Improving Analysis with Information Extraction Technology

Information Superiority/Information Operations

Sarah M. Taylor, PhD
Principal Systems Engineer
Global Information Management Systems
Lockheed Martin Management and Data Systems
3201 Jermantown Road
Fairfax, Virginia 22030
E-mail: Sarah.M.Taylor@lmco.com

Abstract

Information Extraction is an emerging technology with the potential to transform the way textual information is analyzed and used for many information analysis tasks. However, as a relatively new technology, its best uses have not yet necessarily been discovered. This paper provides some background on Information Extraction and discusses several effective uses of the technology from recent Lockheed Martin experience in applications of its AeroText^{TM} software. The paper provides examples of the ways in which different visual organizations of extracted information can improve analysis. Information Extraction enables better analysis by allowing the analyst access to a far greater depth of information than he could read on his own, and by providing the analyst with the capability to view the data in multiple organizations and formats, thus increasing his ability to see trends and to discover the unexpected.

Background

The Information Age challenges us with the effective presentation of the large volumes of information that people must absorb in order to perform their jobs. This issue exists across the board, in many occupations. Data must be consolidated and organized so that the consumer of information can quickly understand and act upon it. Data presentation is a fairly obvious problem, perhaps, when the data in question involves hundreds or thousands of numeric readings from which someone must decide whether to buy or sell stock, determine if a certain subsystem of an aircraft is operating properly, or gauge the progress of supplies and materiel through a distribution system.[1]

However, the problem of how to present data, which is frequently text rather than numeric, to intelligence and information analysts[2] so that they can exploit it most effectively is perhaps less well understood. There is certainly a widespread understanding that the analyst of today must deal with an enormous volume of text, since we hear frequently of the need to exploit large volumes of textual data.[3] Yet we do not fully know today how to present the data contained in text in ways that maximize the opportunities for discovery and understanding.[4] One common intuition favors various

[1] A tantalizing array of visualization tools is available for already structured data and applicable to business, technical, and scientific data; see, for example, the visualizations available with Matlab (Mathworks.com) or from Advanced Visual Systems.

[2] The use of "analyst" throughout this paper refers to any kind of information analyst relying on data and information derived from text. Intelligence analysts, of course, are important examples, but such analysts exist throughout law enforcement and the commercial realm also.

[3] Two indications of the pervasiveness of this idea in the popular culture: a search on the term "information overload" (in quotations) in Google claims roughly 160,000 citations; a quote from the business-intelligence-world promotional material for its 3rd annual conference, "Data is growing exponentially every day and information overload is a reality for most businesses today…..every intelligent business needs to make strategic use of analytic tools to empower its knowledge workers." (http://www.ccworldnet.com/2003/biw_AU)

[4] There are visualizations available for text information, notably those developed at Pacific Northwest National Laboratory, such as the well known Starlight, see www. starlight.pnl.gov, as well as a number of visualizations provided with text analysis tools, such as Semio, see www.entrieva.com. These visualizations are useful for finding documents that discuss similar topics and for finding documents where

forms of “link analysis” - if we could just draw lines between all the related data items in very large data sets, we would come up with interesting, vital, perhaps critical links that were otherwise buried unseen amongst the original pile of documents. However, believing in the analytic power of these kinds of links and making them visible and useful, in practice, are two rather different matters.[5]

The first difficulty in automatically presenting information from text to the analyst in summarized, consolidated or organized formats is that the original text form itself is not particularly useful. From the analyst’s point of view, the information in the text that he wants is likely to be buried within large amounts of material that is irrelevant to his task. His problem is not a matter of simply finding the right documents (a difficult enough task sometimes) but of finding the right items in the documents. Additionally, if the analyst wishes to quantify his analysis at all, understand patterns of relationships, or understand the ways in which events have unfolded over time and through space, he will have to somehow extract the items of interest from the text and put them in a consistent form. For computer systems to be able to show the information to the analyst in any graphic way, or to be able to move easily from one visual presentation to the next, the information needs to be in a structured format – discrete items, consistently labeled, of consistent lengths and data types, arrayed in spreadsheets, databases, tables and the like.

Focus on Benefits of Information Extraction

Information Extraction technology is one of a number of emerging technologies that enable the presentation of information developed from text in a variety of graphic formats.[6] This paper recounts some concrete examples of ways we at Lockheed Martin have been able to use our Information Extraction tool, AeroText™, in conjunction with several analysis and presentation tools,[7] to actually improve both the speed and depth of the analysis our users perform. Although the basis of our experience is AeroText, the lessons we can draw from these examples should be applicable to the integration and use of other Information Extraction tools. While we believe our tool to be the best, demonstrating that point is not the purpose of this paper. The purpose of this paper is to describe some of the ways effective presentation of information from text, enabled by Information Extraction, can improve analysis. The ability to use IE to strengthen analysis is not the dream it was ten years ago. All but one of the examples I use below draws

concepts of interest co-occur. As such, they are typically useful at the beginning of a discovery process, and presume the analyst will either read the documents or otherwise process them further in order to use the specific information provided by the documents. However, it is the effective display and understanding of the specific information contained in the multiple documents that I am referring to here.

[5] See, for example, the concerns and needs expressed in relation to intelligence requirements, Shannon Henry, “In-Q-Tel, Investing in Intrigue” *Washington Post*, July 1, 2002, p. E01.

[6] Phrase identification/extraction and Subject/Verb/Object identification are two other technologies that can be used to put information from text into structured formats. Various methods, usually statistically based, for text clustering can also be used to present high level views of relationships between texts.

[7] We have experience with using AeroText in conjunction with databases, link analysis, cluster displays, and geographic information systems. The specific analysis tools used should not affect the conclusions of this paper.

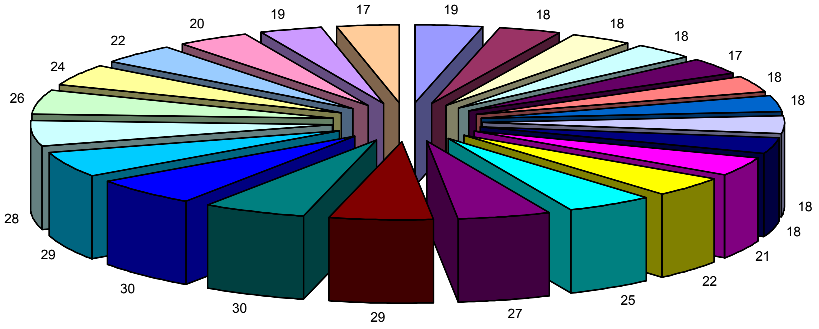


Figure 1: Temperatures – One day in February (There are 24 wedges, one for each hour of the day; the numbers are the temperatures for each hour. The size of the wedge represents the share of that temperature in a total of all temperature readings for the day.)

from our experience with operational systems, although the data I use to illustrate these examples has been changed to protect our clients' confidentiality.[8]

Presenting Data

I begin with a quick illustration of the way in which presentation can influence how an analyst understands the data. This illustration uses weather data – a listing of temperatures and sunrise/sunset times for a week in the month of February in the Virginia suburbs of Washington, DC.[9] This rather straight-forward numeric data provides a useful introductory example because it is something for which most people already have anintuitive understanding. Thus, the unfortunate effects of displaying the data ineffectively are readily understood.[10]

First we'll look at a slice of this data in a summary format that is not particularly useful. Figure 1 shows the contribution of each temperature reading (one every hour of the day) to a total of those temperatures for the day. Each wedge is labeled with the hour of the day for which the reading was taken. Of course, the reader knows instinctively this graphic is wrong for this kind of data, at least for most purposes. Temperatures rise and fall during the day, but this presentation does not show us that activity or the continuous nature of the data. A similar graphic presenting several days worth of data would be even

[8] The collection and use of the example data is fully described in the Appendix. The processing of this data for this paper involved some manual steps that would ordinarily be automated in a large system supporting numerous analysts. These were largely what I would call data transfer steps – such as moving data from MS Excel to MS Access. The key reasoning operations – information extraction itself, and the use of database queries to analyze those extractions – were entirely automated. I did not manually improve on the extraction results or perform any discovery that could not be done with automated tools. In fact, although the document set is only 250 documents, the type of analysis I used would have been extremely laborious and time consuming to perform manually.

