses sentence-selection heuristics to rank text segments [28]. These heuristics strive to reduce redundancy of information in a query-based summary. The summarizer can flexibly summarize Web pages using three or five sentence(s). Users can invoke it by choosing the number of sentences for summarization in a pull-down menu under each result. After which, a new window is activated (shown in Figure 2d) that displays the summary on the left and the original Web page on the right.

Document Categorization. The TKP categorizer organizes the search results into various folders labeled by the key phrases appearing in the page summaries or titles (see Figure 2c), thereby facilitating the understanding of different groups of Web pages. We used the Arizona Noun Phraser (AZNP) to extract meaningful phrases from the titles and summaries of the search results. Developed at the Artificial Intelligence Lab of the University of Arizona, AZNP extracts all the noun phrases from each Web page automatically based on part-of-speech tagging and linguistic rules [40]. An indexing program calculates the frequency of occurrence of these phrases and selects the 20 most frequently occurring phrases to index the results. Each folder shown on Figure 2c is labeled by a phrase that appears in the pages categorized under it. As a page may contain more than one indexing phrase, the categorization is non-exclusive.

Document Visualization. The TKP visualizer was developed to reduce information overload when a large number of search results are obtained, which is a typical situation in many search scenarios. It has of two versions: the Jigsaw and Geographic Information Systems (GIS) versions. Web pages are clustered onto a map, generated using the Kohonen [22] self-organizing map (SOM) algorithm, which is a two-layered neural network that automatically learns from the input Web pages and clusters them into different naturally occurring groups. In the jigsaw SOM visualizer, key terms identified by AZNP were used to label map regions, where the sizes correspond to the numbers of pages clustered in them. Clicking on a map region brings up the list of pages on the right side of the pop-up window (see Figure 2f). Users can then open these pages by clicking on the links. In the GIS SOM visualizer, Web pages are shown as points on a two-dimensional map with their positions determined by the SOM algorithm. The map's background shows contour lines representing the varying values selected by users (e.g., frequency of occurrence of query terms in the Web pages) and is independent of the points' positions. Users can navigate on the map by clicking on the buttons and resize a certain part of the map by dragging a rectangle that will highlight the set of Web pages listed on the bottom right side of the pop-up window (see Figure 2e). Using the TKP visualizer, terrorism researchers can obtain a meaningful and comprehensive picture of a large number of search results.



Fig. 2. Major Components of the Terrorism Knowledge Portal

Table 1. List of Core Terrorism Authors Based on Times Cited in Citation Database

Name	Organization	Specialty	Times cited	# pubs.	# years
1. Wilkinson, Paul	St. Andrews Univ., Center for the Study Terrorism & Political Violence (CSTPV), UK	Terrorism, Europe	229	87	31
2. Gurr, Ted	Univ. Maryland	Political violence	214	52	40
3. Laqueur, Walter	Center for Strategic & Intl Studies (CSIS) (formerly)	Political violence, history	191	38	27
4. Alexander, Yonah	Potomac Institute for Policy Studies, SUNY, Center for Strategic & Intl Studies (CSIS) (formerly)	International studies, media	166	94	31
5. Bell, J.B.	Columbia Univ., Institute, War & Peace Studies (formerly)	Terrorist groups	138	49	34
6. Stohl, Michael	Purdue University	Communication	136	30	27
7. Jenkins, Brian	Rand Corporation (formerly), Kroll Associates, CSPTV council member. Founder terrorism program at Rand	Counterterrorism	96	40	29
8. Ronfeldt, David	Rand Corporation	Cyberterrorism	95	20	29
9. Crenshaw, Martha	Wesleyan Univ., CSTPV council member	Government	90	40	34
10. Hoffman, Bruce	Rand Corporation, CSTPV (formerly)	Terrorist groups	81	121	26
11. Arquilla, John	Naval Postgraduate School, Rand Corporation	Cyberterrorism	75	20	29
12. Mickolus, Edward	National Foreign Assessment Center	Terrorist data	73	26	27
13. Wardlaw, Grant	Australian Institute of Criminology	Criminology	49	25	22
14. Hacker, F.J.	USC Medical & Law Schools	Psychology	38	3	5
15. Rapoport, David	UCLA, Terrorism & Political Violence (founding editor)	Political violence	37	33	33
16. Bassiouni, M.C.	DePaul Univ.	Law	30	8	16
17. Kepel, Gilles	Institut d'Etudes Politiques, Paris	Religion & politics	25	6	4
18. Kupperman, Robert	Kupperman & Associates, Center for Strategic & Intl Studies (CSIS) (formerly)	Crisis management	25	20	19
19. Stern, Jessica	Harvard's Kennedy Institute, Council of Foreign Relations	Public policy	25	24	13

