Voss, Susan and Joslyn, Cliff: (2002) "Advanced Knowledge Integration in Assessing Terrorist Threats", Los Alamos Technical Report LAUR 02-7867

Advanced Knowledge Integration In Assessing Terrorist Threats

S. Voss and C. Joslyn

Overview

This is a report on the FY 02 Reserve LDRD project "Advanced Knowledge Integration in Assessing Terrorist Threats".

Our goal was to apply advanced knowledge integration tools to large data sets of terrorist-related information as a technical collaboration between NIS-8 and CCS-3. Significant progress was made during this past year, and the tools provided new ways of finding direct and indirect relationships between people, groups, terrorist events, and areas of expertise. Additional work is required before these tools can used by individual analysts, but the work completed within this LDRD shows the strength of using knowledge tools on diverse data sets.

The team consisted of: S. Voss (PI), A. Cernicek, K. Gardner, A. Singleton and A. Christensen of NIS-8; C. Joslyn (co-PI), L. Rocha, G. Papcun, S. Mniszewski, and J. Hogden of CCS-3; E. Gavrilov of CCS-1; and J. Gattiker of NIS-7. Thus the LDRD was inherently inter-disciplinary, bringing computer and information scientists, mathematicians, content analysts and specialists, and database and software engineers from the two Divisions; to work with databases from multiple sources; and software tools implementing advanced data mining techniques both developed internally and integrated from external sources.

First we describe the work in constructing the source database in the USentry system, and the results of the data collection and correlation in the system. NIS-8 analysts collected over a 1000 open source reports over a six-month period correlating the data by individual people, terrorist groups, events and expertise. The USentry information system developed by NIS-8 was modified and used. Several key findings were evident using USentry.

We then describe the four methodologies deployed against this database by CCS-3:

- Development and deployment of proximity and semi-metric network analyses showing the direct and indirect relationships among the different entities. These enabled the analyst to determine the probability of co-occurrence and (using the inverse of the proximity measures) to obtain possible associations not directly visible. Based upon the results we were able to determine if some of the predicted relationships were true.
- Development of a data extraction tool based on semantic feature analysis to provide a means of quickly finding patterns within the text and displaying it in a comprehensive manner.
- Deployment of the DataDelver link analytical tool (developed previously) which uses multidimensional database and information theoretical techniques to aid users in guided knowledge discovery in complex databases.
- Deployment of Formal Concept Analysis (FCA) tools to show the hierarchical relations present among collections of data values in the database. These allowed the realization of patterns and relationships among multiple variables not previously detectable.

The results of this LDRD can applied directly to the difficult task of Homeland Defense.

The Database and the Role of the Analyst

A significant amount of open source information was collected on terrorist groups and individual terrorists by A. Singleton and A. Christensen to establish a base of data that could be used with the advanced data mining tools developed by CCS-3. The information was collected and correlated in the USentry tool developed by NIS-8. USentry is a collaborative tool designed to allow the analyst to collect source information and break it into separate entities. For the LDRD, the information was correlated into terrorist groups, individual terrorists and events. The correlated information was provided to CCS-3 as input to their models.

One of the problems encountered by the team was the large number of names and alias used by individual terrorists. The team expanded the collection of information on each individual terrorist within the Sentry system to include unique identifiers to help in tracking information on each person. Additional data was collected on each terrorist including their terrorist physical characteristics, their primary associates and what events they may have been associated with. Examples of events include the September 11th, 2001 attack on the World Trade Center or the Khobar Towers Bombing on June 25, 1996. Currently we have developed a baseline of 40 events, over 350 individuals associated with terrorism and over 50 terrorist-related groups.

Two sets of data were created for the LDRD: first a set of newspaper and web-site reports on the terrorists, events and general evaluation of the information on the 9/11 terrorist attacks; and the second set on individual terrorists, their affiliations, roles and inter-relationships. The information was compiled and linked within the USentry knowledge-base tool developed at Los Alamos National Laboratory. USentry allows the analyst to gather information and search for patterns and anomalies.

The document data set was compiled from a number of different open sources. The database has 996 documents. The documents are from the Foreign Broadcast Information System (FBIS), a system limited to government agencies and their contractors, and newspaper, magazine, and journal articles, which were obtained primarily through reviews of FBIS profiles, through information sent by other specialists and through profile searches on specific web sites.

As part of the study we established a set of terrorists and terrorist-related people profiles. This "people" data set focused on three primary groups of individuals: the 9/11 hijackers; those featured in the FBI's most wanted list; and five important people in the Al-Qaeda network. Five important people within the Al-Qaeda list were chosen because of their connection to bin Laden and their leadership roles in the many Al-Qaeda cells. The datasets were chosen to try out new ways of using data mining tools to determine unique relationships between the three sets of individuals that were not obvious.

The nineteen hijackers, the associated flight number and the destination are listed in the Table 1 below.

Table 1:

Hijacker	Flight number and Destination
Mohamed Atta	AA Flight 11 7:45 a.m. departed Boston for Los Angeles, 8:45 a.m. crashed into North Tower of WTC
Abduloziz Alomari	AA Flight 11 7:45 a.m. departed Boston for Los Angeles, 8:45 a.m. crashed into North Tower of WTC
Wail M. Alshehri	AA Flight 11 7:45 a.m. departed Boston for Los Angeles, 8:45 a.m. crashed into North Tower of WTC
Waleed M. Alshehri	AA Flight 11 7:45 a.m. departed Boston for Los Angeles, 8:45 a.m. crashed into North Tower of WTC
Satam Al Suqami	AA Flight 11 7:45 a.m. departed Boston for Los Angeles, 8:45 a.m. crashed into North Tower of WTC

Marwan Al-Shehhi	UA Flight 175 7:58 a.m. departed Boston for Los Angeles, 9:05 a.m. crashed into South Tower of WTC
Fayez Banihammad	UA Flight 175 7:58 a.m. departed Boston for Los Angeles, 9:05 a.m. crashed into South Tower of WTC
Ahmed Alghamdi	UA Flight 175 7:58 a.m. departed Boston for Los Angeles, 9:05 a.m. crashed into South Tower of WTC
Hamaza Alghamdi	UA Flight 175 7:58 a.m. departed Boston for Los Angeles, 9:05 a.m. crashed into South Tower of WTC
Mohand Alshehri	UA Flight 175 7:58 a.m. departed Boston for Los Angeles, 9:05 a.m. crashed into South Tower of WTC
Saeed Alghamdi	UA Flight 93 8:01 a.m. departed Newark for San Francisco, 10:10 a.m. crashed into Stony Creek Township, PA
Ahmed Al Haznawi	UA Flight 93 8:01 a.m. departed Newark for San Francisco, 10:10 a.m. crashed into Stony Creek Township, PA
Ahmed Alnami	UA Flight 93 8:01 a.m. departed Newark for San Francisco, 10:10 a.m. crashed into Stony Creek Township, PA
Ziad Samir Jarrah	UA Flight 93 8:01 a.m. departed Newark for San Francisco, 10:10 a.m. crashed into Stony Creek Township, PA
Khalid Almidhar	AA Flight 77 8:10 a.m. departed Washington Dulles for Los Angeles, 9:32 a.m. crashed into the Pentagon
Majed Moqed	AA Flight 77 8:10 a.m. departed Washington Dulles for Los Angeles, 9:32 a.m. crashed into the Pentagon
Nawaf Alhazmi	AA Flight 77 8:10 a.m. departed Washington Dulles for Los Angeles, 9:32 a.m. crashed into the Pentagon
Salem Alhazmi	AA Flight 77 8:10 a.m. departed Washington Dulles for Los Angeles, 9:32 a.m. crashed into the Pentagon
Hani Hanjour	AA Flight 77 8:10 a.m. departed Washington Dulles for Los Angeles, 9:32 a.m. crashed into the Pentagon

