Novel Information Discovery for Intelligence and Counterterrorism

D.B. Skillicorn N. Vats
School of Computing
Queen's University
{skill,vats}@cs.queensu.ca

September 2004
External Technical Report
ISSN-0836-0227-
2004-488

School of Computing
Queen's University
Kingston, Ontario, Canada K7L 3N6

Document prepared September 30, 2004

Abstract

Intelligence analysts construct hypotheses from large volumes of data, but are often limited by social and organizational norms and their own preconceptions and biases. The use of exploratory data-mining technology can mitigate these limitations by requiring fewer assumptions. We present the design of the ATHENS system, which discovers novel information, relative to a specified set of existing knowledge, in large information repositories such as the world wide web. We illustrate the use of the system by starting from the terms "al Qaeda" and "bin Laden" and running the ATHENS system as if on September 12th, 2001. This provides a picture of what novel information could have been known at the time. This is of some intrinsic interest, but also serves to validate the performance of the system since much of this novel information has been discovered by conventional means in the intervening three years.
Keywords: intelligence analysis, counterterrorism, information discovery, novelty, al Qaeda.

Novel Information Discovery for Intelligence and Counterterrorism

D.B. Skillicorn and N. Vats
skill@cs.queensu.ca

Abstract: Intelligence analysts construct hypotheses from large volumes of data, but are often limited by social and organizational norms and their own preconceptions and biases. The use of exploratory data-mining technology can mitigate these limitations by requiring fewer assumptions. We present the design of the ATHENS system, which discovers novel information, relative to a specified set of existing knowledge, in large information repositories such as the world wide web. We illustrate the use of the system by starting from the terms "al Qaeda" and "bin Laden" and running the ATHENS system as if on September 12th, 2001. This provides a picture of what novel information could have been known at the time. This is of some intrinsic interest, but also serves to validate the performance of the system since much of this novel information has been discovered by conventional means in the intervening three years.
Keywords: intelligence analysis, counterterrorism, information discovery, novelty, al Qaeda.

1 Introduction

It is helpful to think of an intelligence analyst as interacting with two different spaces: an *information space* and a *hypothesis space*. The information space contains facts of various kinds ('data'); it is typically extremely large. There are many difficult issues in the design and use of such a space, for example whether its content is controlled by the analyst (a pull model) or from some outside source (a push model); whether there are gatekeepers who control what information can appear within it; and what is the relationship between organizational barriers and separation of information spaces. Good solutions to these problems are necessary to good analysis, but they are in the domain of information retrieval, and we will not consider them further. The critical feature is that the information space is a passive object, providing no guidance about how its contents are to be understood or interpreted. So improvements in information spaces do not necessarily lead to improvements in analysis.

The hypothesis space contains the hypotheses ('knowledge') derived from the information space. These hypotheses have a natural quality metric based on (a) the evidence that supports them in the information space, and (b) their explanatory power. A good hypothesis is one for which there is good evidence in the information space, and which has some predictive power (in Popper's terms, is falsifiable [11]). Such hypotheses become the basis for action in the real world. A poor hypothesis is one that is not supported by the data, or one that doesn't explain much (it is either bad or useless).

The central task of an intelligence analyst is to populate the hypothesis space with high-quality hypotheses. The standard way to do this is to use the analyst's skills, intuition and hunches to develop hypotheses and then validate them (or not) against the data in the information space. This process is often iterative; an initial hypothesis is partly validated, but the evidence against it suggests an alteration that creates a better hypothesis.

It is clear what the limitations of this process are. Hypotheses that are inconceivable to the analyst are never examined against the knowledge base. In practical terms, possible but unlikely hypotheses are never examined either. The process is fundamentally limited by analyst (and organizational) norms and preconceptions. This seems to be at the heart of what the 9/11 Commission

[5] described as a "failure of imagination" in the runup to the September 2001 attacks against the World Trade Center.