4.1.2 Terrorism Expert Finder Portal

Because of the information overload problem and the increasing number of dispersed researcher groups studying terrorism, we will build a collection containing data on core terrorism authors, their bibliographic citations, and their cited references in the Terrorism Expert Finder Portal. The portal will provide a much-needed mechanism for identifying, mapping, collaborating, and stimulating a convergence in thinking about research challenges such as terrorism definitions and theoretical models. Silke [38] described contemporary terrorism works as being characterized by a marked absence of conceptual agreement and a wide diversity of views on even basic issues. This portal is important because many researchers and students waste a lot of time since they often do not know where to start their terrorism research, who are the current experts, what are their empirical studies, and insights on specific issues. It provides a mechanism to map the terrorism knowledge domains and answer several research questions:

- Who are the core terrorism authors?
- What are their influential terrorism publications?
- What are the dominant knowledge creation, dissemination and communications processes used in terrorism research?

The portal is an extension of Reid's [32, 33, 34] research on identifying the invisible college of terrorism researchers and diffusion of their ideas. Reid, a former terrorism specialist at the Central Intelligence Agency, used bibliometric, citation and content analysis to identify and analyze core terrorism experts and their terrorism publications.

Of the 44 core terrorism authors identified, the top 19 authors are enumerated in Table 1 together with pertinent citation counts, number of terrorism-related publications, and number of years they authored terrorism publications. To calculate the number of years that an author has been writing terrorism publications (based on our database), his latest publication year is subtracted from his earliest publication year.

The core terrorism authors' data generated 1,324 unique titles that covered about 35 years of research on contemporary terrorism. The data were parsed into many fields including author, title, journal, publication type, publication year, and the number of citations per title. The 44 core authors and their 250 coauthors published in 148 different journals, the largest number of which was published in *Terrorism and Political Violence* (141 articles).

We also conducted subject searches in ISI Web of Science to retrieve bibliographic and citation data on the topic: terrorism. About 7,590 records were retrieved and will be used to analyze citation, co-authorship, and social network patterns. The records were parsed into several fields such as author, cited author, cited source, and publication date of cited reference.

The data will be used for identifying, analyzing, summarizing, and visualizing core researchers' knowledge creation patterns. Table 2 presents methodologies that can be applied to the data. It is a revision of Boyack's [2] research that demonstrated the use of different techniques for analyzing 20 years of papers published in the *Proceedings of the National Academy of Sciences* (PNAS).

Table 2. Summary of Methodologies for Analyzing the Terrorism Knowledge Domains

Unit of Analysis	Methodology	Research question associated with:
Authors	Citation analysis, Co-authorship analysis, Content analysis, Factor analysis, Multidimensional scaling, Social network analysis	Communities of practices, Intellectual structure & history of terrorism, Levels & types of collaboration Pathfinder network for visualization Social structure of terrorism
Documents	Co-citation analysis, Content analysis, Various clustering methods	Development of paradigms
Journals	Co-citation analysis, Bradford distribution	Diffusion between fields, Sociology of science
Words	Content analysis, Relational extraction, Semantic analysis	Cognitive structure of terrorism Convergence in ideas Topical classification
Indicators such as economic activity level indicators (counts of papers, patents, & citations)	Combination of methodologies	Impact of scientific outputs & funding, Impact of scientific outputs on policy

4.2 Addressing Challenges Associated with Terrorist: Dark Web Portal and Terrorist Social Network Portal

4.2.1 Dark Web Portal Portal

In addition to analyzing terrorism research, we will also focus on the information access and methodological problems in analyzing terrorist groups. We propose to develop a Dark Web Portal which helps researchers locate, collect, and analyze Dark Web data. By “Dark Web”, we mean the alternate side of the Web which is used by terrorist and extremist groups to spread their ideas.