[9] See the Appendix for a description of how this data was collected.

[10] Edward R.Tufte, The Visual Display of Quantitative Information, Graphics Press, Cheshire, Connecticut, 1983, discusses in detail the ways in which excellent graphics reveal the meaning in the data and has fine examples of how a graphic can reveal multiple trends at the same time, such as the Charles Joseph Minard graphic depicting the fate of Napoleon's army in Russia (p. 40). See also Nahum Gershon, Presenting Visual Information Responsibly, SIGGRAPH Computer Graphics Newsletter, Vol. 33 No. 3 Augaust 1999; http://www.siggraph.org.

more difficult to use. However, a standard line graph works well for a single day or multiple days.[11]

With this presentation, shown in Figure 2, we can readily see the typical daily pattern: how the temperature falls steadily from midnight to a low point in the early morning, rises to a high point in the mid –afternoon and then again begins to fall. We can also see the anomalous line, for this data, which continues to rise after the normal mid-afternoon peak, telling us that something different happened that day - the arrival of a warm front. To drive the point home, we can add further data of another type to this chart. By juxtaposing new information with the temperature information, we can understand the behavior of the temperature even more clearly.

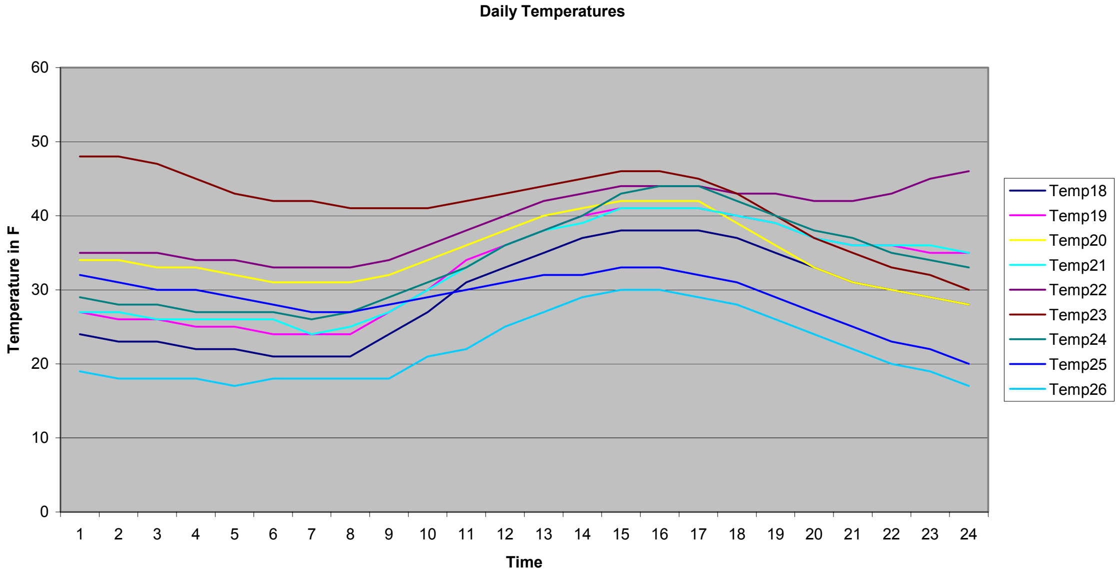


Figure 2: Temperature – Nine days in February

[11] The original spreadsheet can also be formatted to be useful. See the Appendix, p 23.

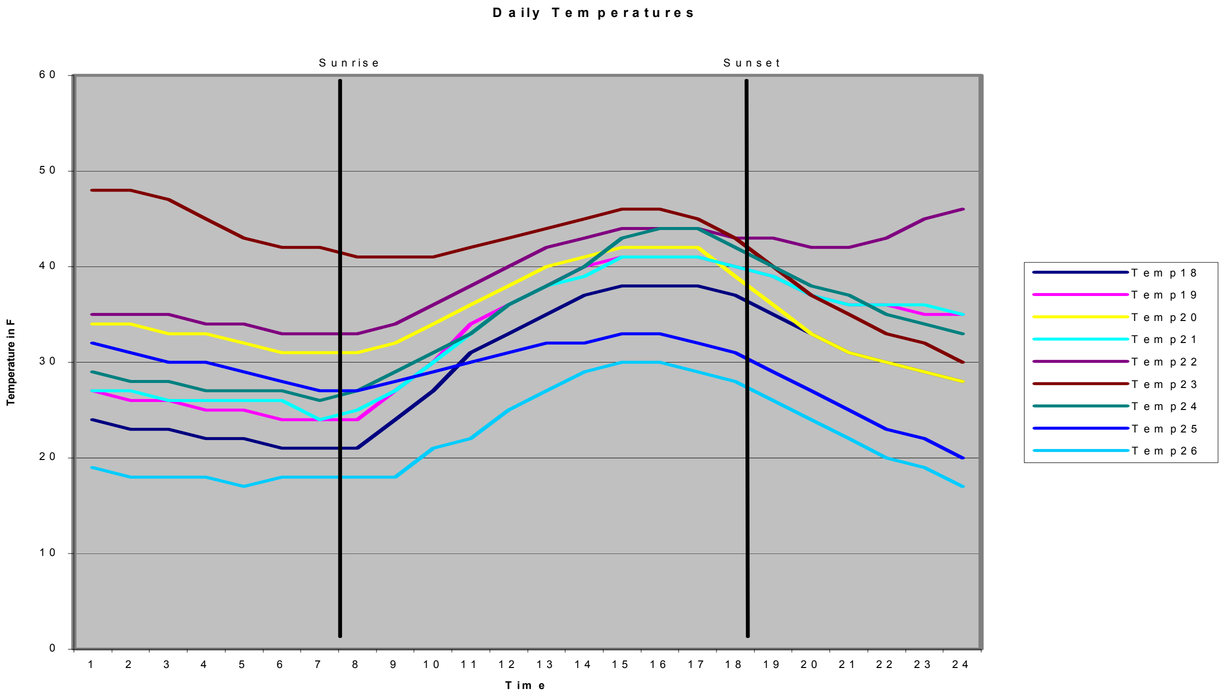


Figure 3: Temperature and Sunrise, Sunset Times – Nine Days in February

With the data presented as line graphs, and the addition of the approximate times of sunrise and sunset, as in Figure 3, we can now see the effect of the sunrise and sunset on the daily pattern that week in February, with the lowest temperatures normally occurring right before dawn and lasting for a short time after, and the warmest couple hours of the day starting about seven hours after sunrise and ending about two hours before sunset.

From this basic illustration, we can draw a number of simple preliminary conclusions about data presentation for analysis. The presentation has to suit both the data and the kind of information the analyst needs to get from the data. The pie chart was a poor choice for the temperature data for a number of reasons, not the least of which is that the pattern of daily temperature change, so obvious from the line graph, does not show in a pie chart, which displays shares of a total. Changing the way data is presented can make a major difference in what patterns the analyst can see in the data. Furthermore, being able to combine or compare data of different types can help the analyst see even more patterns.

However, combining and presenting information in multiple graphic formats, if the original data is buried in text, requires the text be first processed and the information of interest stored in a structured format easily manipulated by analysis applications. That is why today, Information Extraction technology, and some other kinds of Natural Language processing, hold such promise for the improvement of many areas of text data analysis.