One possible enhancement for intelligence analysis is exploratory data mining, a family of techniques that are able to build models from data without prespecified hypotheses. Of course, this approach contains its own preconceptions, both in the choice of model building technique, and in the parameters used to build each model. However, these preconceptions are of a different kind, assumptions about *structure* in the data, rather than the social and political norms that can often affect a human analyst. Techniques such as social network analysis and link analysis have been used to search for criminal behavior [1, 8], for patterns in communication [13, 16], and for counterterrorism [3, 12]. Techniques such as matrix decompositions [14] have also been used for counterterrorism. Overviews of this approach are Popp *et al.* [10] and Taipale [15].

The techniques mentioned above assume that the information space is a single dataset that can be processed directly using flat file or database storage. They cannot be directly applied to an information space such as the world wide web. The content of the entire world wide web is surely no more than a few hundred petabytes (which is well within the range of high-performance data mining systems) but the content is arranged in an awkward and highly distributed way. Techniques are needed that can handle both the pragmatics of size and distribution, and also the fact that almost all of the content of the world wide web is irrelevant to most hypotheses.

We present the ATHENS system, which can be used to attack the hypothesis-generation problem using large information repositories such as the web as its information space. Search engine technology is adequate for discovering all of the information about a topic for which reasonable descriptors (keywords) are known. Several search engine enhancements are even capable of organizing the results of a search in useful ways, for example clustering similar pages [6], or ordering so that the most interesting pages (by some metric) are presented first. This already represents a step towards the intelligence analyst's goals. However, the major drawback with the search engine approach is the need for descriptors; a search engine cannot search for something that the analyst does not know about, and so is limited by the analyst's preconceptions in the same way as other directed tools.

The ATHENS system is explicitly designed to find novel information (that is whose existence is not known to the analyst), but novel information contextualized by what the analyst does already know. In other words, ATHENS does not produce random nuggets of new information, but rather enables answers to questions such as: "I know all about topic X; which other topics Y, Z are related to X but are not easily findable knowing only X".

To use ATHENS, an analyst provides a set of keywords representing knowledge that is already familiar. The system returns clusters of new information that is relevant to the familiar knowledge but too indirectly connected to be easily discovered by browsing from the pages containing the familiar knowledge.

We illustrate the use of ATHENS by showing the results from the initial keywords "bin Laden" and "al Qaeda", executing the system as if on September 12th, 2001. These results are of considerable intrinsic interest, since they show what knowledge would have been readily accessible had ATHENS been in existence then. These results also allow us to validate the ATHENS system because much of the information related indirectly to the query concepts has been discovered in the intervening three years [4, 7]. Overall, much useful information that has come to light over the past three years would have been available contemporaneously using ATHENS. Significantly, however, these results support the claim that, at least from public data, the September 11th attacks could

not have been predicted.

2 The Athens system

Search engines are designed to find and retrieve the most relevant documents corresponding to a user query. However, this does not suit them well for open-ended, exploratory knowledge discovery. Suppose a user searches using keyword X. The list of retrieved documents will contain references to other topics, some of which may be new to the user. However, the ordering of the search results will tend to spread these references to new topics Y and Z at random through the list, so it is hard for the user to notice them.

Furthermore, if a user does notice topic Y, a search on Y is likely to produce another large list of documents which will mention further new topics. In other words, the occurrence of new information in search results is both random and growing in size.

For example, a Google search for "osama bin laden" returns 582,000 pages. Only 9440 of these pages also mention Hambali, a major link between al Qaeda and South-East Asian Salafist terrorists, the first page ranked at position 121. A new search on "hambali" returns 20,100 pages. It is apparent that the process of elucidating the important connections within al Qaeda using this approach quickly becomes impractical. The problem is that the world wide web, considered as a graph whose edges represent co-occurrence of terms, has high degree. Searching out from a known set of terms or pages reaches a huge number of pages within only a few steps. Furthermore, there are few hints about which 'directions' in this graph are likely to be productive.