Dark Web Testbed. The goal of this part is to build a high-quality, up-to-date Dark Web testbed which contains multilingual information created by major terrorist groups in the world. We started the process by identifying the groups that are considered by reliable sources as terrorist groups. The main sources we used to identify US domestic terrorist groups include: Anti-Defamation League (ADL), FBI, Southern Poverty Law Center (SPLC), Militia Watchdog, and Google Directory. To identify international terrorist groups, we relied on the following sources: Counter-Terrorism

Committee (CTC) of the UN Security Council, US State Department report, Official Journal of the European Union, and government reports from the United Kingdom, Australia, Canada, Japan, and P. R. China. These sources have been identified following the recommendations of core terrorism authors.

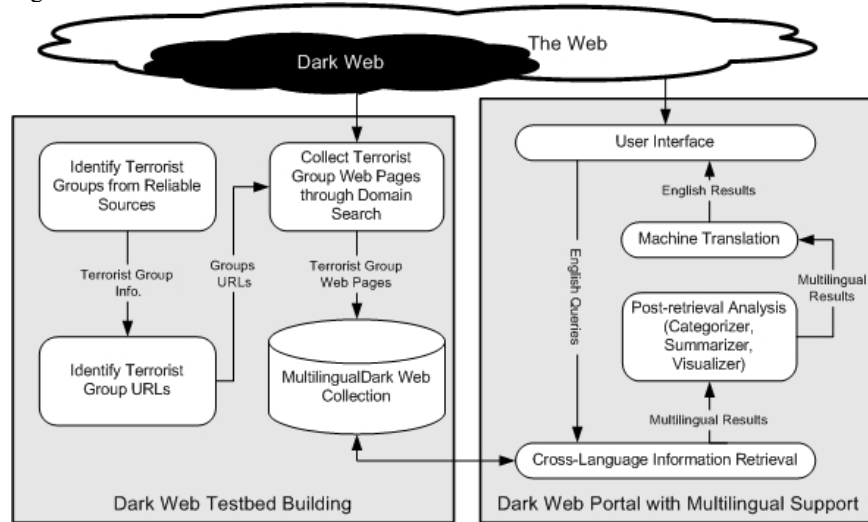


Fig. 3. The Components of the Dark Web Portal Project

We manually identified the URLs of the web sites from the reports alluded to by the sources mentioned above then searched the web using the group names as queries. A total of 94 US domestic terrorist groups and 440 international terrorist groups have been identified. To ensure that our testbed covers all the major regions in the world, we sought the assistance of language experts in English, Arabic, Spanish, Japanese, and Chinese to help us collect URLs in different regions. After the URL of a group is identified, we used the SpidersRUs toolkit, a multilingual Digital Library building tool developed by our own group, to collect all the Web pages under that URL. So far, we have collected 500,000 Web pages created by 94 US domestic groups, 300,000 Web pages created by 41 Arabic-speaking groups, and 100,000 Web pages created by Spanish-speaking groups. The process of building the Dark Web testbed is ongoing.

Intelligent Portal with Multilingual Support. The purpose of the Dark Web Portal is to help terrorism researchers easily access, analyze, and understand the multilingual information in the Dark Web testbed. To address the information overload problem, the Dark Web Portal will be fitted with post-retrieval components (categorizer, summarizer, and visualizer) similar to those of the Terrorism Knowledge Portal. However, without addressing the language barrier problem, researchers are limited to the data in their native languages and cannot fully utilize the multilingual information in the testbed. To address this problem, we plan to add a Cross-language Information Retrieval (CLIR) component into the portal. CLIR is the study of retrieval information in one language through queries expressed in another language. Based on our previous research, we have developed a dictionary-based CLIR system for use in the

Dark Web Portal. It currently takes English queries and retrieves documents in English, Spanish, Chinese, Japanese and Arabic. Another component that we will add to the Dark Web portal is a Machine Translation (MT) component, which will translate the multilingual information retrieved by the CLIR component back into the experts' native languages.