Information Extraction: a quick overview

Information Extraction (IE) technology, as we know it today, really began with the Message Understanding Conferences (MUC) established in the late 1980's under DARPA sponsorship. Several experimental Information Extraction systems existed at the time. However, it was the focus on testing and evaluation brought by the MUC and the concentrated effort of the MUC leaders to understand and define the types of tasks for which Information Extraction might be useful to the Government that really started IE systems down a path toward operational utility. In the early through mid-1990's, Information Extraction research was heavily supported by the DARPA Tipster Program.[12] Since that time, Government sponsorship of IE related research has been focused by ARDA, the DARPA TIDES and EELD programs, support from the National Science Foundation, and continued evaluations managed through the NIST.[13]

Commercialization of IE technology began in the late 1990's with the launching by InXight of the ThingFinder product from technology originally developed by Xerox PARC. At roughly the same time, SRA began commercial applications of their NetOwl extractor. Lockheed Martin's Information Extraction technology, AeroText, was launched in February of 2001, although Lockheed Martin's experience with IE extends back through the decade of the 1990s and joint work with General Electric under the Tipster Program. Within the last few years a substantial number of other Information Extraction tools have entered both the commercial and government markets. Lockheed Martin now counts approximately two dozen direct IE competitors.

Information Extraction technology automatically skims through text and identifies predefined types of information. For example, it can find all the person names in a text, or all the organization names. IE can also determine relationships, such as which people are members of which organizations, according to that text. At its most complex level, it can also determine events described in text, such as merger or acquisition activities among firms, the launching of new products, or the movements of military units. An IE application is usually configured to mark this information, or tag it, with the information type, and often location in the text, so the information then becomes available for further processing, human or automatic. A typical use of Information Extraction is to "extract" this information and put it in a database format; so, for example, text documents, describing in natural English language, the movement of various corporate executives from one position to the next in different companies, can be converted into a structured database with fields such as: Executive Name, Old Company, Old Position, New Company, New Position, Type of Transfer, Date of Transfer.[14] Such databases have

[12] For more detail on the history of the Message Understanding Conferences, see Ralph Grishman and Beth Sundheim, "Message Understanding Conference – 6: A Brief History", in the *Proceedings of the 16th International Conference on Computational Linguistics,* Copenhagen, June 1996, available at http://acl.ldc.upenn.edu/C/C96.

[13] ARDA – Advanced Research and Development Activity; DARPA – Defense Advanced Research Projects Agency; EELD – Evidence Extraction and Link Discovery; TIDES – Translingual Information Detection, Extraction and Summarization; NIST – National Institute of Standards and Technology.

[14] Corporate executive transfers were the subject of the MUC-6 Scenario Template task, Fall 1995. A brief description of the task can be found at http://www.itl.nist.gov/iaui/894.02/related_projects/tipster.

obvious utility in analysis, not the least of which is they enable various kinds of presentation, such as link analyses, timeline presentations, geo-location, and graphing of trends. But Information Extraction is also used today to support and enhance other kinds of text handling applications such as Information Retrieval, Question Answering, Text Categorization, and even Machine Translation of Natural Languages.

Information Extraction technology recognizes predefined types of information in text, without the specific instances of those types having to be defined ahead of time. It is not searching solely for known strings. Rather, it can recognize the string "Northern Virginia" as the designation of a place, although it may never have encountered that phrase previously. Software that extracts items solely on the basis of lists or of preexisting tags, such as web page HTML, which have been manually supplied, cannot be considered true Information Extraction.[15] Systems today use one of three strategies for finding the specified information in text – human developed patterns to find key information, based on grammatical structures and domain expressions, patterns of key information automatically learned from manually tagged examples in the text, and hybrid systems. Most current systems are in fact based almost entirely on human developed patterns.[16] Some of these, including Lockheed Martin's, are experimenting with automated learning in combination with manually developed patterns and can be considered hybrid.

On the whole, IE technology has reached levels of accuracy and adaptability to new tasks that make it useful in operational settings. Out of the box accuracy levels for simple entity tagging of English (finding the people, places, times, and organizations, for example, in open source news text) is fairly high, reliably above 80%, often above 90% in good systems, certainly sufficient to form a basis for many kinds of quantitative analysis.[17] Tagging is readily available in a range of languages.[18] However, tasks that are more complex – such as the example cited above, of tracking changes in high corporate office holders – require several months of tuning the patterns before they can be accomplished at all, and may stay in the 60% - 80% accuracy range for several months of effort, before they can be improved beyond that point.[19]

[15] Of course, systems that use patterns may also use lists. It is common for systems to make use of first and last name lists, as well as gazetteers, as information for their pattern knowledge bases.

[16] Of our two dozen competitors we know of only three with substantial or complete reliance on automated pattern generation.

[17] For example, MUC-7 Proceedings report scores of 14 systems on the named entity task; these include both mature and new systems. The range of F scores was 69.67 to 93.39, with 79% (11) of the systems achieving scores F scores of over 80. *Message Understanding Conference Proceedings*, MUC-7, 1998, available at www.itl.nist.gov.

[18] AeroText has current capabilities in four languages (Chinese, Japanese, Spanish, Arabic) in addition to English. However, the software is language independent and can be operated in Unicode, so that knowledgebases can be built to extract from any language, in the original language.

[19] Accuracy levels for the Scenario template in MUC-7 ranged from F scores of 42.09 through 50.79 (eliminating the score of one very immature system). (MUC-7 Proceedings, 1998.) However, Lockheed Martin work on operational projects since 1998 suggests that a 60% - 80% accuracy range, at least, is achievable on tasks of this complexity with a longer development effort, and starting from the higher accuracy baselines available today.

Likewise, the type of text being analyzed makes a considerable difference to the speed with which accuracy above the 80% mark can be achieved. Text which is straightforward, reasonably predictable in its structure and well edited is considerably easier to work with than other kinds. Difficult text poses problems such as:

- errors and omissions introduced by poor quality originals and a scanning and ocr process;
- inconsistently spelled names, introduced, for example, when names are transcribed into English from non- Roman alphabet languages;
- poor proof reading and typographical errors;
- loss of document formatting from a document conversion process;
- telegraphic writing styles, in which subjects or verbs may be omitted;
- styles relying on many parenthetical and parallel grammatical constructions;
- frequent use of semi-formatted structures, such as outlines, bullets, and lists, especially if they vary from document to document;
- poor grammar, such as incomplete sentences or unclear pronoun references.

These issues are all mostly addressable today, but they complicate the pattern development process and can make systems built to handle text with these characteristics somewhat time consuming to develop.

Use of Information Extraction in Operations Settings to improve Analysis

Lockheed Martin has, over the past few years, deployed a number of systems that use Information Extraction to provide analysts of many subject domains with data from text in a "ready to analyze" format. Through these systems we are helping to move analysts away from the "old" way of doing analysis to a new way. That is not to say that the old way must be thrown out; it is still necessary for many kinds of analytic tasks. But the "new way" is proving more efficient and beneficial for certain tasks, especially those of two types:

- Tasks requiring the rapid location of specific types of entities or events in very large volumes of data, and the subsequent analysis of those, such as compilation of summary statistics, or analysis of trends in the resulting data.
- Tasks requiring the identification of relationships of specific types between entities, such as the association of people with certain kinds of organizations, again in large volumes of data.

In the "old" way of approaching these tasks, an analyst would have available in most cases an Information Retrieval system, office tools, and perhaps in some cases a tool such as a Geographic Information System (GIS) or a link analysis tool, which worked with a database and could be used to query, display and manipulate information in a database. However, there was no automated way to transfer information from the text into either database or spreadsheet. Specifically, if an analyst wanted to know, for example, how many CEO's had resigned from corporations in 1998 compared with 1999, he would construct queries to his document store to retrieve documents from 1998 which had words like CEO, corporate executive, president of company, and resign, left, moved on, and so forth. The returns from his queries were documents, which he would himself skim, and if his query was reasonably good, he would find among them many examples of presidents and CEO's of companies who had resigned or moved to other positions. He would then manually record these examples in whatever tool he was using – highlighted printed copies, index cards, spreadsheets, databases, text document lists or tables - and count the number of resignations, manually or with a spreadsheet or database function. Then he would perform the same process for 1999 and he could compare the two numbers. The process is slow, laborious, and error prone, so much so as to prevent most analysis of this type from getting done, unless the topics in question are of extremely high priority, and the analytic question of proven utility. This "old way" discourages trial lines of analysis because the time and effort wasted, if the answer does not prove critical, is just too great.[20]

In the "new" way of tackling such a task, a system is built in which the Information Extraction component selects the documents to be analyzed (sometimes through its own functionality, sometimes in connection with queries built in a Retrieval or Document distribution system), pulls out the required information and puts it into a database, spreadsheet or other structured format. This information can then be analyzed, as is, errors and all, if desired, for a "quick and dirty" answer to the question of whether it looks like more CEOs left their positions in 1998 or 1999. Or if an accurate count is required, the analyst or other personnel, can clean up the data, that is, correct for the mistakes of the Information Extraction system, and the analysis is then re-run.