The Athens system provides a focused way to look for novel information in systems such as the web. It begins with a set of terms, representing the user's existing knowledge. This initial information is used to create a representation of the user's background knowledge, from which new, contextualized search queries are constructed. The results of these queries are clustered, both to remove less useful information and to organize what is found. After this phase, the content retrieved by the system is a good representation of what the user knows or could easily discover using standard techniques. The entire process is now repeated, starting from each of the first-level clusters, to retrieve content that is both relevant (because of the contextualized search) and novel (because it goes beyond what can be easily discovered).

The key steps of the Athens discovery procedure are:

- Closure: This step identifies the central content that the user's list of keywords represents. It is implemented by searching using the keywords, selecting the most relevant pages returned, and extracting a concise description of their content. In the current implementation, this description is a set of nouns ranked by importance. Closure ensures that the starting point for information discovery is not skewed by the user's particular choice of keywords (or from a different perspective, permits users to be casual in their choice of initial keyword lists).

- Probe: This step begins the process of acquiring new information from the foundation of the closure. New queries are generated by combining terms from the original keywords with terms from the closure.

- Cluster: This step organizes the information returned by the probe queries. Pages are clustered using a spectral partitioning technique. Those pages that do not fall into clusters are discarded; each remaining cluster is presented as a unit of novel information, with a set of descriptive words extracted from it.

- Iterate: The iterate step repeats the three steps above, using the cluster descriptors as the starting points for the second iteration.

2.1 Algorithm

Given an initial search query Q (a set of keywords), the following operations are applied:

1. Closure:
   - Retrieve a subset S of the most relevant web pages for Q using an underlying search engine. In the current implementation, ATHENS uses the Google WebAPI to retrieve search results. This API enables searches to include phrases as well as single words, and to restrict queries to particular domains or time ranges. (Other search engines and other information repositories can be used by making a few low-level changes in the system.)
   - Create a list of nouns and their frequencies for each page using the MontyTagger, a parts-of-speech tagger [9]. The nouns from the original query (which must necessarily be present) are removed at this stage. (The motivation for using nouns is that, in English, they best capture the content of the page.)
   - Combine the noun lists from all pages into a single list, summing their frequencies. Note that this automatically gives longer pages (those with more content) more influence. A stopword list is used to remove common words.
   - Eliminate the less discriminating nouns by comparing their relative frequency in the combined list to their relative frequency in the BNC corpus [2], a large collection of written and spoken English. Only those nouns whose relative frequency in the retrieved pages is greater than in ordinary English are retained.
   - Order the list by descending differential relative frequency (i.e. how much the relative frequency differs from that of the BNC).
2. Probe:
   - Form a set, Q, from the original search terms by leaving out one term each time. Form the cartesian product of Q with the list of nouns constructed during the previous step, ordering the product by the order of the list.
   - Select some prefix of this ordered list and use each set of terms as a search query. Create noun lists from the returned pages as in Step (1).
   - Create a page-page matrix, P, whose ijth entry represents the similarity between page i and page j. Let L_i and L_j be the noun lists for pages i and j respectively, and f_{n_i} be the frequency of noun n in page i. Then the *Jaccard similarity* between pages i and j is

     $$\frac{i \cap j}{i \cup j}$$

     where

     $$i \cap j \ = \sum_{nouns} min(1, f_{n_1}, f_{n_2})$$

     and

     $$i \cup j \ = \mid L_1 \mid + \mid L_2 \mid$$

3. Cluster:

   - (a) Compute L, the normalized adjacency matrix of P, whose off-diagonal elements are

$$L_{ij} = \frac{P_{ij}}{\sqrt{d_i d_j}}$$

     where d_i is the *degree* of page i, the row sum in P. The diagonal elements of L are set to 0. L is a normalized representation of P.
   - Perform SVD on L and truncate to k dimensions so that

$$L \approx U_k S_k V_k'$$

   - Cluster the pages by putting two pages in the same cluster if the magnitude of their vector sum exceeds α of the sum of their magnitudes. A page which does not fall within any existing cluster becomes the seed of a new cluster.
   - For each cluster, generate a descriptive set of nouns, and a web page consisting of links to the pages in the cluster.