One of the purposes of the LDRD was to determine if unique connections and relationships could be identified using open source data. To this end, the nineteen hijackers remain critical in determining the relationships between the hijackers and the others in the Al-Qaeda network.

The FBI's most wanted were chosen because they represent such a broad range of Al-Qaeda members and their actions, particularly pre-9/11. They provided interesting trails to follow as information on their relationships became available in the open press. The following individuals selected for the LDRD analysis are listed in Table 2 below along with their affiliation and role. Each of the individuals listed are wanted for the US Embassy bombings in Kenya and Tanzania on August 7, 1998.

The following individuals are wanted for the US Embassy bombings in Kenya and Tanzania on August 7, 1998.

Table 2:

Name	Affiliation and Role
Osama bin Laden	Al-Qaeda, head of the terrorist network
Ayman Al-Zawahiri	Al-Qaeda advisor and physician to ObL, Shura Council, founder of Egyptian Islamic Jihad, spiritual leader of al-Gama'a al-Islamiya
Abdullah Ahmed Abdullah	Al-Qaeda, Shura Council, ran Al-Qaeda training camps
Muhsin Musa Matwalli Atwah	Al-Qaeda, explosives expert
Anas al-Liby	Al-Qaeda Lt., Shura Council, Libyan Islamic Fighting Group, computer expert
Ahmed Khalfan Ghailani	Al-Qaeda
Ahmed Mohammed Hamed Ali	Al-Qaeda, agriculture expert
Fazul Abdullah Mohammed	Al-Qaeda, computer expert
Mustafa Mohamed Fadhil	Al-Qaeda
Sheikh Ahmed Salim Swedan	Al-Qaeda,
Fahid Mohammed Ally Msalam	Al-Qaeda
Muhammad Atef	Al-Qaeda cofounder, Shura Council, al-Jihad, military strategist, advisor, head of terrorist operations
Saif Al-Adel	Al-Qaeda security chief, Egyptian Islamic Jihad

Ali Saed Bin Ali El-Hoorie; Ahmad Ibrahim Al-Mughassil; Ibrahim Salih Mohammed Al-Yacoub; Abdelkarim Hussein Mohamed Al-Nasser, are wanted for the Khobar towers military complex bombing on June 25, 1996:

Ali Saed Bin Ali El-Hoorie	Al-Qaeda
Ahmad Ibrahim Al-Mughassil	Al-Qaeda
Ibrahim Salih Mohammed Al-Yacoub	Al-Qaeda
Abdelkarim Hussein Mohamed Al-Nasser	Al-Qaeda

Imad Fayez Mugniyah; Ali Atwa; Hasan Izz Al-Din, are wanted for the June 14, 1985 hijacking of a commercial airliner. Abdul Rahman Yasin is wanted for the bombing of the World Trade Center on February 26, 1993. Khalis Shaikh Mohammad is wanted for involvement in a plot to bomb US commercial airliners flying to the US from Southeast Asia in January of 1995:

Imad Fayez Mugniyah	Hizbollah, possible Al-Qaeda, head of Hizbollah security
Ali Atwa	Hizbollah, possible Al-Qaeda
Hasan Izz Al-Din	Hizbollah, possible Al-Qaeda
Abdul Rahman Yasin	Al-Qaeda
Khalis Shaikh Mohammad	Al-Qaeda

Five of the most important people in the Al-Qaeda network are: Zayn-al-Ibidin Abu Zubaydah, the new leader of Al-Qaeda operations; Sami Essid Ben Khemais, leader of the Milan, Italy cell; Imad Eddin Barakat, head of the Spain cell; Ahmed Ressam, believed to be the head of the Montreal, Canada cell; Haydar Abu Doha, the alleged leader and paymaster of the London, England cell.

Data Collection and Correlation

Overview of the Available Information

The type of information found initially dealt primarily with Osama bin Laden and Al-Qaeda, what they stood for and what their capabilities were. The articles speculated greatly about what Osama bin Laden and Al-Qaeda were going to do next. The majority of news articles focused on Osama bin Laden's actions and that of Al-Qaeda. Finding articles that provided the names of the people involved also proved challenging. These people were referenced by what region of their country they came from and how they were related to Osama bin Laden and Al-Qaeda. Recent articles have placed more emphasis on Al-Qaeda and how vast the network is and their ties to other Islamic groups. The current articles have started using the names of the terrorists and their known aliases. This has been a useful development, but has led to more confusion in putting the important and correct information into USentry.

As new documents were created in USentry, synopses for the documents were created to pull out the most important information. Most synopses have information about specific terrorists, their actions, how they are involved and with whom they associate.

USentry was used to baseline numerous terrorist groups affiliated with Osama bin laden and Al-Qaeda. At present there are approximately 47 different groups that have been identified and that participate directly or indirectly with Osama bin Laden and Al-Qaeda. Base-lining the groups and the events have linked terrorists, which helps identify relationships, groups, and events.

USentry helped the analyst identify and track the terrorists. Each terrorist was entered as a separate document with characteristics unique to that person. Characteristics captured included place of their origin, mosques they attended, significant events that they participated in, universities they attended, expertise, military history, and other groups they are affiliated with.

The physical characteristics captured include hair color; eye color; weight; height; build. Also captured were certain deformities; facial features; and the language they spoke and their language proficiency.

Terrorist Baseline

Approximately 300 known or alleged terrorists have been entered into USentry. Because the data changes daily, updates and revisions are constantly being added. Creating a person in USentry is similar to creating an autobiography, as we compiled all the articles on that person and sort through to capture the relevant characteristics as detailed above.

Other pertinent characteristics captured included company(s), international programs, passport details, last known locations, with whom they associate, and what other terrorist/events actions they have participated in. Once that person has been created, we are able to sort through the information and retrieve details about that person and the events in which they have been involved. Creating a person involves gathering past and present information known about that person and entering as a profile.

In one case the information that we entered on a particular person and who that person associates with, helped to establish this was the same person. Two reports gave information about this person, but the names and spelling were different for each one. Assuming they are the same person, one cannot be positive until you start looking at known associates and past travel routes. Putting this key information together enabled us to feel confident that is was the same person.