4.2.2 Terrorist Network Portal

Besides collecting terrorist data from the Web, we will also use an authoritative terrorism monograph and its articles as input data for our Terrorist Network Portal. The monograph, entitled "*Understanding Terror Network*", which analyzed the characteristics of 171 terrorists, was written by Dr. Marc Sageman, a forensic psychiatrist and expert on Al-Qaeda. Since terrorists are often unavailable for interviews, Crenshaw [13] recommended using case studies, diaries, and group histories as preferred sources of data for psychological analysis. By using the full-text version of Sageman's monograph and selective articles, we will have testbed data to design and analyze our terrorist network methodology for identifying and visualizing terrorist social networks.

Terrorists are not atomized individuals but actors linked to each other through complex networks of direct or mediated exchanges [26, 35]. Our terrorist network analysis research focuses on discovering and presenting previously unknown patterns in terrorist networks, which involve people, resources, and information tied together by family, religious, personal, financial, operational, and many other relationships. Specifically, we propose to analyze network members in terms of their personal, social, educational background. So, we want to identify how relationships among terrorists are formed and dissolved. To achieve this, we need to identify the relations' types, characteristics, and strengths, operational groups, their assigned tasks, and planned activities. This will enable us to discover the structure and organization of terrorist networks, capture network dynamics, and predict trends of terrorist attacks. We will employ various visualization techniques to intuitively present these patterns to facilitate the interpretation, comprehension, and application of the results.

We will design our Terrorist Network Analyzer and Visualizer (TNAV) using advanced social network analysis methodology. TNAV will provide statistical analysis, cluster analysis, social network analysis (SNA), and visualization. SNA has recently been recognized as a promising technology for studying criminal and terrorist networks [26, 39], which we plan to use in all five levels of analysis.

A network can be treated as a graph in which nodes represent individuals and links represent relations between the individuals. We propose to analyze terrorist networks at five different levels, namely, node, link, group, overall structure, and dynamics. For instance, Sageman partitioned the terrorist network, Global Salafi Jihad, into four groups: Central Staff, Maghreb Arabs, Core Arabs, and Indonesians. In each group, there are a hub and several gatekeepers. Figure 4 shows Osama bin Laden as the hub of the Central Staff cluster and issues commands to the whole network through his gatekeepers.

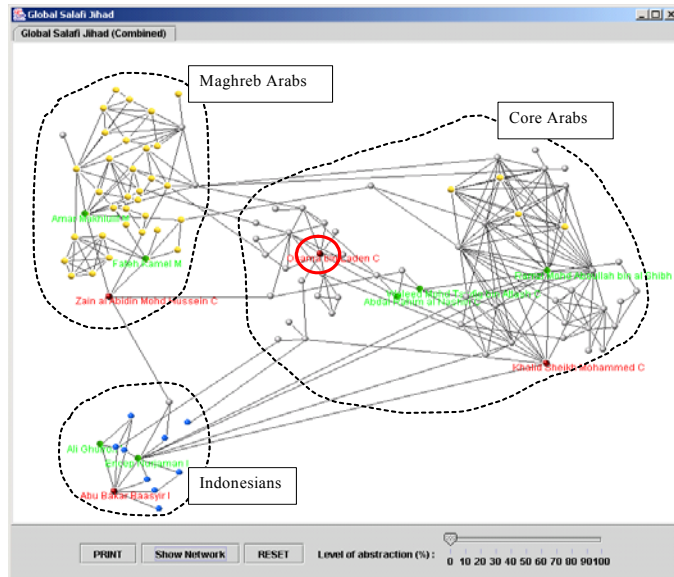


Fig. 4. Global Safafi Jihad Partition into Four Groups (red circle is Osama bin Laden)