4. Iterate: Repeat the steps above for each cluster, using a prefix of the descriptive set of nouns as the initial keyword set.

ATHENS requires an underlying environment that is able to produce ranked lists of responses to a search query. Hence it is easily portable to other settings. It is also, at present, limited to English because of the dependencies of the tagger and the use of the BNC to determine how unusual each word is. These deficiencies are pragmatic rather than fundamental.

3 Experiment

We now illustrate the application of ATHENS by beginning from the initial keywords "bin Laden" and "al Qaeda" *as the system would have performed on September 12th, 2001* [1]. The purpose of this example is to illustrate what novel information would have been available immediately after the attack. The presumption is that data immediately relating to Osama bin Laden was known and understood. We now know that this was not entirely the case, but that aspect of the problem is not addressed here. ATHENS provides an answer to what information might have been missed, and what information might have been underappreciated because it was too diffuse to be detected.

The following parameters were used: number of pages in Closure: 10, number of pages in Probe: 5, number of new queries in Probe: 20, α: 1.92, and cluster representation: 3 terms (first phase), and 15 terms (second phase).

We first show the three-word queries generated as cluster centers after the first phase (Figure 1). These clusters are not part of the output of the system, but they are useful in better understanding the output of the second phase. All of the descriptors at this stage have an obvious relevance, although it is clear that Set 8 are very general, so we might expect that its derivatives at the next level might be insufficiently contextualized to be useful. The problem here is that the ATHENS system is purely syntactic. Hence it correctly discovers that "New York" is a relevant term but

cluster	descriptor
1	Kherchtou Nairobi Mohamed
2	Mohamed Odeh United
3	Pakistan Taliban Afghanistan
4	Afghanistan American United
5	Kosovo Islamic Western
6	Muslim Peninsula Americans
7	Mullah Rabbani Taliban
8	York States United
9	Pakistan India Terrorism

Figure 1: First level cluster descriptors

does not understand that it is a phrase and its words should be kept together. "York" as a search term loses context.

Table 1 shows the 15-term descriptors generated for each second-level cluster. The system generates HTML pages containing the complete list of URLs corresponding to each cluster. These pages are the most useful way to interact with the results of the ATHENS system. However, the 15-term descriptors provide a way to summarize the system's output.

The 15-term descriptors are generated from the complete list of nouns in the pages of a cluster, with stopwords removed. Apart from acting as a compact description of each cluster, they are also useful as a set of search terms to find further content related to each cluster.

Humans tend to assume that the structure of the web approximates a human ontology, but we have evidence from the use of ATHENS that this is not the case. In particular, well-defined clusters exist whose content is coherent but is not of the form that a human would have constructed. Sometimes the clusters generated by ATHENS do not seem obviously coherent to humans; and yet their existence is clear. This is a drawback to 15-term descriptors; although they are good descriptors for the content of a cluster, humans sometimes find them opaque.

Humans using ATHENS also exhibit *ontological bias* – they expect the system to return clusters that are not only novel and relevant, but are also ontologically similar to the initial query. Because ATHENS, and the tools on which it is built, are essentially syntactic, clusters sometimes surprise users because they seem (superficially) to be "off topic". Although this can be perceived as a weakness of ATHENS, it is arguably one of its strengths, since it presents information that is less biased by user expectations than a more semantic system would.

Table 1: Descriptors for second level clusters

Label	Cluster identifier and search terms
1.1	Odeh Hage Rick Halperin York Albright United Embassy States Judge American Florida Americans Kenya Tanzania
1.2	Blair Hutchinson Poage Piper Rick Halperin Texas Ashley Flaherty Engleton Chester County Elmore Allan Rensch

[1]Google does not maintain snapshots of the web at different times, so the possibility of capturing a page that has changed more recently exists. In fact, only one anachronistic page was discovered among the results.