This new overall process is not instantaneous, due to the system development time required, including pattern development, and any time required for data cleanup. More time is required, primarily for pattern development, the more complex the extractions are that need to be made. However, it is demonstrably more efficient than the "old" way for large volume tasks: (1) the substance of far more data can be reviewed; and (2) the burden of the work is shifted off the analyst, and away from the time period when the analysis work is being done (often a crisis), to work that can be done by system developers in a largely preparatory stage. If designed properly, the databases are then ready for the analysis when they are needed and should be flexible enough to handle major analytic questions in the subject domains for which they have been prepared.[21] Additionally, because the data is waiting in the database, this more automated process can encourage exploratory analysis and it enables the use of statistical and visualization tools to support that analysis. These features both make the analyst more efficient, and support more accurate analysis.

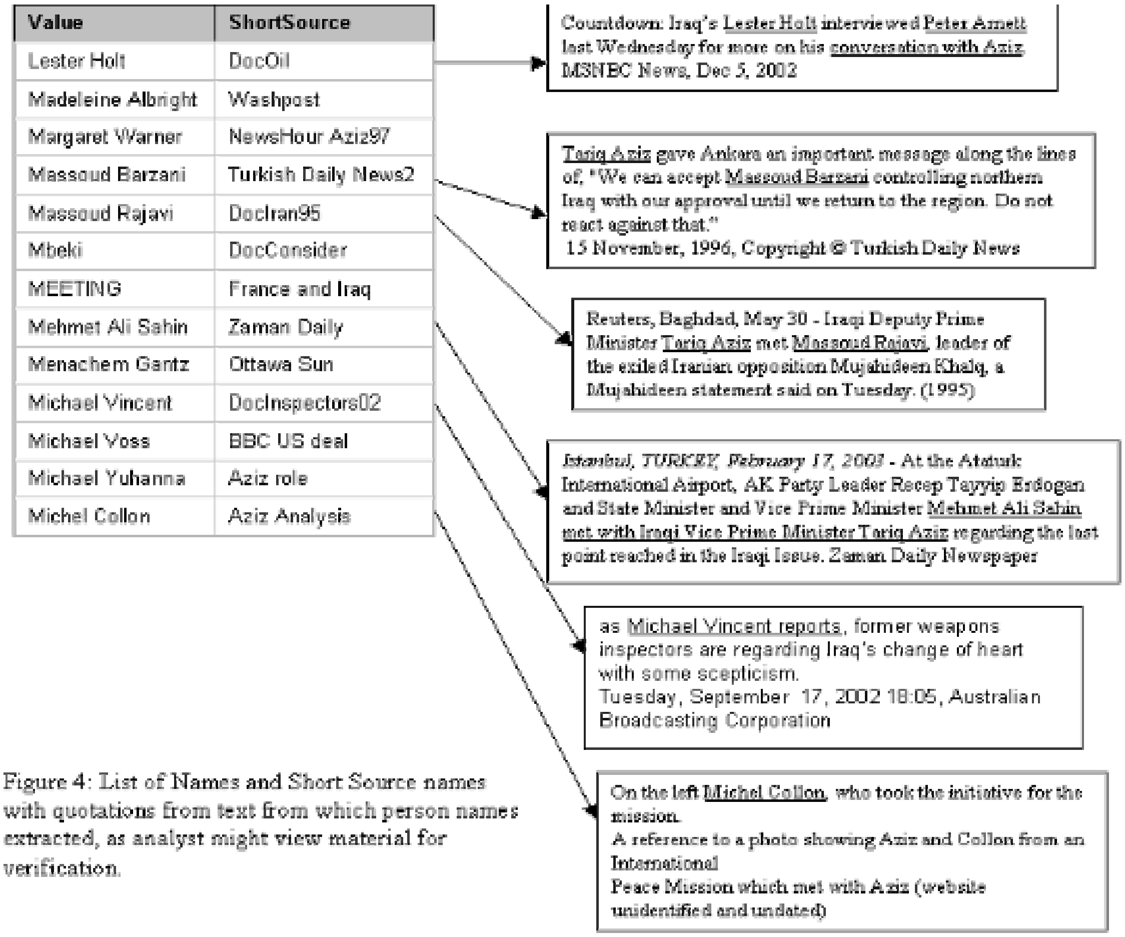
[20] Douglas MacEachin, in his Foreword to Richard Heuer's book, *Psychology of Intelligence Analysis*, Center for the Study of Intelligence, Central Intelligence Agency, 1999, points out the problem that "…too often, newly acquired information is evaluated and processed through the existing analytic model, rather than being used to reassess the premises of the model itself." We are discussing here the technical underpinnings of that same problem. MacEachin and Heuer focus on the human thought processes, which cause people to be reluctant to reassess their assumptions, their premises, their interpretations of data. However, without considerable automated support there are also practical hurdles to overcome.

[21] How one designs such a system, particularly the database, so that it is flexible enough to handle the questions that need to be asked and so that the information is ready for the analyst when he needs it, is of course a systems engineering task for any system development.

Lockheed Martin and its customers have had success with Information Extraction in several roles supporting analysis. I discuss and illustrate these below. In all the examples given, the Information Extraction component was part of a larger, integrated system in which the analyst had access to several tools, in which information developed by the Information Extraction component was automatically loaded into a structured form, such as a database, and in which the analyst had easy access, linked back from the extracted data, to the textual source of the extraction.

Presentation of hypotheses for review

A presentation of extracted information to the analyst for review is perhaps the most basic application of Information Extraction. It simply replaces an entirely manual step in the "old" analysis process, where the analyst extracts information himself into a structured format. For example, if the analyst is trying to develop a list of people who may have met with Tariq Aziz, using a very simple application of IE with only entity extraction, he could get a list of names of people from documents concerning Aziz[22], and then check each name against the text to see if there indeed was a meeting.



Value	ShortSource
Lester Holt	DocOil
Madeleine Albright	Washpost
Margaret Warner	NewsHour Aziz97
Massoud Barzani	Turkish Daily News2
Massoud Rajavi	DocIran95
Mbeki	DocConsider
MEETING	France and Iraq
Mehmet Ali Sahin	Zaman Daily
Menachem Gantz	Ottawa Sun
Michael Vincent	DocInspectors02
Michael Voss	BBC US deal
Michael Yuhanna	Aziz role
Michel Collon	Aziz Analysis

Countdown: Iraq's Lester Holt interviewed Peter Arnett last Wednesday for more on his conversation with Aziz. MSNBC News, Dec 5, 2002

Tariq Aziz gave Ankara an important message along the lines of, "We can accept Massoud Barzani controlling northern Iraq with our approval until we return to the region. Do not react against that."
15 November, 1996, Copyright © Turkish Daily News

Reuters, Baghdad, May 30 - Iraqi Deputy Prime Minister Tariq Aziz met Massoud Rajavi, leader of the exiled Iranian opposition Mujahideen Khalq, a Mujahideen statement said on Tuesday. (1995)

Istanbul, TURKEY, February 17, 2003 - At the Ataturk International Airport, AK Party Leader Recep Tayyip Erdogan and State Minister and Vice Prime Minister Mehmet Ali Sahin met with Iraqi Vice Prime Minister Tariq Aziz regarding the last point reached in the Iraqi Issue. Zaman Daily Newspaper

as Michael Vincent reports, former weapons inspectors are regarding Iraq's change of heart with some scepticism.
Tuesday, September 17, 2002 18:05, Australian Broadcasting Corporation

On the left Michel Collon, who took the initiative for the mission.
A reference to a photo showing Aziz and Collon from an International
Peace Mission which met with Aziz (website unidentified and undated)

Figure 4: List of Names and Short Source names with quotations from text from which person names extracted, as analyst might view material for verification.