Figure 5 shows the system architecture of the TNAV. It consists of five analyzers corresponding to the five levels of analysis. The analyzers include:

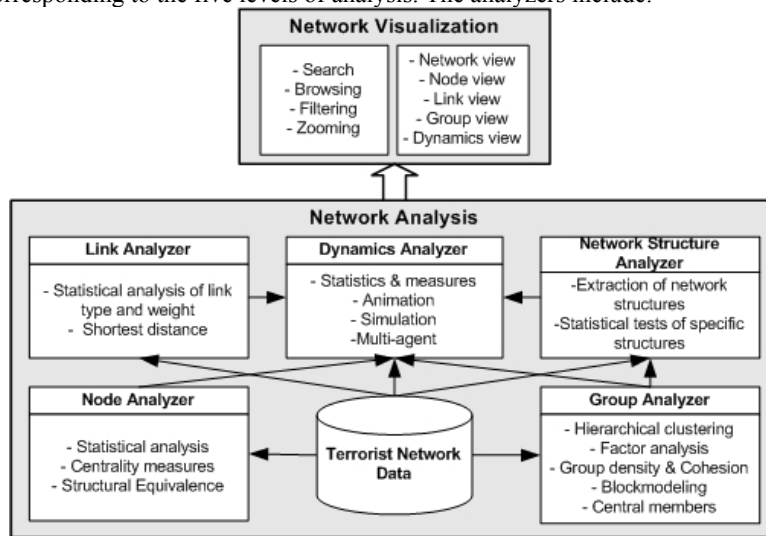


Fig. 5. System Architecture of the Terrorist Network Analyzer and Visualizer (TNAV)

Node Analyzer. The basic statistical analysis function in this analyzer summarizes the characteristics of terrorists in terms of their demographics, background, socio-

economic status, etc. In addition, we plan to use the centrality [19] and structural equivalence measures [25] from SNA to measure the importance of each member. Centralities such as degree, prestige, betweenness, and closeness can automatically identify the leaders, gatekeepers, and outliers from a network [1, 39].

Link Analyzer. In addition to the statistical summary of link type and weight distribution, our link analyzer will measure the ease or difficulty for one member of the network to connect to other members. We plan to use traditional single source-destination [17] and all-pair shortest-path algorithms [18] to identify this property of a terrorist network.

Group Analyzer. We plan to employ two methods to find groups in a terrorist network: hierarchical clustering [20, 30] and factor analysis. We will test which method is more likely to generate meaningful and valid groupings that match the actual organization of a terrorist network. In addition, with a rich set of approaches and metrics from SNA, we will extract interaction patterns between groups using blockmodeling approach [43], analyze and predict the stability of a terrorist group using group density and cohesion [41] measures, and detect the central members from a specific group.

Network Structure Analyzer. After the groups have been identified and their interaction patterns extracted, we will establish whether a terrorist network takes a specific structure (centralized or decentralized) and its degree of hierarchy. The structure found can help reveal the vulnerability of a terrorist organization.

Dynamics Analyzer. Adding a time dimension to all the statistics and measures in the previous levels, our dynamics analyzer will extract the patterns of change in terms of individual member's importance and operational roles; group(s) that are merging or splitting; and recruitment and turnover. We will also use animation to intuitively show how a network changes over time. More importantly, we hope to be able to predict the operations of a terrorist network based on the patterns extracted from historical data and use hidden Markov models, simulations, and multi-agent technology to foresee how a network reacts to changes such as removal of central members, disruption of certain groups, etc.

GUI and Visualizer. We propose to use several network layout algorithms such as multidimensional scaling (MDS) [3], pathfinder scaling [8], and force-directed spring algorithms to portray a terrorist network. Node view, link view, and group view can show the details of individual terrorists, their relations, and their group memberships. The dynamics view will present animated visualization of a network or a specific group over several periods of time. Statistical summary and test results will be presented in traditional formats such as charts and tables. The GUI will take user input such as a query for a seed terrorist and present the analysis result about the part of network surrounding the terrorist.