Tracking the information in this manner has been key in helping analysts decide the accuracy and relevance of the material. Much of the information that comes from open sources is sensational and biased. Our method of tracking terrorists and their characteristics allows the analyst to sort through the information and determine what is credible and what is not. Once this has been done, we then can start to identify patterns and connections.

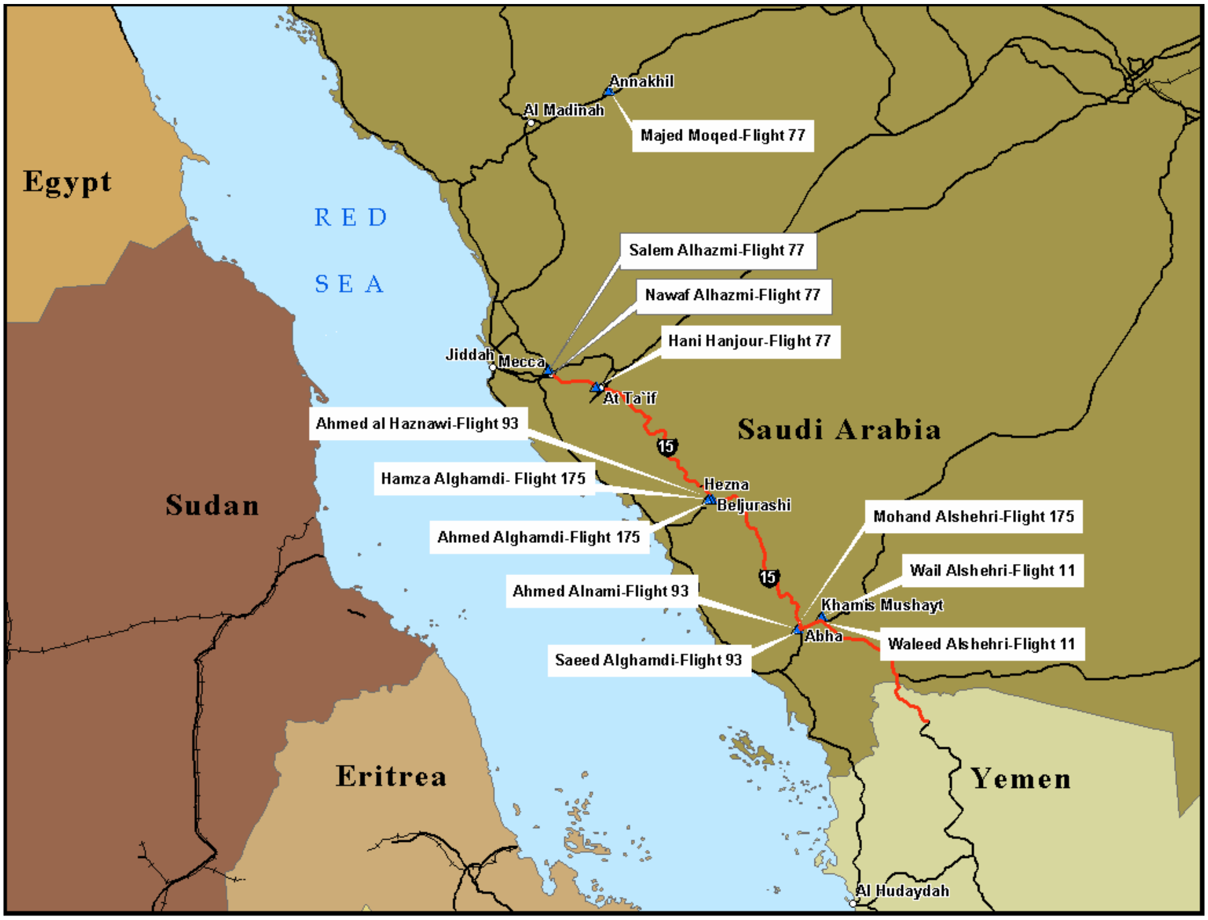
When the information was correlated it became clear there were numerous ties and connections between each of the individual terrorists. A VisioPro Chart was created to show the relationships between people, terrorist groups, terrorist cells, religious leaders, Mosques and terrorist events. Links were created between each person and their relationship to each of the other entities. This allowed a visual representation of the information. Additional ties were seen when reviewing the VisioPro chart. One particular tie that became apparent was the connection between Al-Qaeda and Hizbollah. It was reported that Imad Fayez Mugniyah; a Hizbollah member and head of security, also a possible Al-Qaeda member, and Osama bin Laden had contact with each other as early as 1993. This contact connects other terrorists and links their connections to different groups, organizations, and terrorists. Visio allows you to visualize the events, groups, cells, people, and mosques and how they are connected to each other.

Patterns

In order to find patterns in data on terrorists, articles from various sources were scanned for pertinent data that could be entered into a reference database. Gradually patterns began to arise from the data captured. These patterns were present in the articles themselves, however they were much more apparent when condensed and cross-referenced within a database. Examples of some of the patterns discerned are those dealing with person-to-person relationships, histories of involvement in terrorist activities, and places of origin or of occupation. Finding these patterns involved tracking individuals, consciously analyzing data, and searching for connections between people, places, and events. To further aide this process, the database could filter the data so that similarities in personal relations, events, and places could be more readily seen. An example of a pattern revealed in this data-mining process pertains to how specific terrorists can be related by their origins.

A more detailed analysis was done of the nineteen hijackers involved in the September 11th, 2001 attacks on the United States. It was found that fifteen of the nineteen hijackers came from Saudi Arabia; twelve of them from the southwestern Asir province of Saudi Arabia. The twelve hijackers are from towns that are strung out along a single highway, Highway 15, which is considered to be the masterpiece of construction by Mohammed bin Laden, father of Osama bin Laden. Because this area was passed over in the economic surge during the oil industry boom, very few of the well-educated youth of the area find satisfying jobs.

Highway 15 runs through the middle of an Asir town, Khamis Mushayt, home to two brothers, Waleed and Wail Alshehri, who were both hijackers on American Flight 11 that crashed into the North Tower of the World Trade Center. The brothers came from the prominent Seqeley family, a branch of the well-respected Alshehri tribe who assisted in the building of the highway and donated the Seqeley Mosque to Khamis Mushayt. It was at this mosque that the Alshehri brothers probably met Ahmed Alnami, a hijacker on American Flight 93 that crashed in the Pennsylvania woodlands, from near by Abha. The trio was joined by an Islamic Law student named Saeed Alghamdi , a school friend of Alnami's who was also from Abha and would later accompany Alnami on Flight 93. Together, the four took an oath of jihad in the Seqeley Mosque in the spring of 2000.



Mohand Alshehri was a hijacker on United Flight 175, which crashed into the South Tower of the World Trade Center. He also appears to have ties to Abha. According to his family, he attended Imam Mohammed bin Saud University in Abha, after graduating religious high school. He later dropped out to join in jihad.