[22] The query used to select these names lists all people names found in close proximity to Aziz in the sample documents.

Figure 4 shows a sample from the list of 460 instances of person names closely associated with Aziz in the sample document set. The figure illustrates what an analyst might see in an actual system with an integrated Information Extraction capability – a list of person names, to be thought of as "hypotheses" that each of these people met with Aziz in that time period. Some names are obvious errors, or may be of no interest, and the analyst can eliminate them without even checking the text. But the remainder of the names he checks by "clicking" on the name and having the relevant piece of text appear (illustrated in the figure by the text boxes on the right), from which in a properly set up system he can then cut and paste, or even drag and drop, corrections.

The improvements for the analyst are several:

- He does not have to read all the text. In most cases he only has to review the sentence or two near the extraction to find the information he requires. Our experience is that there will be some items that are difficult to check, because of ambiguities in the text, but these are a relatively small number instances.
- What is already correct is also already in the format required by the task, so it does not have to be retyped or otherwise copied into the format, reducing the possibility of human error. To save time on those items that must be corrected, the system has to be designed to make correcting as simple as possible.
- The task has become a more passive "review" task, rather than an active "find the information" task. The analyst, in fact, has far fewer decisions to make in the passive task, than in the active task. If he is actively reading all the text, even skimming, he is constantly processing, for each phrase, is this item relevant to my task, yes, or no. For those who are easily distracted, it also requires constant effort to stay on task. However, the use of IE reduces both the distractions and the number of relevance decisions the analyst must make down to the number of names on the list, far fewer than the number of words and phrases in the text. So, the task is actually cognitively easier, once the analyst has adapted to it.
- The task can be managed more effectively since there are a known number of items to be reviewed (rather than perhaps a known number of pages to be read resulting in an unknown number of items to be validated.)

We have used such a system for a well-defined analysis task and recorded metrics for two different analysts. Both used the same analysis platform, with integrated retrieval and information extraction capabilities. However, one relied heavily on the available extractions, following a verification process like that already outlined. The other analyst focused her method almost solely on the retrieval system, looking for relevant documents, and then reading them more or less in their entirety. Both analysts addressed the same task and the same document set. The differences in results were dramatic, shown in Table 1.

Metric	Analyst 1 (used IE)	Analyst 2 (without IE)
Hours spent	48	200
Entities Verified	350	200
Pages used	2601	Unrecorded Estimate - 2600

Table 1: Comparison of analyst productivity - with and without reliance on Information Extraction

Other similar applications, in use by customers, have reduced data preparation time, in one case, by a half, and in another from "days to hours". These applications provide substantially reduced data sets[23] for the analysts to correct, as they extract events and relationships, not just entities.

The issue that inevitably arises when these types of applications are discussed is: since the analyst is only checking what is put before him, and not looking further, what about the instances missed by the system and how does the analyst find them? To express the problem in terms of the Aziz data, we need to know if we miss any recorded instances of people meeting with Tariq Aziz by depending upon the IE system to pre-screen the text for us. In this sample, based on a review of 10% of the documents, I estimate an analyst would have missed between 5% and 16% of the meetings referenced.[24]

The issue just described is, of course, the issue of Recall. Even well tuned IE components do not have perfect Recall, although as indicated above Recall levels are now fairly high for well understood tasks.[25] However, the inability to expect perfect

[23] Reduced, that is, from the volume of text in which those relationships and events are found. But also reduced in comparison to the volume of data that can be expected in a simple case of entity extraction, like the example of persons potentially associated with Aziz described previously, where no attempt was made to *automatically* determine whether the people had met Aziz or not. That step was left to the analyst.

[24] In the first 25 documents (listed in alphabetical order by file name) I found 19 instances of meetings between Aziz and a named person. 13 of these (68%) were found by the extraction method exactly as I found them. (Note that the extraction method for finding a possible meeting was based on proximity of the name of the person to Aziz's name in the text, not on a more complex set of patterns, which would be the more usual method.) Of the remaining 6 instances, 3 - while missed by the system in these first 25 documents- were found in other documents in the total sample set. The remaining 3 instances were missed completely. Of these three, two were interviews with Aziz by newspaper reporters who were only designated by their by-line at the top of the article, and thus might be considered special cases. If they are counted, the missing amounts to 16%, if not, the missing amounts to 5%.

[25] Additionally, measurement of Recall is not a totally cut and dried business. For example, when IE systems are evaluated for "filtering" in MUC, that is determining which documents have relevant information to the task at hand, they tend to score much higher on filtering Recall than Recall for the template as a whole (see *MUC-7 Proceedings*, Scenario Template results). From a practical systems implementation point of view, this means that if the analyst is checking the results anyway for accuracy at a detailed level, he will be almost certainly viewing most of the documents that contain the kinds of information he wants to look at. (In MUC-7, filtering scores ranged from 68% to 92%). Because extraction tasks frequently involve finding relationships or events, which are made up of multiple entities, there is also a high probability that the analyst will be directed toward the correct portion of a text, even if not all the

Recall of IE applications has to be balanced against the undeniable fact that even highly skilled analysts do not have perfect Recall either. One could speculate analysts' Recall drops as the volumes of material they scan get larger, because of fatigue, boredom, and confusion, while, an application will perform consistently over even larger volumes of data. Additionally, to an as yet unknown extent, the inaccuracies of an automated system can be off-set by the fact that the system can process vastly greater amounts of material than the human. Most information sources are highly redundant, both within and across documents, so that what is not picked up automatically from one source can still be recovered from another.[26]

Exploring trends in large data sets

A second way in which we have found Information Extraction to be of practical value is in exposing trends in large text data sets. Finding such trends is impossible for the analyst to do on his own, except under the most exceptional circumstances, due to the extremely long time (years, decades)[27] required to manually extract the data items from the text. Additionally, the discovery of any unusual trends depends upon the analyst's ability to ask many questions of the data, based on his intuitions about what the data might be able to tell him, until he identifies trends of significance to his task. This discovery process requires structured data in a data base format. Because of the difficulties of preparing the data, trend analysis of information from text is typically not done without automated support. The availability of IE provides the capability to the analyst to address many more issues from a quantitative view than has been possible to date, thus significantly augmenting the arsenal of available analysis techniques.

To illustrate using the example data set: we might like to know the places of interest to Aziz, over a period of years covered by the data. Are there, for example, any significant differences from one year to the next? One could postulate that there might be, and that these differences could reflect differences in Aziz's responsibilities or in Iraqi foreign policy. Queries for places associated with Aziz in each year produce a set of comparisons shown in the following table.[28]

correct entities have been found, because at least some of them are likely to be found. However, in strict testing Recall scores will not capture this effect.

[26] This statement is based on observation during the marking of text and the error analysis of IE system output. It is certainly supported by the material in the Appendix, Table 5; however, I have not yet had the opportunity to substantiate it in a more thorough way.

[27] Extracting information manually on one topic at an uninterrupted rate of 10 pages an hour (this would allow there to be considerable irrelevant information in the text) an analyst could theoretically review and extract information from about 18,000 pages (or possibly 270Mb, at 15Kb per page) per year on that one topic. That would be equivalent to extracting the information from all the documents returned from one moderate sized Google search.

[28] This table has been edited to remove mistakes produced by the extraction, to keep names of cities and countries only, to eliminate references to places in Iraq, and to use only one name for each entity (U.S. for U.S. and USA, United States, etc.). The raw table is in the Appendix, Table 6. The query is again based on proximity: closeness of occurrence within the text is relied on as an indicator of possible association between Aziz and the place.