Table 1: (continued)

1.3	Texas Halperin Rick United States Gaudin Kenya Salazar Press Hage Clinton Messages Odeh Calif District
2.1	Nairobi York Kenya Osama Tanzania Americans Franken States Washington Roger Cossack Embassy Khamis August Khalfan
2.2	American States Somalia Islamic president General Nations Saudi August Ladin Egypt Israel Morocco UNOSOM State
2.3	Fazul Dalitz Hage Islamic American Owhali Nairobi Osama Jews Afghanistan Arab Sudan States EmergencyNet ERRI
2.4	Reza Washington York Embassy Bombing Prosecution State April Department Hage Saudi Federal States Government East
2.5	Chair Abdul Razak Professor States Malaysia Ohio Sulaiman University Scholar America Prime Minister Southeast
3.1	Iran Islamic Kabul United Islam Sunni States Shia Muslim India Sharif American Russian Alliance Tehran
3.2	RAWA Peshawar Women Kabul Afganistan International Association Revolutionary April Afghans Secretary Minister NWFP Chief Party
3.3	Islamic United Indian Kandahar Kabul States Omar Islam Bamiyan Buddhas Secretary Buddha December India Taleban
3.4	Iran Islamic Taleban September Tehran Sharif Islamabad Saudi Mazar Republic York Aziz Arabia Kabul Islam
3.5	Iran Sharif Kabul United Mazar President Nations Teheran Afghans York Mission Shia August Security States
3.6	Islamic Kabul Kandahar Islam Kashmir Afghans General United Mujahideen Iran India Sharif Asia Americans Muslim
3.7	Islami Islamic Hikmatyar Hizb Rabbani Khalis Islamist Pashtun Jamiat Mujahideen Kabul Burhanuddin Muslim Party Mohammad
3.8	Islamic Muslim Muslims Quran Allah Prophet Quranic Mohammed Ambassador Islam Shukriya America Hashemi Kabul Hindus
3.9	Government August Opinion Business Brig Imtiaz Bank Hindu India Intelligence Nangarhar Tech Catalyst Investment Banking
3.10	Islamabad Sudan Internet Friday Osama Lahore InfoTimes Career Service Karachi Kashmir Services Peshawar Quetta Nation
4.1	Hoover Taliban Kandahar Kabul Soviets Taliban Islamic Azad Jamiat Jihaad Straight Communist Mike Journalists Cindy
4.2	Taliban States Iran Pakistan Middle East Islamic Brown University Americans William Beeman html Economic Bombings
4.3	President Islamic Kyrgyzstan Kyrgyz Taliban Central States Akayev Republic Veterans Libya Iran Egypt Situation Bush

Table 1: (continued)