Ahmed al Haznawi was raised near the town of Beljarushi. His father, the local imam, forbade him to join jihad in 1999. However in 2000, al Haznawi trained at the Al-Qaeda's Al Farouq camp in Afghanistan. When he returned, al Haznawi began recruiting his cousins and other members of his tribe and saw success when Hamza and Ahmed Alghamdi, also from Beljarushi, were successfully recruited. Hamza soon left his job as a stockboy in a housewares shop, a job considered humiliating by most Saudis. Al Haznawi died aboard Flight 93 and the two Alghamdi brothers were aboard Flight 175.

Four more hijackers are from cities along Highway 15. Further north of Beljarushi are the holy cities Mecca and Medina and the major metropolises of Jeddah (where 13 hijackers obtained their visas) and Taif. Hani Hanjour, the hijacker and pilot of the American Flight 77 that crashed into the Pentagon, was the son of a prosperous Taif businessman and was the only September 11th pilot from Saudi Arabia. Mojed Moqed, the son of a leader of the Baniauf Tribe and another 9/11 hijacker, was raised in the village of Annakhil, located near the holy city of Medina. Two brothers named Nawaf and Salem Alhazmi were from Mecca had joined jihad in Chechnya years before September 11th 2001.

Analytical Methodologies

CCS-3's goal for this project was to develop tools to help intelligence analysts do their jobs more rapidly, efficiently and effectively. This entails aiding analysts in rapidly and efficiently inferring from databases, texts and other materials certain relationships among individuals or organizations that may be involved in terrorist activity. These capabilities are important because of the enormous amounts of information that are available every day from publications and reports the world over, including newspapers, the Internet and reports from the field. No individual or even a team small enough to communicate amongst each other can possibly keep up with, let alone digest and synthesize, all this information.

Proximity and Semi-Metric Analysis of Social Networks: Uncovering Latent Associations (L. Rocha)

Document Networks (DNs) contain many possible relations among documents and between documents and semantic tags (e.g. documents in the WWW are related via a hyperlink network). Furthermore, documents can be related to semantic tags such as keywords used to describe their content. Mathematically, such structures are binary relations, that is relations between two sets of objects. Our semi-metric method analyzes the mathematical properties of these relations, conceived of as networks, to identify latent or indirect associations among specific elements not evident from a simpler examination, and to assess overall properties of the network with respect to the quantity and strength of such indirect connections. In this project, we used connections among the people, groups, and events as reflected in the database.

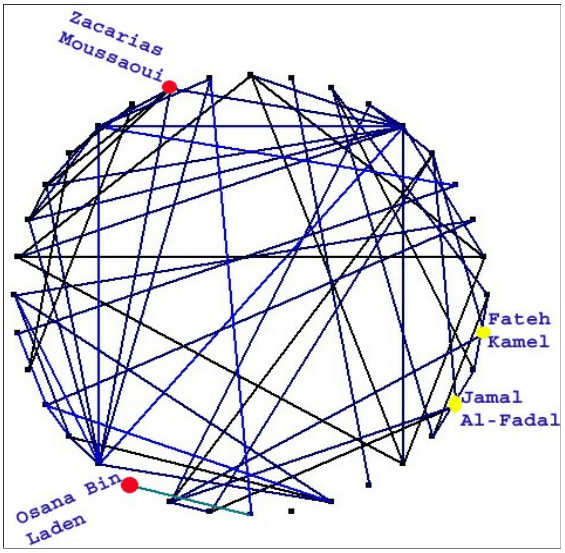


Figure 1: subset of **PDP** showing only links with proximity = 1

For each binary relation, our method calculates two proximity graphs, one for each set of objects. A proximity graph is a fuzzy graph that is reflexive and symmetric. The two proximity graphs associate objects according to how often they are related to the same elements of the other set in the relation. Specifically, given a binary relation R between sets X (of n elements x) and Y (of m elements y), we extract two complementary proximity relations: XYP and YXP. $XYP(x_i, x_j)$ is the probability that both x_i and x_j are related in R to the same element y 0 Y. Conversely, $YXP(y_i, y_j)$ is the probability that both y_i and y_j are related in R to the same element x 0 X.

In this project, an obvious relation to extract from the collection of news documents is the relation between people names and documents[1]. Given this relation, we calculate the proximity among people given how often they co-occur in the same documents (PDP), and the complementary proximity among documents given how often the same people are mentioned in them (DPP). We developed code to easily identify pairs of elements (e.g. people names) above a certain α 0 [0, 1]. We also studied the typical distribution of proximity weights in these graphs, and discussed them vis-a-vis graphs obtained from other datasets. Figure 1 depicts a subset of the PDP graph.

[1] In addition to this relation, we also analyzed the following binary relations **CITIES_DOCUMENTS**, **EVENTS_GROUPS**:, **EVENTS_PEOPLE**, **GROUPS_PEOPLE**, and **PEOPLE_RELATEDPEOPLE.**

From the proximity graphs, we generated distance graphs via an inverse transformation. The distances between elements of these graphs are not always metric. That is, for some pairs of elements, the shortest distance is not the direct edge, but some indirect path in the graph – thus breaking the triangular inequality required of metrics. Based on evidence compiled elsewhere [Rocha 2002a,2002b], we know that such semi-metric pairs denote a latent association in the data. That is, an association that is not grounded on direct evidence provided by the binary relation, but rather implied by the overall network of associations in this relation. In our context, a latent association is an implication that two elements (e.g. people names) are potentially related via other elements, but direct evidence for the association is not present in the available documents. This could be because direct evidence is missing, or the association, while very possible, has not occurred.

The methodology for capturing semi-metric behavior is based on three parameters defined in the full report. We applied this methodology to the several binary relations extracted from the project's dataset. We observed that most distance graphs obtained do not possess much nor strong semi-metric behavior. That is, they are close to metric graphs with small number of strong latent associations. This means that the documents from which we extracted the binary relations already capture most of the indirect associations. Given that these documents are media reports, this behavior is expected since journalists have essentially the same data available to them. We expect the situation to be different in real intelligence reports.

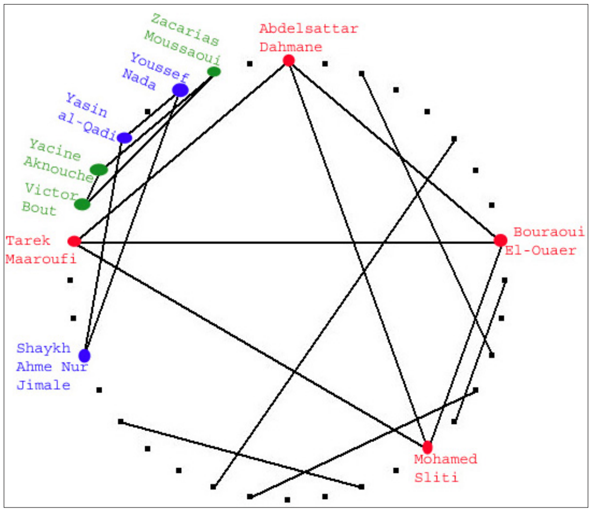


Figure 2: Two very semi-metric pairs in subset of PDP.