98	99	00	01	02	03
Afghanistan	Amman	Amman	Afghanistan	Ankara	Afghanistan
Amman	Beirut	Assisi	Britain	Assisi	Ankara
Assisi	China	Beirut	China	Bahrain	Assisi
Bahrain	Egypt	China	Israel	Britain	Bahrain
Britain	Iran	Damascus	Kuwait	China	Britain
China	Italy	Israel	Moscow	Damascus	Cairo
France	Jordan	Jordan	New York	Egypt	China
Geneva	Lebanon	Kuwait	Pakistan	France	Damascus
Iran	South Africa	Lebanon	Russia	Germany	Egypt
Israel	Syria	Moscow	South Africa	Hollywood	France
Italy	U.S.	Prague	Syria	Iran	Germany
Johannesburg	U.A.E.	Rome	U.S.	Israel	Hollywood
Jordan	Washington	Russia		Italy	Iran
Kuwait		Syria		Johannesburg	Israel
Lebanon		U.S.		Jordan	Italy
Moscow		Washington		Kuwait	Johannesburg
New York				Lebanon	Jordan
Pakistan				Marrakesh	Korea
Paris				Morocco	Kuwait
Russia				Moscow	Lebanon
Saudi Arabia				New York	Marrakesh
Switzerland				Northampton	Milan
Syria				Paris	Morocco
Turkey				Rome	Moscow
U.S.				Russia	New York
Vatican				Saudi Arabia	Northampton
Vienna				South Africa	Ottawa
Washington				Syria	Pakistan
				Turkey	Paris
				U.S.	Rome
				U.A.E.	Russia
				Vatican	Saudi Arabia
				Vienna	South Africa
				W.Va.	Syria
				Washington	Turkey
					U.S.
					Vatican
					Washington

Table 2: Places associated with Aziz - from 1998 through 2003

Table 2 shows one view of the places associated with Aziz from 1998 to 2003. Given this data, the analyst might perform a verification step similar to that described in the first example. However, the analyst might also be willing to accept the possible errors in this list, if he was simply looking for a broad trend. From the comparative listing alone, the analyst might conclude that 1998 and 2002-3 were years of considerably greater international importance than the intervening years.[29]

If we normalize the list to country names, organize them by region, and then align countries, across years, we get the following table.

98	99	00	01	02	03
Africa					
			Morocco	Morocco	
South Africa	South Africa		South Africa	South Africa	South Africa
Americas					
					Canada
U.S.	U.S.	U.S.	U.S.	U.S.	U.S.
Central and South Asia					
Afghanistan			Afghanistan		Afghanistan
Iran	Iran			Iran	
Pakistan			Pakistan		Pakistan
East Asia					
China	China	China	China	China	China
				Korea	
Europe					
Austria				Austria	
Britain			Britain	Britain	Britain
		Czechoslovakia			
France			France	France	France
				Germany	Germany
Italy	Italy	Italy	Italy	Italy	Italy
Russia		Russia	Russia	Russia	Russia
Switzerland					
Vatican				Vatican	Vatican
Middle East					
Bahrain				Bahrain	Bahrain
	Egypt			Egypt	Egypt
Israel		Israel	Israel	Israel	
Jordan	Jordan	Jordan	Jordan	Jordan	
Kuwait		Kuwait	Kuwait	Kuwait	
Lebanon	Lebanon	Lebanon	Lebanon	Lebanon	
Saudi Arabia				Saudi Arabia	Saudi Arabia
Syria	Syria	Syria	Syria	Syria	Syria
Turkey				Turkey	Turkey
	U.A.E.			U.A.E.	

Table 3: Countries associated with Aziz - organized by Region and Aligned across Years

[29] The reader should not attempt to draw any substantive conclusions from these graphics. They are meant to be indicative only. The limitations of the data set described in the Appendix mean that any patterns visible in these conclusions are not necessarily of significance.

From this arrangement of the data, it becomes quickly obvious which regions have been completely ignored (Latin America, Southeast Asia), and something of the relative importance of those regions mentioned emerges (a somewhat greater interest paid to countries in the Middle East than in Europe, for example). The presentation also highlights countries that seem the object of the most regular attention – China, Italy, Syria, and the US.

In Table 4 the same countries are grouped and ordered differently with the result that the apparent close ties with Italy and South Africa are highlighted in the non-Muslim world and the apparent close ties with Syria, Lebanon and Jordan are highlighted in the Muslim world.

98	99	00	01	02	03
Predominately non-Muslim					
China	China	China	China	China	China
Italy	Italy	Italy	Italy	Italy	Italy
U.S.	U.S.	U.S.	U.S.	U.S.	U.S.
Russia		Russia	Russia	Russia	Russia
South Africa	South Africa		South Africa	South Africa	South Africa
Britain			Britain	Britain	Britain
France			France	France	France
Israel		Israel	Israel	Israel	
Vatican				Vatican	Vatican
Austria				Austria	
				Germany	Germany
					Canada
		Czechoslavakia			
				Korea	
Switzerland					
Predominately Muslim					
Syria	Syria	Syria	Syria	Syria	Syria
Jordan	Jordan	Jordan	Jordan	Jordan	
Lebanon	Lebanon	Lebanon	Lebanon	Lebanon	
Kuwait		Kuwait	Kuwait	Kuwait	
Afghanistan			Afghanistan		Afghanistan
Bahrain				Bahrain	Bahrain
	Egypt			Egypt	Egypt
Iran	Iran			Iran	
Pakistan			Pakistan		Pakistan
Saudi Arabia				Saudi Arabia	Saudi Arabia
Turkey				Turkey	Turkey
			Morocco	Morocco	
	U.A.E.			U.A.E.	

Table 4: Countries closely associated with Aziz - grouped by Muslim and Non-Muslim association, ordered by number of years of close association

The analytic techniques suggested in this section – comparison across some time span, organization by geography, and grouping information by other possibly significant attributes such as cultural ties – are all well known and widely used in analysis. The difference in their use when supported by Information Extraction is, however, significant. IE allows these kinds of analyses to be performed on a far larger scale than in the past. The larger scale of the analysis improves the identification of trends, that might not show up in smaller data sets, or which may be skewed if the range of information used is too small. IE allows, as in this example, the easy transformation of data sets from one view to another, again improving understanding of the data and improving the chances that trends of significance can be found. The ability to make use of really large data sets in many cases will also increase the tolerance for error found in the original data itself, and in the results of the Information Extraction process.[30]

Finding links

Link analysis is of course a key technique in the analyst's toolbox. Relationships, especially between people and organizations, and the patterns of those relationships, are critical to many analytic tasks, such as in law enforcement, counterterrorism and counterdrug missions, or in competitive market analysis. Lockheed Martin has had success in feeding link analysis systems with information extracted from text in two different analysis scenarios. In one instance, we feed the link analysis tool with unverified information, and the analyst explores the information, looking for relationships and series of links or relationships that are of interest. Those links of potential use are next verified by the analyst, to see if they are, in fact, correctly extracted and of interest. This approach has the advantage of the "large numbers" types of effects that accrue in the preceding discussion of finding trends. Since material does not have to be initially verified, substantial volumes of data can be processed and reviewed by the analyst at a high level, without incurring the costs of verifying every single extraction. In one such system, we processed several thousand documents, with entity extraction, on a particular topic, with resulting extractions amounting to tens of thousands of entities. This volume of documents and entities is obviously far larger than what an individual analyst could process himself in any reasonable period of time. Yet from this data set, within a couple hours, a single analyst was able to discover several key entity relationships.

We have also used link analysis in a different type of scenario where only verified material is fed to the link analysis tool. This scenario is applicable to a situation in which the data set is of manageable size[31], and where the analyst's task is already extremely well defined: certain types of data and relationships, and the same trends and types of structures, are repeatedly investigated. In this instance the time savings and productivity enhancements are in the data preparation stage of the process, as already discussed concerning the verification of hypotheses.

[30] We do not understand yet what these tolerances might be. However, any large data set investigation expects some error. Since the question bears so heavily on the costs of system development it is one that should be investigated.