4.3 Addressing Challenges Associated with Terrorized Victims: Chatterbot Portal

In addition to supporting the analysis of terrorist groups, the Terrorism Knowledge Discovery Project will provide a testbed for victims and citizens who are trying to

respond to terrorism. To help in the dissemination of information to the victims of terrorism, we will create a computer-driven natural language chatterbot that will respond to queries about the terrorism domain and provide real-time data on terrorism activities, prevention, and preparation strategies. The goal of this testbed is to create an intimate atmosphere where individuals can converse with the chatterbot and receive meaningful and immediate responses to their terrorism queries.

In our exploration of the chatterbot technology, we discovered that ALICE chatterbots could handle more domain-specific knowledge topics better than the general conversational chatterbots. After concluding tests using telecommunications domain knowledge, we decided that a chatterbot testbed would be appropriate for supporting general conversation and information exchange about terrorism.

The major thrust of the Chatterbot Portal segment is the development of the Terrorism Activity Resource Application (TARA). TARA uses a modified version of the ALICE engine to make the programming and maintenance of the system easier as well as providing some value-added features. The development phases of TARA are as the follows:

First, TARA will be benchmarked against search systems, such as Google, to explore the effectiveness and efficiency of using a natural language dialog system to return terrorism-related answers in a friendlier manner. This approach will eliminate some of the ‘hunting and pecking’ skills necessary for filtering out search engine information. Using the “official” websites that experts on terrorism deem to be trustworthy will ensure the reliability of TARA’s responses.

Second, TARA will be built as a ‘friend and a companion’ to listen to users who have been affected by terrorism activities. We plan to define the knowledge in the Terrorism domain as static (e.g., ‘Who is bin Laden?’ or ‘What should I do in case of a gas attack?’), and dynamic (e.g., ‘What is the current Terror Threat Level?’ or ‘What has Hezbollah done recently?’). Because of TARA’s genealogical relationship to ALICE, TARA possesses a natural language interface and all of the conversational dialogue components inherent in ALICE. We believe that these elements will be a vital part of TARA’s flexibility in responding to queries that may not directly relate to Terrorism.

The last piece we envision applying is the use of multimedia content in conjunction with the conversational dialogue. We plan on using pictures, audio, and video feeds to enhance the knowledge delivery of the system.

5 Concluding Remarks

Information overload, uncertain data quality, and lack of access to integrated datasets and advanced methodologies for studying terrorism are major hurdles and challenges which both traditional and new counter-terrorism researchers have to overcome. Understanding the ramifications of each component hurdle and how existing information technology approaches can be used to address such hurdle is the over-arching objective of this project. In this report, we exhaustively analyzed component issues confronting contemporary approaches counter terrorism efforts and lined-up potent technology solutions to help fill in existing gaps. Taking the key suggestions from the

report “*Making the Nation Safer*”, we are developing a Terrorism Knowledge Discovery Project that provides advanced methodologies for analyzing terrorism research, terrorist and their social networks, and the terrorized groups (victims and citizens). The project consists of custom-built multilingual portals (testbeds): Terrorism Knowledge Portal, Terrorism Expert Finder, Dark Web (Web resources used by terrorists), Terrorist Network Portal, and Terrorism Activity Resource Application chat-terbot.

Although the Terrorism Knowledge Discovery Project is aimed at supporting the counter-terrorism researcher community, we have reason to believe that the law enforcement, intelligence, and security communities can also benefit from using it. It is a critical first step in demonstrating the feasibility of developing a national terrorism research infrastructure. Most importantly, we believe that it will help facilitate knowledge creation and discovery processes because it will provide a seamlessly integrated environment of functional modules that fully support the collaborative efforts of various, culturally diverse terrorism research teams. Lastly, it can become the learning resource and tool that the general community can use to heighten their awareness of the global terrorism phenomenon, to learn how best they can respond to terrorism and, eventually, garner significant grass root support for the government’s efforts to keep America safe.