In any case, some latent associations between pairs of objects (e.g. people names) do exist in the data. We extracted the strongest of these for each distance graph. Figure 2 depicts two such pairs. The quality and relevance of these pairs was studied with user tests, which are described in the detailed report. The user tests ascribed higher relevance to the semi-metric pairs, well above random.

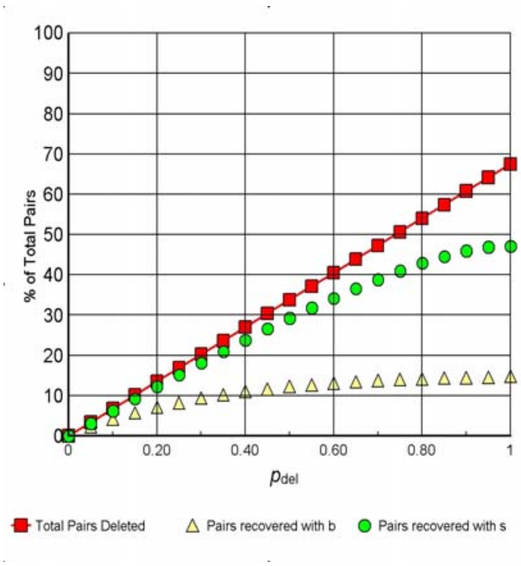


Figure 3: Percentage of pairs deleted and recovered with parameters s and b for several probabilities of partial deletion

However, given that the semi-metric methodology aims to discover latent associations, that is, associations between elements that the experts may not be aware of, user tests are not the best way to test the validity of the methodology. Therefore, in order to identify "holes" in the data, we conducted random deletion tests. First we computed the transitive closure of the proximity graphs obtained from binary relations. The transitive closure, a similarity relation, captures all possible indirect associations in the data. We start with this similarity relation and assume it to represent "perfect knowledge" – that it captures all real associations between elements. We then randomly deleted associations in the derived distance graph, with different probabilities of deleting an association, and observed how well the semi-metric methodology recovered the deleted associations.

As Figure 3 shows, the semi-metric methodology works extremely well to recover deleted associations. All details of this deletion analysis can be found in the full report [Rocha 2002c]. We are thus justified in claiming that the semi-metric methodology is indeed successful at uncovering missing knowledge in distance graphs obtained from document networks.

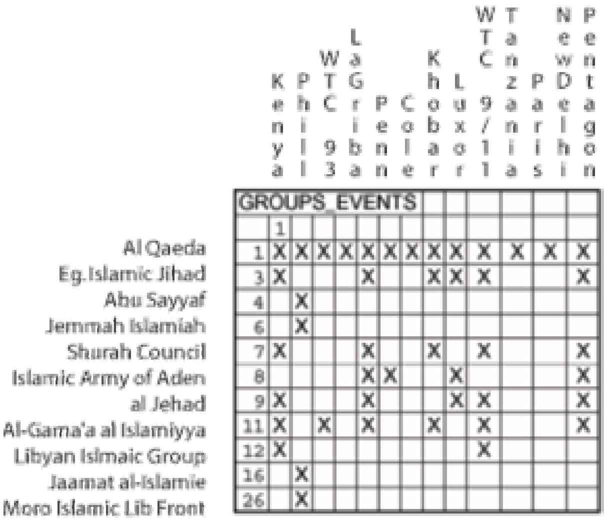
Formal Concept Analysis (C. Joslyn)

Formal Concept Analysis (FCA) [Ganter and Wille 1999] is a methodology for computer representation and analysis of boolean, binary relations derived from database tables. This representation supports both automatic inference and user-guided discovery and exploration of hypotheses, and provides an unbiased, visual display of the sub-relations present in complex relational data. FCA shows great promise for data analysis by supporting both manual inspection and query of relational data, and automated hypothesis generation and analysis to identify indirect associations among groups of elements

FCA depends on the subfield of combinatorics called Galois theory, and is also closely related to methods in association rule extraction and other knowledge representation techniques such as conceptual graphs. FCA is becoming established in a number of areas of information science, for example in natural language processing for information retrieval, biology, chemistry, environmental science, and for structuring phenotypes/genotypes in behavior genetics. Extensions to basic FCA include “fuzzy concept lattices” and “iceberg concept lattices”, which have been used for for database marketing, configuration space analysis, transformation of software class hierarchies, ontology learning, and database tuning.

In this component we investigated the formal basis for FCA, and began extensions in a number of directions (see below). We also performed an FCA analysis on the data concerning groups, events, and people using existing third-party tools, and consulted with the content experts about interpretation and usefullness. For complete details, see the full report [Joslyn and Mniszewski 2002b]. Here we review some issues using appropriate examples.

FCA works by taking a boolean, binary relation; calculating the connections, called **concepts**, among distinct groups of rows and columns; and then calculating the hierarchical relations between these concepts. We begin by analyzing the binary join tables represented explicitly the database: people x events, people x groups, and people x expertise. In addition, we constructed some relations, including some forms of ternary relations among three entities, special-purpose for this investigation: Groups x Events, People x (Events 4 Groups), Expertise x Groups, People x (Groups 4 Expertise), People x (Events 4 Expertise), and People x (Events 4 Groups 4 Expertise).



	GROUPS_EVENTS	Kenya	Phill	WTC93	LaGriba	Penn	Cole	Khobar	Luxor	WTC9/11	Tanzania	Paris	NewDelhi	Pentagon
		1												
Al Qaeda	1	X	X	X	X	X	X	X	X	X	X	X	X	X
Eg.Islamic Jihad	3	X				X		X	X	X				X
Abu Sayyaf	4		X											
Jemmah Islamiah	6		X											
Shurah Council	7	X				X		X		X				X
Islamic Army of Aden	8					X	X		X					X
al Jehad	9	X				X			X	X				X
Al-Gama'a al Islamiyya	11	X		X		X		X		X				X
Libyan Islmaic Group	12	X								X				
Jaamat al-Islamie	16		X											
Moro Islamic Lib Front	26		X											

Figure 4: Groups x Events

As an example, first see Figure 4, showing the connections among terrorist groups and events in the Groups x Events view. Then Figure 5 shows the corresponding concept lattice produced by the CONIMP tool, and Figure 6 the lattice produced by the GRAPHPLACE tool. A simple examination shows facts evident from the original table, for example that all events are related through Al-Qaeda, and that the Shura Council was involved in Pennsylvania, Pentagon, Kenya, Tanzania, and Luxor. But combining extracted rules and the lattice itself, even in this limited example we can derive a number of important, but less obvious, results, such as:

- Whereas Luxor was not related in the People x Events view, they are here.