[31] What is manageable will, of course, depend upon the number of people available to verify the extraction and the period of time available for verification.

Extractions from the Tariq Aziz material also illustrate how extracted material can support link analysis. The two step, or second level, link is frequently of interest in analysis, that is, to use our example, having identified organizations connected with Tariq Aziz, the analyst may want to know if any informative connections exist between those organizations and other entities. Figure 5 shows several organizations of possible interest, extracted as being associated with Aziz, and then the people associated with those organizations (with Aziz removed from that list, of course). Often in this process, the analyst simply follows his intuitions. He already knows, from domain knowledge, that certain first level connections are likely of more interest than others. He follows those to the second level, and some prove fruitful while others do not. By displaying multiple two level connections at once the analyst will get a better picture of the significance of the connections. From a link diagram like that displayed in Figure 5, the analyst would proceed to verify those person names of interest to him. With a system like that described above, where he can simply click on the person's name and see the source text, this initial verification step is comparatively painless.

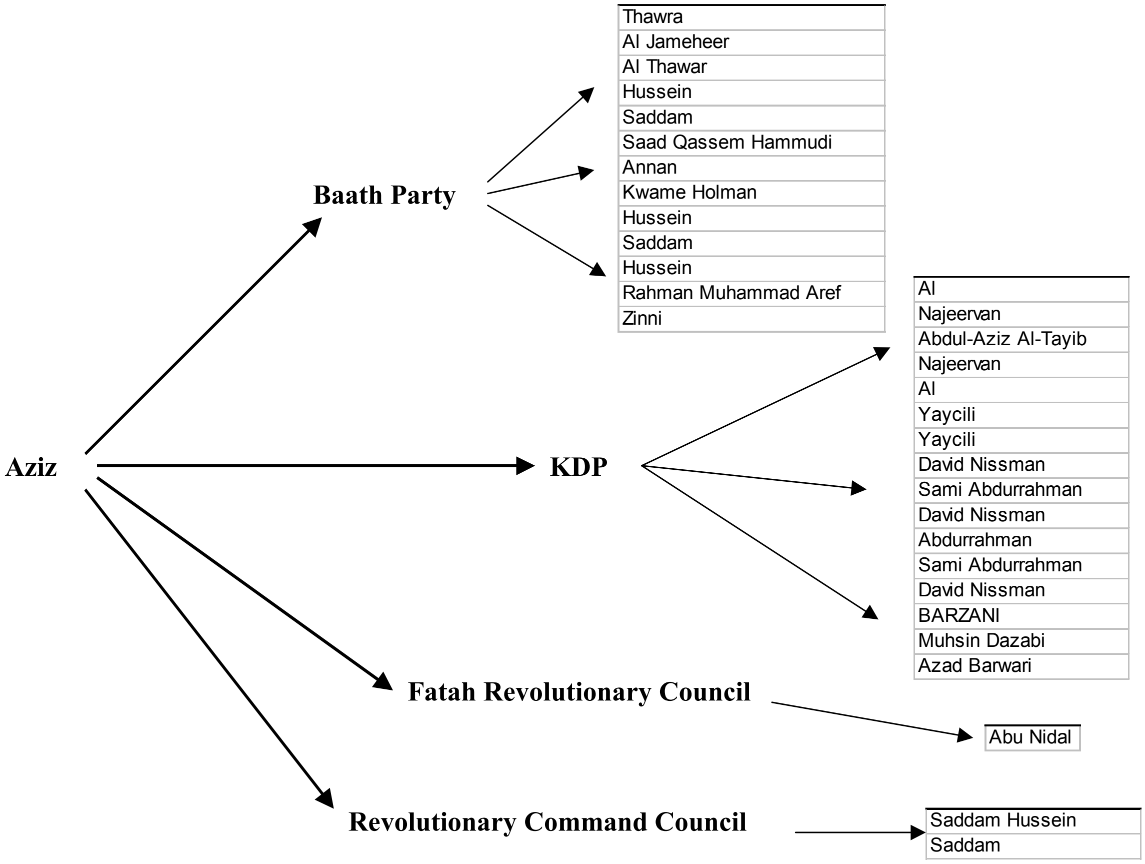


Figure 5: Example of Link Analysis diagram filled directly from Information Extraction

Once the data has been extracted into the database, these link discovery processes can then be repeated to multiple levels, although carried too far they may no longer be useful. With a good visualization tool the portions of interest of even a very large database can be rapidly explored. However, to use link analysis in this manner as a discovery tool, Information Extraction is required. If the analyst must first find and extract the information from text himself, link analysis becomes primarily a way to record, organize, and present to others what the analyst has already come to understand from reading the text.[32] Having analysts review the data first, before it goes into the link tool, also restricts the size of the data set to whatever can be handled by the human resources available for the task. This restriction, for practical reasons, tends to limit the information available for discovery to that derived from text either already known to be useful or felt to have a very high probability of relevance. Under these circumstances, discovery is less likely to produce something unexpected than to confirm the expected.

Finding the unexpected

The Holy Grail is the ability to find something both unexpected and useful, whether it be a pattern in the data or a single chain of important links. This kind of discovery is actually fairly difficult to achieve even with good automated tools. Once the technical problems are solved, reality may be that there is no key relationship or trend to be discovered. Even if such connections may exist in reality, they may not be captured in the data. So what has to be pursued by the analyst, the tool developer, and the system designer is a set of tools and a system configuration that will facilitate and encourage the discovery of such connections if they exist, with the realization that there is nothing to guarantee it. This section of the discussion extrapolates from our experience as defined in the previous three sections, to discuss what I believe is possible to accomplish today.

Finding the unexpected and useful pattern in the data depends upon three things:

The first is having a reasonable set of data upon which conclusions can justifiably be drawn. For example, any patterns that we might find, using the data set developed for this paper, are completely suspect, although individual links found based on this data are not. The reasons are that the data sample is relatively small and has been skewed by the selection method. The Google indexer itself only covers a portion of what is available in the open source; since it is not intended as an archive, information more than a few months old will be less well represented than information of more recent vintage. Thus historical trends, such as I illustrated in the second section, cannot really be developed on the basis of this information. Finally, the Google ranking methods are developed to push toward the top of the return list material of a mainstream interest. We do not have good measures for what constitutes sound bodies of text upon which trends conclusions can be based. However, to provide a basis for the discovery of historical trends, for example, text data would have to be collected according to reasonably consistent criteria from comparable sources for the period of years in which the comparisons are to be made.

[32] This statement is not meant to dismiss this role of link analysis tools as insignificant. The organization and presentation of complex, interconnected data are also extremely important, and can aid in the analyst's discovery of the significance of what he has found. But without IE the role of link analysis in discovery, as opposed to presentation, is limited. See, for example, the Call for Papers for the Text-Mining and Link-Analysis Workshop at the 18th International Joint Conference on Artificial Intelligence, August, 2003, organized by Marko Grobelnik, Natasa Milic-Frayling, and Dunja Mladenic which is focused on the "intersection of the two increasingly important areas of research: Text-Mining and Link-Analysis," available at http://www-2.cs.cmu.edu.

The second dependency is upon the analyst himself. In order to be able to recognize the unexpected, he has to recognize its indicators and be willing, and have time, to pursue those trains of reasoning. In order to understand whether or not the pattern is valid, he has to have the skills to not only evaluate and verify individual sources but to understand the possible biases introduced by his tools and methods. Finally, in order to understand the significance of the trend, he must understand his subject area: what is already known about it; what data is normally used and how it is regularly analyzed, and the reasons why what is unexpected has not previously been discovered.

But the third dependency is on the tools the analyst uses to look at his data. These tools, and how they are configured, as demonstrated throughout this paper, can have a considerable effect on what the analyst can see, whether he has time and flexibility to pursue multiple trains of reasoning, any of which may turn into dead ends, how easily he can verify individual pieces of data, and how easily he can understand the biases of his tools. Information Extraction lies at the core of these tools capabilities, enabling the key database and visualization functions the analyst requires.