6 Acknowledgements

This research has been supported in part by the following grants:

- DHS/CNRI, “BorderSafe Initiative,” October 2003-September 2004.
- NSF/ITR, “COPLINK Center for Intelligence and Security Informatics – A Crime Data Mining Approach to Developing Border Safe Research,” EIA-0326348, September 2003-August 2005.

We would like to thank the officers and domain experts of Tucson Police Department, Arizona Department of Customs and Border Protection, and San Diego ARJIS for their support throughout the project. We would also like to thank all members of the Artificial Intelligence Lab at the University of Arizona who have contributed to the project, in particular Homa Atabakhsh, Cathy Larson, Chun-Ju Tseng, Ying Liu, Wei Xi, Charles Zhi-Kai Chen, Guanpi Lai, and Shing Ka Wu.

References

1. Baker, W. E., Faulkner, R. R.: Social Organization of Conspiracy: Illegal Networks in the Heavy Electrical Equipment Industry. *American Sociological Review* 58(12) (1993) 837-860
2. Boyack, K. W.: Mapping Knowledge Domains: Characterizing PNAS. *PNAS Early Edition* (2003)
3. Breiger, R. L., Boorman, S., Arabie, P.: An Algorithm for Clustering Relational Data, with Applications to Social Network Analysis and Comparison with Multidimensional Scaling. *Journal of Mathematical Psychology* 12 (1975) 328-383

4. Carley, K. M., Lee, J. S.: Destabilizing Networks. *Connections* 24(3) (2001) 31-34
5. Carnegie Mellon University: Web Portal and Portal Taxonomies. Final Report of Consequence Management Program Integration Office (2001)
6. Caterinicca, D.: Data Mining Aims at National Security. *Federal Computer Week* (2002)
7. Center, T. R.: About the Terrorism Research Center (2003)
8. Chen, C., Paul, R. J., O'Keefe, B.: Fitting the jigsaw of Citation: Information Visualization in Domain Analysis. *Journal of American Society of Information Science and Technology* 52(4) (2001) 315-330
9. Chen, H., Lynch, K. J.: Automatic Construction of Networks of Concepts Characterizing Document Databases. *IEEE Transactions on Systems, Man, and Cybernetics* 22 (5) (1992) 885-902
10. Chung, W., Zhang, Y., Huang, Z., Wang, G., Ong, T. H., Chen, H.: Internet Searching and Browsing in a Multilingual World: An Experiment on the Chinese Business Intelligence Portal (CBizPort). *Journal of the American Society for Information and Science and Technology*, Accepted for publication (forthcoming)
11. Colby, K. M. W., Sylvia, H., Dennis, F.: Artificial Paranoia. *Artificial Intelligence* 2 (1971) 1-25
12. Council, National Research: Making the Nation Safer: the Role of Science and Technology in Countering Terrorism. Washington, D.C. (2002) 339
13. Crenshaw, M.: Psychology of Terrorism: An Agenda for the 21st Century. *Political Psychology* 21(2) (2000) 405-420
14. Cutter, S., T. J. Wilbank, (ed.): *Geographical Dimensions of Terrorism*, Taylor & Francis, Inc. (2003)
15. DARPA: Terrorism Knowledge Base Proposal. (2002)
16. De Angeli, A. J., Graham, I., Coventry, L.: The Unfriendly User: Exploring Social Reactions to Chatterbots. *Proceedings of The International Conference on Affective Human Factors Design*, London, Asean Academic Press. Han, S. K., YounGi (2001)
17. Dijkstra, E.: A Note on Two Problems in Connection with Graphs. *Numerische Mathematik* 1 (1959) 269-271
18. Floyd, R. W.: Algorithm 97: Shortest Path. *Communications of the ACM* 5(6) (1962) 345-370
19. Freeman, L. C.: Centrality in Social Networks: Conceptual Clarification. *Social Networks* 1 (1979) 215-240
20. Jain, A. K., Dubes, R. C.: *Algorithms for Clustering Data*. Upper Saddle River, NJ, Prentice-Hall (1988)
21. Kennedy, L. W., Lunn, C. M.: Developing a Foundation for Policy Relevant Terrorism Research in Criminology (2003)
22. Kohonen, T.: *Self-organizing maps*. Springer-Verlag, Berlin (1995)
23. Krebs, V. E.: Mapping networks of terrorist cells. *Connections* 24(3) (2001) 43-52
24. Lawrence, S., Giles, C. L.: Accessibility of Information on the Web. *Nature* 400 (1999) 107-109
25. Lorrain, F. P., White, H. C.: Structural Equivalence of Individuals in Social Networks. *Journal of Mathematical Sociology* 1 (1971) 49-80
26. McAndrew, D.: Structural Analysis of Criminal Networks. *Social Psychology of Crime: Groups, Teams, and Networks*, Offender Profiling Series, III. L. Allison. Dartmouth, Aldershot (1999)
27. McBryan, O.: GENVL and WWW: Tools for Taming the Web. *Proceedings of the First International Conference on the World Wide Web*, Geneva, Switzerland (1994)
28. McDonald, D., Chen, H.: Using Sentence-selection Heuristics to Rank Text Segments in TXTRACTOR. *Proceedings of Second ACM/IEEE-CS joint conference on Digital libraries*, Portland, Oregon, USA (2002)