4.4	Islamic State Embassy International Kabul Department Asia Pakistan Travelers Sharia Globe Washington Medical Information Global
4.5	Taleban Islamic Pakistan Taliban Iran Osama International Amnesty Nations Kabul General States Islam Saudi Muslim
4.6	Iran States India Islamic Tehran Pakistan York Russian State President Anglo Department World Britain Washington
4.7	States State Islands President Vidal Okinawa Address Japan Union Soviet Sutton Island Inaugural America January
4.8	States Rights Human July April Colombia March Argentina Detention Death Committee Torture Americas International Violations
4.9	Islands Jan28 Island Mar31 Republic Jan4 Apr31 South North Codes Country Guinea Arab Cape British
4.10	Islands Africa Assigned ASIA Republic Island South Coded East Arab Yemen French Azerbaijan Central West
4.11	Dollar Franc Pound Islands French States Peso East Countries Caribbean Dinar Zealand Rupee Island Guinea
4.12	Muslim Islamic Muslims Quran Taliban Allah Prophet Quranic Mohammed Shukriya Islam Kabul Hindus Sikhs Perfect
5.1	Albanians Albanian Orthodox Serb Church Serbian Serbs Muslim Metohija Europe Bosnia Christian Muslims Serbia Catholic
5.2	Albanians Albanian Albania Muslims Muslim Serbian NATO Serbs Europe Kosovars Kosovar Macedonia West Yugoslavia Balkans
5.3	NATO Yugoslavia Serbs Bosnia Albanians Serbian Albanian Milosevic Serbia Yugoslav Serb Muslims Bosnian Muslim Balkans
5.4	Albania Bosnia Muslim Iran Serbs Iranian Saudi Albanians Saudis Israel Afghanistan April Croatian Yugoslav Albanian
6.1	Islam Islamic Muhammad Allah United States World Christian Mohammad Melungeon English Christians Elijah Prophet North
6.2	Almighty Jihad Islamic Islam Jews Crusaders Shaykh Muhammad Group Egypt Arabian Front Statement Iraq Imam
6.3	Philippines Marcos Filipinos Base Pacific Indigenous Filipino President Clark Force Island Subic Naval Japan Army
6.4	Islam Philippines Filipinos Mindanao Moro Christian Peace Agreement Filipino Christians Islamic ARMM South Philippine Province
6.5	Republic Vietnam Vung Mindanao Philippines China Pacific Saigon Japan Empire United Japanese Singapore World Dairen
7.1	RAWA Afghanistan Peshawar April Afghans Pakistan Tuesday Saudi Police Revolutionary Association Kabul Women Khalili Road

Table 1: (continued)

7.2	Mirror China Afghan Afghanistan President Japan Islamic State Chinese Tech Advanced Sitemap Parliament European Massoud
8.1	School District Local Nevada Middle Schools Public High County City Unified Junior Shelley Valley East
8.2	Census County Pennsylvania Listings Genealogy Records Data USGenWeb Septennial Update Ancestry Sign MyFamily Software Policy
8.3	County Census Genealogy GenSource Guide Found Records Common Archives Policy Pennsylvania Ancestry Directory Highlighted Tree
8.4	Pennsylvania Times Daily Journal Post Herald Gazette Australia Star Sunday Morning Pittsburgh National Sharon Pakistan
8.5	Gazette Boston Pennsylvania Advertiser Journal London Massachusetts Weekly South Carolina American Virginia Bath Hampshire Great
8.6	University College Michigan Illinois Virginia Xavier Wisconsin Tech Boston Univ Millersville Southern School Tennessee Central
9.1	South Asia Bush President States Kashmir United American China Russia Ambassador Foreign Brookings Afghanistan Policy
9.2	Kashmir October Jammu Kashmiri Killings Times Terrorists Information Poonch Terrorist Islamic Pandits Network Tribune Srinagar
9.3	Kashmir Islamic Islamabad State Muslim Department Afghanistan Singh China Delhi Asia Nuclear Kashmiri South Government
9.4	South Reuters Durban Paul Africa Sept Richardson Aligned Movement 12th Delegates Comprehensive Nuclear Test Treaty
9.5	Kashmir Jammu Islamic Clinton Sharif Hindus Hindu Times Kashmiri Iraq President Lord Avebury Indian Minister

One of the striking things about these clusters is that they mention almost all of the countries that have turned out to be important in the history of al Qaeda and the fight against terrorism: Afghanistan, Albania, America, Australia, Azerbaijan, Bosnia, Egypt, India, Iran, Iraq, Israel, Japan, Kenya, Morocco, Pakistan, Philippines, Russia, Saudi Arabia, Serbia, Somalia, Sudan, United States, Tanzania, Yemen, Yugoslavia, as well as regions such as Kashmir.