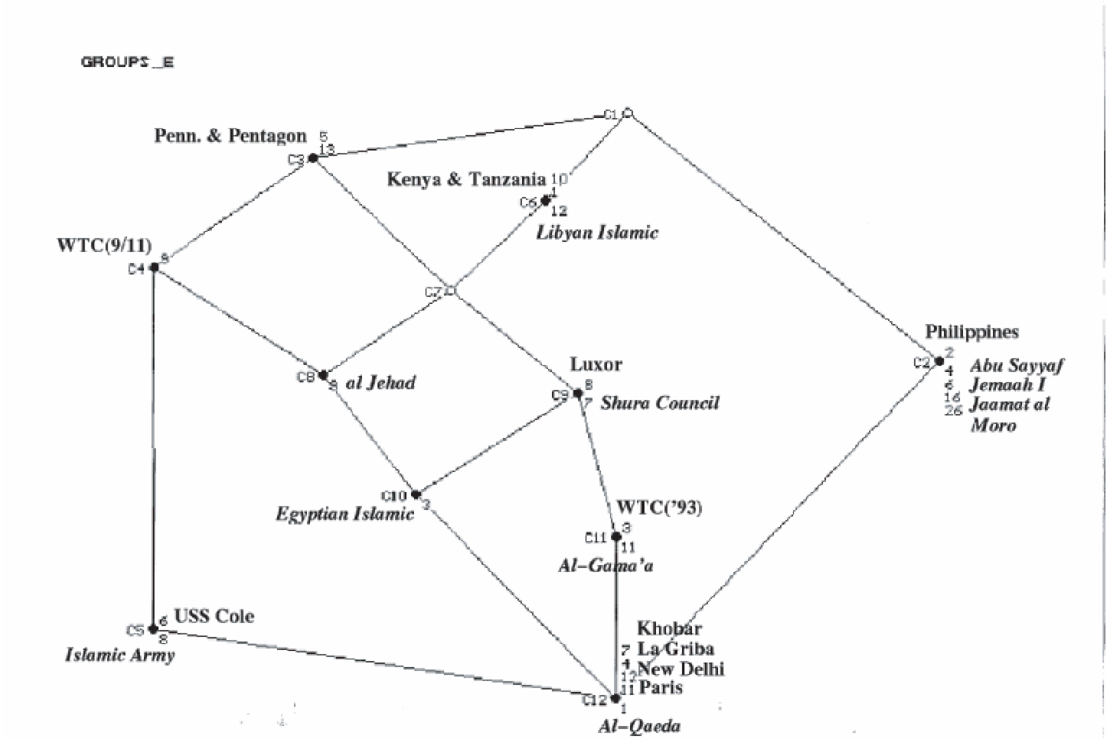


Figure 5: Concept lattice of Groups x Events: CONIMP

- The regular structure of the groups in could be interpreted as suggesting a hierarchy of subgroups within Al-Qaeda.

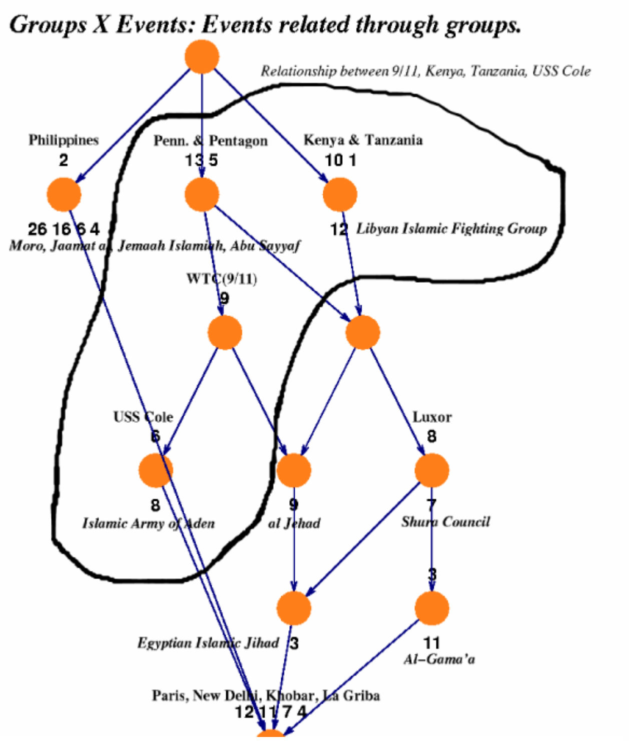

Groups X Events: Events related through groups.
Relationship between 9/11, Kenya, Tanzania, USS Cole
Philippines
2
Penn. & Pentagon
13 5
Kenya & Tanzania
10 1
26 16 6 4
Moro, Jaamat al, Jemaah Islamiah, Abu Sayyaf
12
Libyan Islamic Fighting Group
WTC(9/11)
9
USS Cole
6
8
Islamic Army of Aden
9
al Jehad
Luxor
8
7
Shura Council
3
Egyptian Islamic Jihad
3
11
Al-Gama'a
Paris, New Delhi, Khobar, La Griba
12 11 7 4

Figure 6: Concept lattice of Groups x Events: GRAPHPLACE

- Pennsylvania, the Pentagon, and 9/11 are related.
- The USS Cole bombing is related to 9/11 through the Islamic Army of Aden
- Kenya and Tanzania are related through the Libyan Islamic Fighting Group
- Kenya, Tanzania, and 9/11 are related through al Jehad
- Luxor, Kenya, Tanzania, and 9/11 are related through the Egyptian Islamic Jihad and the Shura Council of Al-Qaeda
- And finally, the World Trade Center, Luxor, Kenya, Tanzania, and 9/11 are related through Al-Gama'a al-Islamiyya

In addition, we have identified and begun exploration of some significant research questions, such as:

- Multidimensional FCA: Extension of FCA to n-ary relations, where n > 2.
- Non-Boolean relations: Extension to "fuzzy" cases where traditional "scaling" methods fail.

- Formal measures on hypotheses in concept lattices
- FCA tools in the context of an overall suite of knowledge discovery tools.
- And finally the relation to link analytical techniques.

Link Analysis and DataDelver

DataDelver (originally called VisTool) was developed in prototype form originally for a research project sponsored by the IRS to identify patterns of criminal fraud within databases of electronically filed tax returns. It was developed for the dual purposes of providing a schema-specific visualizing front end for analysts to examine the source database, and to provide a platform within which to implement and explore our research algorithms in user-guided knowledge discovery and link analysis. In this project, we recovered a prior prototype implementation, and deployed it against the current project database. For details on DataDelver, see the full report [Joslyn and Mniszewski 2002b]. Here we introduce some of the link analytical concepts which DataDelver approaches, and which are described in publications resulting from this project [Joslyn 2002a, 2002b; Joslyn and Mniszewski 2002b].