29. Moore, R. G., Graham: Emile: Using a Chatbot Conversation to Enhance the Learning of Social Theory. Huddersfield, England, Univ. of Huddersfield (2002)
30. Murtagh, F.: A Survey of Recent Advances in Hierarchical Clustering Algorithms Which Use Cluster Centers. *Computer Journal* 26 (1984) 354-359
31. National Science Foundation: Data Mining and Homeland Security Applications. (2003)
32. Reid, E. O. F.: An Analysis of Terrorism Literature: A Bibliometric and Content Analysis Study. School of Library and Information Management. Los Angeles, University of Southern California (1983) 357
33. Reid, E. O. F.: Using Online Databases to Analyze the Development of a Specialty: Case Study of Terrorism. 13th National Online Meeting Proceedings, New York, NY, Learning Information (1992)
34. Reid, E. O. F.: Evolution of a Body of Knowledge: an Analysis of Terrorism Research. *Information Processing & Management* 33(1) (1997) 91-106
35. Sageman, M.: Understanding Terror Networks. Pennsylvania, University of Pennsylvania Press (2004)
36. Salton, G.: Recent Trends in Automatic Information Retrieval. Proceedings of the 9th Annual International ACM SIGIR, Pisa, Italy (1986)
37. Selberg, E., Etzioni, O.: Multi-service Search and Comparison Using the MetaCrawler. Proceedings of the 4th International World-Wide Web Conference (1995)
38. Silke, A.: Devil You Know: Continuing Problems with Research on Terrorism. *Terrorism and Political Violence* 13(4) (2001) 1-14
39. Sparrow, M. K.: Application of Network Analysis to Criminal Intelligence: An Assessment of the Prospects. *Social Networks* 13 (1991) 251-274
40. Tolle, K. M., Chen, H.: Comparing noun phrasing techniques for use with medical digital library tools. *Journal of the American Society for Information Science (Special Issue on Digital Libraries)* 51 (4) (2000) 352-370
41. Wasserman, S., Faust, K.: *Social Network Analysis: Methods and Applications*. Cambridge, Cambridge University Press (1994)
42. Weizenbaum, J.: Eliza - A Computer Program for the Study of Natural Language Communication between Man and Machine. *Communications of the ACM* 9(1) (1966) 36-45
43. White, H. C., Boorman, S. A. Breiger, R. L.: Social Structure from Multiple Networks: I. Blockmodels of Roles and Positions. *American Journal of Sociology* 81 (1976) 730-780
44. Vrajitoru, D.: Evolutionary Sentence Building for Chatterbots. Genetic and Evolutionary Computation Conference (GECCO), Chicago, IL (2003)