Recall that the African embassy bombing suspects were on trial in the U.S. during the summer of 2001. The clusters 1.1–1.3 concern the death penalty issue, with some connections to the embassy bombings, but to other death-penalty cases as well. Clusters 2.1–2.4 focus on the embassy bombing trials, as well as related bombing operations in Somalia. At this time, the connection between these bombings and Osama bin Laden was not considered well-established, and there was a tendency to regard him as a financial backer, rather than as a terrorist leader.

Clusters 3.1–3.10 can be summarized as dealing with Afghan and regional politics. Cluster 3.1 concerns Islamic religious politics; cluster 3.2 concerns the role of women in Afghanistan; cluster 3.3 concerns the destruction of statues of the Buddha in Afghanistan by the Taliban; cluster 3.4 concerns the Taliban; cluster 3.5 concerns relations between the Taliban and Pakistan; cluster 3.6

has similar content but in a wider context; cluster 3.7 concerns mujahideen groups and leadership; cluster 3.8 consists of pro-Taliban propaganda; cluster 3.9 concerns the role of heroin in the region; and cluster 3.10 concerns perceived U.S. aggression towards Pakistan.

Clusters 4.1–4.12 can be summarized as history and geography. Clusters 4.1–4.6 provide background on Islamic countries from the former Soviet Union through to India. Cluster 4.7 concerns U.S. history and geography. Cluster 4.8 concerns human rights, while cluster 4.12 contains anti-Taliban propaganda. Clusters 4.9–4.11 are nice examples of shifts in ontologies: 4.9 gives telephone country codes; 4.10 contains details of country-specific properties such as postage stamps; and 4.11 contains details of currencies. These clusters might be considered as indicating the extent to which the initial search terms reflect a global phenomenon.

Clusters 5.1–5.4 have content specific to the Balkans, both the conflicts following the breakup of Yugoslavia, and the subsequent civil war in Kosovo.

Clusters 6.1–6.5 can be summarized as history. Cluster 6.1 is about clashes between religions, particularly in the past two centuries. Cluster 6.2 contains some of the historical material used by Salafist Islam to justify its jihad against the west. Cluster 6.3 concerns the older history of the Philippines, and cluster 6.4 the more recent Islamic history. Finally, cluster 6.5 is a summary of Pacific history. This set of clusters is particularly interesting because it makes the connection between Middle Eastern al Qaeda and the Islamic terrorist organizations in the Far East, including the Moro Islamic Liberation Front.

Clusters 7.1–7.2 are the least consistent. Cluster 7.1 contains further material on the oppression of women in Afghanistan; while cluster 7.2 contains Chinese commentary on the Taliban.

Clusters 8.1–8.6, as expected, are internally cohesive, are not sufficiently contextualized to be useful. The occurrence of the word 'York' produces several clusters concerning York County in Pennsylvania. Cluster 8.4 and 8.5 concern newspapers, while cluster 8.6 concerns radio stations.

Clusters 9.1–9.5 are generally concerned with the situation in Kashmir and its connections. Cluster 9.1 concerns Indian, Pakistani, and Afghan terrorism. Cluster 9.2 focuses on terrorism in Kashmir. Cluster 9.3 concerns terrorism in the wider Asian context. Cluster 9.4 concerns the nuclear non-aligned movement, as a result of the development of nuclear weapons by both India and Pakistan. Finally, cluster 9.5 concerns that likelihood that Pakistan will follow the path of Iraq towards rogue statehood.

How effective would the content of these clusters have been at guiding decision making in the immediate aftermath of the September 11th attacks? Almost certainly, most of this content was available, in some form, to intelligence organizations. The structure imposed by textscAthens might have been suggestive about the relative importance of various aspects. For example, it is clear from these results how widespread the connections were between al Qaeda and Islamic terrorism groups and factions in other settings (for example, Salafist terrorists who are non-Arabs, or are geographically remote from the historical center of Islamic terrorism); and how weak the connections between al Qaeda and Iraq were. These results might also have provided further evidence for the importance of bin Laden as a guiding hand behind many trends that seemed, at the time, to be unconnected. It is certainly the case that the clusters collected here provide a fairly complete primer on Salafist Islamic terrorism that would have informed, for example, the media of the scale of the problem in the immediate aftermath of the attacks.