We've discussed how in this data set we have four variables of interest: people, events, groups, and expertises. The network analysis and FCA methods consider two variables at a time, for example people against events or people against groups (in the FCA, where a single variable was considered against a union of two or more others, for example People x (Events 4 Groups), the union Events 4 Groups is considered as a "single variable"). What we call **link analysis** [Joslyn 2002a, 2002b; OTA 1995] concerns questions about how *collections* of records are distributed over *collections* of fields. So, for example, given such a collection of records, how do they implicate one collection of fields or another? Similarly, how do they implicate other connected collections of records, perhaps being more, fewer, or somehow overlapping? This is a much more computationally difficult task, such that for large databases even the largest and most sophisticated computer-based systems are not now, and may never be, able to provide complete, automatic, answers to our questions. Instead, we aim at methods that are appropriate for moderately sized databases (10^2 to 10^5 records), semi-automatic, and expert user guided.

Assume a database of *M* data records and *N* field types, denoted D_{NM}. We use the concept of a **view** as the projection of D_{NM} to a particular subset of dimensions *n* *{ 1, 2, ..., N }* and restriction to a particular subset of records *m* *{ 1, 2, ..., M }*, denoted D_{nm}. **Chaining** then consists of moving from one particular view D_{nm} to another $D'_{n'm'}$, where *n 3 n' ! ℘, m 3 m' ! ℘*, or both, so that there are some rows and/or columns which are "held over" from the prior view. Conceptually, first an analyst considers certain aspects (*n)* of a certain group of records (*m*), for example including the place of birth of a group of people who all went to the same school. She then chains to consider another aspect, say the zip codes of those of that group who went to Harvard.

For more details, see [Joslyn and Mniszewski 2002b]. For here, first consider the illustration in Figure 7, showing a database table illustrated as a "contingency table",

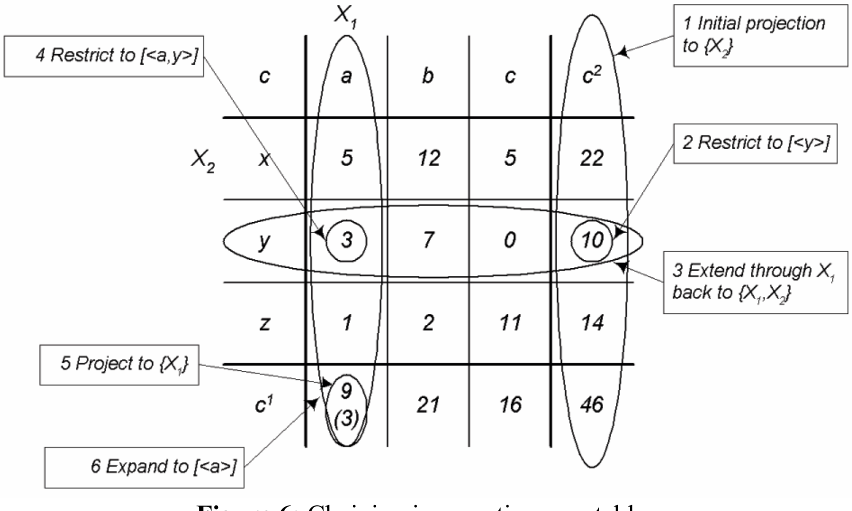


Figure 6: Chaining in a contingency table.

indicating the number of observations of each record type. The cells indicate the number of records with a certain vector value, and the marginal counts are included on the edges of the matrix. The chaining process begins with an initial view D_{2M} , all records projected onto the second dimension. The second step restricts this to D'_{2m}, where $m \quad \overline{M}$ now indicates those ten records such that $x_2 = y$. In the third step, D''_{Nm} indicates the same set of records, but now extended back both dimensions $N = \{ 1, 2 \} \rho \{ 2 \}$. Similar other steps are indicated.

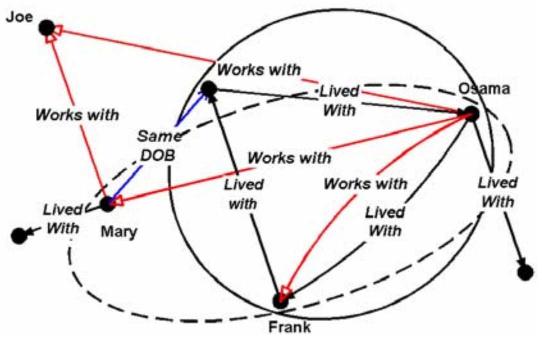


Figure 7 Chaining in a typed-link network.

Figure 8 shows a database using a quite different representation, namely a **typed-link meta-network** [Joslyn 2002b]. Here nodes are data records, links are fields held in common, and the type of the link indicates the type of the field. The concept of chaining is quite similar. The solid boundary indicates a collection of records $m = \{w,y,z\}$ viewed through the single field $n = \{f\}$, yielding $D_{\{f\},\{w,y,z\}}$. The dashed boundary indicates the transition to a new view on a different field type and somewhat different records $D_{\{g\},\{x,w,z\}}$, so that $n \; 3 \; n' = \wp$, but $m \; 3 \; m' \, ! \; \wp$.

We have previously developed the link analytical methodology called **Data Exploration through Extension and Projection (DEEP)**, which uses multidimensional database and information theoretical principles to help guide a user from one view to another [Joslyn and Mniszewski, 2002b]. DataDelver both implements DEEP, and provides other valuable services, in particular a **schema specific interface** customized by a knowledgeable administrator to reflect a particular relational schema, and a **view query manager** supporting the ability to construct and store queries supporting multi-dimensional views.

Semantic Feature Analysis (G. Papcun)

The goal of semantic feature analysis is to *find specific relationships* among the items discussed in the text. As opposed to finding co-occurrences, or other methods of determining whether two items are related, this method makes explicit the relationships among people, organizations, processes, *etc*.

The goal for this component comprises three main steps: (1) extracting information from databases, texts and other sources, (2) inferring relations and connections from the information extracted, and (3) representing and displaying the resulting information in ways that can be quickly and easily comprehended by analysts. We produced tools to extract information and present aspects of that information in ways, especially graphically that allow a user to become aware of events and relationships of interest.

Semantic feature analysis can answer specific questions, but has extended applications as well. For example, suppose a given individual is suspected of being a terrorist. Open-source newspaper accounts reveal that others were born in the same town. Some reports reveal that several of those who were born in the same town as the suspect also go to the same mosque. Immigration and Naturalization Service records reveal that the group entered the United States at the same time at the same border crossing. They take flying lessons together. Even without the 20-20 hindsight of 9/11, these kinds of facts could raise concern about a cadre that bears investigating.

Of course, many other individuals were born in the same town, and many other individuals entered the United States on the same day, and many other individuals go to the same Mosque as does the person of interest. However, a properly programmed computer, because it can tirelessly sort through immense quantities of data, is more likely to "notice" and bring to the attention of an analyst the fact that a group of individuals shares *all* these connections to the person of interest. Ultimately, these features can be interpreted as proximities, and used as input for statistical analysis and graphical presentation.

In this project we apply an approach called “construction grammar,” a linguistic theory in which phrasal patterns as well as words are viewed as learned pairings of form and meaning (Goldberg, 1995). Furthermore, we apply a radical approach to construction grammar in which we categorize components in terms of their relations to constructions (Croft, 2001) rather than in traditional categories such as parts of speech such as noun, verb, adjective, *etc*., or as sentence components such as subject, object, *etc*.