What is missing is any connection between al Qaeda in its Middle East incarnation, and Salafist terrorist groups in Europe, including the related groups in countries such as Algeria and Morocco. This appears to be justified by the scarcity, at that time, of any web content making connections

between these two subgroups; such pages as do exist are mostly about the Balkans (which are contained in Clusters 5.1–5.4).

4 Conclusions

Intelligence analysts need tools that allow them to work with large information spaces, and which help them to break out of preconceptions to consider a larger fraction of the hypotheses that the available data may support. Exploratory data mining can help with the second problem, but is not directly useful for information repositories such as the world wide web.

The ATHENS system is designed to address both problems by piggybacking on existing tools such as search engines to fetch appropriate subsets of the huge available data; and by using contextualized searches to go beyond the limitations of an analyst's existing knowledge.

We have demonstrated the use of the system by generating the knowledge it would have produced if started from "al Qaeda" and "bin Laden" on September 12th, 2001. The results demonstrate the effectiveness of the system, and are also of some inherent interest.

Software: The ATHENS system is available from `www.cs.queensu.ca/home/skill/athens.html`.

References

[1] W.E. Baker and R.B. Faulkner. The social organization of conspiracy: Illegal networks in the heavy electrical equipment industry. *American Sociological Review*, 58:837–860, December 1993.

[2] British National Corpus (BNC), 2004. **www.natcorp.ox.ac.uk**.

[3] T. Coffman, S. Greenblatt, and S. Marcus. Graph-based technologies for intelligence analysis. *CACM*, 47(3):45–47, March 2004.

[4] J. Corbin. *Al-Qaeda: In Search of the Terror Network that Threatens the World*. Thunder's Mouth Press, 2002.

[5] United States Government. *Final Report of the National Commission on Terrorist Attacks Upon the United States*. 2004.

[6] M. Granitzer, W. Keinreich, V. Sabol, and G. Dosinger. WebRat: Supporting Agile Knowledge Retrieval through Dynamic, Incremental Clustering and Automatic Labelling of Web Search Results. In *Twelth IEEE International Workshops on Enabling Technologies: Infrastructure for Collaborative Enterprises (WETICE'03)*, 1080–1383, 2003.

[7] R. Gunaratna. *Inside al Qaeda*. Berkley Publishing Group, 3rd edition, 2003.

[8] D. Jensen and J. Neville. Data mining in social networks. Invited presentation to the National Academy of Sciences Workshop on Dynamic Social Network Modeling and Analysis, November 2003.

[9] H. Liu. MontyTagger v1.2, 2003. **web.media.mit.edu/hugo/montytagger**.

[10] R. Popp, T. Armour, T. Senator, and K. Numrych. Countering terrorism through information technology. *CACM*, 47(3):36–43, March 2004.

[11] K. Popper. *The Logic of Scientific Discovery.* Hutchinson, London, 1959.

[12] M. Sageman. *Understanding Terror Networks.* University of Pennsylvania Press, 2004.

[13] D.B. Skillicorn. Detecting related message traffic. In *Workshop on Link Analysis, Security and Counterterrorism, SIAM Data Mining Conference*, pages 39–48, 2004.

[14] D.B. Skillicorn. Finding unusual correlation using matrix decompositions. In *Second Symposium on Intelligence and Security Informatics*, 2004.

[15] K. A. Taipale. Data mining and domestic security: Connecting the dots to make sense of data. *Columbia Science and Technology Law Review*, 2, December 2003.

[16] J.R. Tyler, D.M. Wilkinson, and B.A. Huberman. Email as spectroscopy: Automated discovery of community structure within organizations. HP Labs, 1501 Page Mill Road, Palo Alto CA, 94304, 2003.