Why are we not using these time-honored syntactic categories? A simple answer is that by using the alternative categories to be described below we achieve better results with cleaner, more logical computer code. A deeper answer is that careful examination of the criteria for establishing the traditional categories results in contradictions or arbitrary choices. In “apple pie” is “apple” a noun or an adjective? If it is a noun, it should take a plural, but you cannot say “apples pie.” If it is an adjective, it should be capable of being modified by an adverb, but you cannot say “a very apple pie.” Selecting one or the other criterion would be arbitrary and is not in accord with any principled linguistic evidence. The fact is that we know what “apple pie” is. The criterion for whether something fits in the “—pie” slot is whether it is something you can make a pie out of, even mud.

We illustrate the method by finding the answer to “Where was X born?” There are many ways of stating that someone was born somewhere, each of which may be considered a construction:

- [person name] was born in [place name] as in “Osama bin Laden was born in Saudi Arabia.”
- [place adjective] –born [person name] as in “Saudi-born Osama bin Laden”
- [place adjective] –born [descriptive noun phrase] [person name] as in “Saudi-born terrorist Osama bin Laden”
- [person name], who was born in [place name], as in “Osama bin Laden, who was born in Saudi Arabia”

We have built a fully functioning Homeland Analysis System that is convenient and easily used. For complete details, see the full report [Papcun 2002a, 2002b]. Here we summarize some of the main components.

- **AllReferences:** This program takes as input the name of a document and a search term, which can be a word or a phrase. It returns a new document that lists every occurrence of the term together with various identifying information and a hyperlink back to the place in the original document where the instance of the term was found. An example of a results document is:

  Abu ZubaydahAbu Zubaydah   Abu Zubaydah.   A Palestinian from the Gaza Strip, where he was born in DocID: F87967EF510F257787256B65007A45B0

- **OntologyMarkupPlaces:** This program inserts markup symbols as hidden text into a document. It requires a separate table of the items to be marked up; in this case, a table of places. Although the program is now specialized for places, it could clearly be generalized for other kinds of markup, e.g., names, events, etc. As written now, it does not produce any additional output like documents with links such as AllReferences does, but could be extended to do so. This program is preliminary to and must be run before “Birthplace,” which is another program in this suite. Whether it ought to be triggered automatically by Birthplace is a human factors issue that should be considered on the basis of discussion and experience with users. A sample of a marked-up document is:

  > London Sunday Telegraph
  > June 24, 2001
  > 'Bin Laden's Bomber' Arrested
  > By Tim Brown in Madrid and Julian Coman in Paris Spanish police claimed yesterday to have foiled a plot to bomb the European Parliament with the arrest of the alleged European operations

```
chief of the wanted terrorist Osama bin Laden.Mohammed Bensakhria, 34, was arrested in a raid
on a telephone and fax centre in Alicante, on the south-east coast of Spain, on Friday. Another
```

- **Birthplace:** This program takes as input the name of a person and a document to be analyzed. Using advanced linguistically-based algorithms, it scans the document to find information about the birthplace of the person. To run Birthplace, the program OntologyMarkupPlaces must first have been run on the document to be analyzed. An example of a result document is:

```
Saudi Osama bin Laden
DocID:  A453E41BE126511A87256B660075BA0F
```

An example of a Birthplace Evidence Result Document is:

```
U.S. officials have said {o{placeNounSaudi}}-born dissident Osama bin Laden, now a
DocID:  A453E41BE126511A87256B660075BA0F
```

References

[Bybee *et al.* 1994] Bybee, Joan, Revere Perkins, William Pagliuca *The Evolution of Grammar: Tense, Aspect and Modality in the Languages of the World*. Chicago: University of Chicago Press.

[Croft 2001] Croft, William, *Radical Construction Grammar: Syntactic Theory in Typological Perspective*. New York: Oxford University Press.

[Davis 2001] Davis, Anthony R. *Linking by Types in the Hierarchical Lexicon.* New York: Oxford University Press.

[Ganter and Wille 1999] Ganter, Bernhard and Wille, Rudolf, *Formal Concept Analysis*, Springer-Verlag

[Goldberg 1995] Goldberg, Adele E. *Constructions: A Construction Grammar Approach to Argument Structure*. Chicago: University of Chicago Press.

[Joslyn, 2002a] Joslyn, Cliff: "Network Worlds: From Link Analysis to Virtual Places", *Proc. 2002 Conf. on Virtual Worlds and Simulation*, ftp://ftp.c3.lanl.gov/~joslyn/vwsim02f.pdf

[Joslyn 2002b] Joslyn, Cliff: "Link Analysis of Social Meta-Networks", *2002 Conf. on Computational Analysis of Social and Organizational Systems (CASOS 02),* ftp://ftp.c3.lanl.gov/~joslyn/casos02f.pdf

[Joslyn and Mniszewski 2002a] CA Joslyn and SM Mniszewski: "Relational Analytical Tools: VisTool and Formal Concept", LAUR 02-7697, ftp://ftp.c3.lanl.gov/pub/users/joslyn/hl1.pdf

[Joslyn and Mniszewski 2002b] CA Joslyn and SM Mniszeiski, "DEEP: Data Exploration through Extension andProjection", LAUR

[OTA 1995] Office of Technology Assesment: (1995) "Information Information Technologies for Control of Money Laundering", OTA-ITC-630, US GPO, Washington DC, http://www.wws.princeton.edu/~ota/disk1/1995/9529_n.html

[Papcun 2002a] George Papcun, "Homeland Analysis System User's Manual", LAUR

[Papcun 2002b] George Papcun, "Computational System for Homeland Defense Analysis", LAUR

[Papcun 2003] George Papcun, Kari Sentz, Andy Fulmer, Jun Xu, Olaf Luk, Murray Wolinsky, "A Construction Grammar Approach to Extracting Regulatory Relationships From Biological Literature," *Pacific Symposium on Biocomputing*, Lihue, Kauii, Hawii, January 2-7, 2003.

[Rocha 2002a] Rocha, Luis M. "Combination of Evidence in Recommendation Systems Characterized by Distance Functions". In: *Proceedings of the 2002 World Congress on Computational Intelligence: FUZZ-IEEE'02.* Honolulu, Hawaii, May 2002. IEEE Press, pp. 203-208. LAUR 02-154

[Rocha 2002b] Rocha, Luis M. "Semi-metric Behavior in Document Networks and its Application to

Recommendation Systems". In: *Soft Computing Agents: A New Perspective for Dynamic Information Systems*. V. Loia (Ed.) International Series Frontiers in Artificial Intelligence and Applications. IOS Press. In Press. LAUR 02-3316.

[Rocha 2002c] Rocha, Luis M. "Proximity and Semi-Metric Analysis of Social Networks". Internal Report of Advanced Knowledge Integration In Assessing Terrorist Threats LDRD-DR – Network Analysis Component. LAUR 02-6557.