The two most important capabilities that tools must have for the analyst to uncover the unexpected are (1) the ability to easily reconfigure and rapidly rerun data queries, often with only slight changes to the query, and (2) the ability to easily reconfigure and review the results of those queries in different formats.

The example data set provides an illustration of these two principles. The discussion above, *Exploring trends in large data sets* (p.13), showed that Aziz was closely associated with a number of countries on a yearly basis from 1998 through 2003. The identity of three of those countries outside the Middle East – Russia, China, and France – is certainly no surprise to followers of current events. I decided to determine if I could discover anything more about his associations with these countries: Was one significantly more important than another? Did their relative importance change over time? First, I explored a number of possible measures, or indicators of relative importance. Initially, I attempted to look at numbers of times Aziz was mentioned in close conjunction with key figures in these three countries, but the data was too sparse. However, judging by the names of the people associated with Aziz in conjunction with these three countries, the data seemed heavily weighted toward discussing contacts with Russia. As a follow-on to that insight, I decided to use the number of articles where Aziz is mentioned in association with each of these countries, as a measure of the country's relative importance over time to Aziz and presumably therefore to Iraqi policy. From the point at which I determined to explore the relationships with Russia, China, and France more thoroughly to the completion of Figure 6, which can be considered one answer to that question, required the development, running, and reviewing of at least two dozen queries and subqueries.

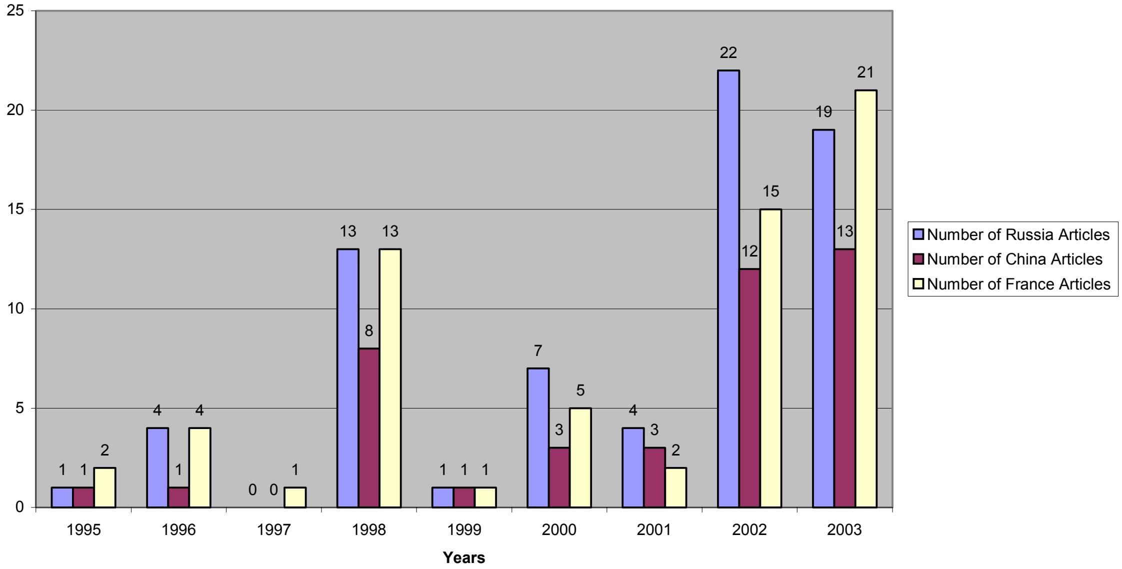


Figure 6: Numbers of articles - where Aziz associated with each country from 1995 to 2003

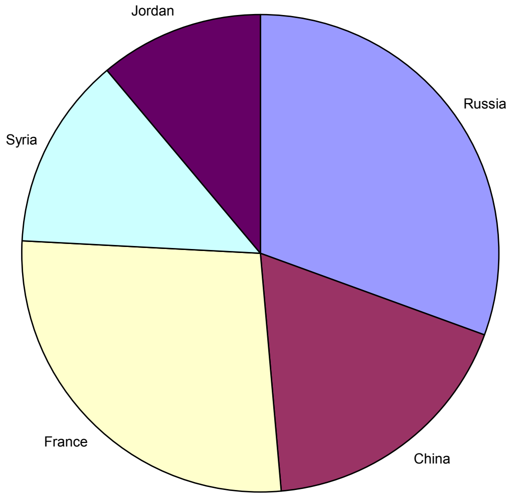


Figure 7: Share for each country - articles associating Aziz with each country, from 1995 to 2003, including Jordan and Syria

Figure 6, comparing the numbers of articles relating to France, Russia, and China, does show a slight dominance of the Russian articles. However, the significance of these numbers is hard to evaluate, given the unevenness of the data. To gain a better picture of the possible significance of the numbers, I looked also at the numbers of articles over the same period associating Aziz with Jordan and Syria, two countries neighboring Iraq that received fairly consistent attention according to the data already presented above, Tables

2, 3 and 4. Looking at the total number of articles associating Aziz with these five countries between 1995 and 2003, Figure 7 clearly shows a lower percentage of articles associating Syria and Jordan with Aziz than France, China, or Russia. However, it loses the historical information. The historical perspective is reintroduced in Figure 8, where the line represents the average of the Syria and Jordan numbers, which can now be seen to clearly follow the same trend as the other articles, but to be consistently lower than Russia and France throughout the period, and generally below China.

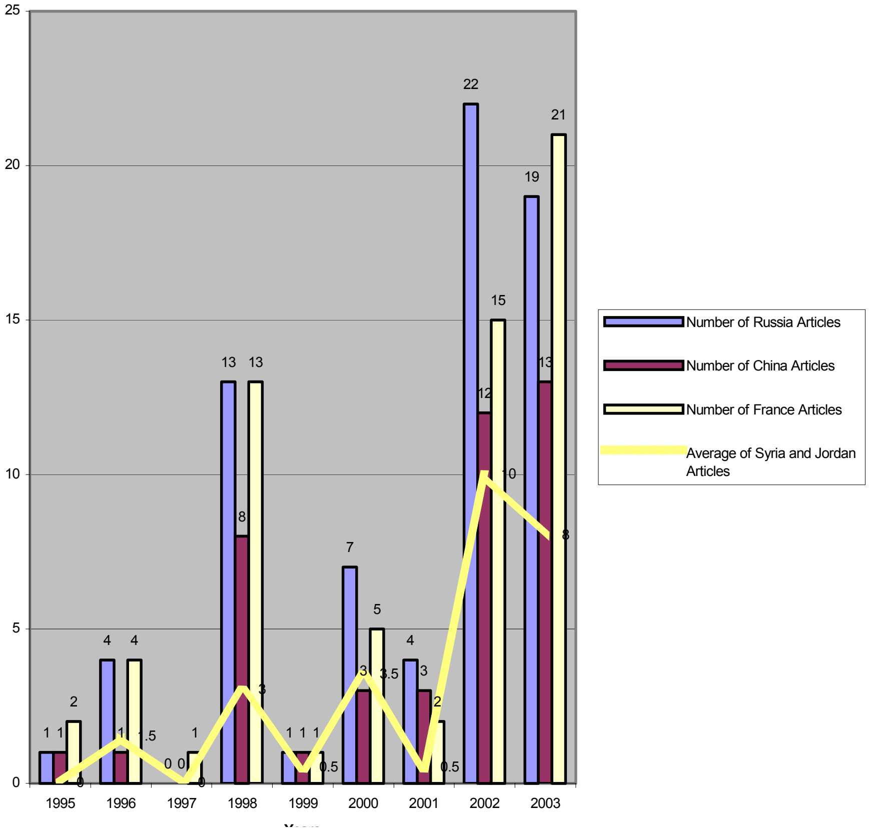


Figure 8: Numbers of articles per year - associating Aziz with China, Russia, and France, compared to a baseline of the average number of articles per year associating Aziz with Jordan and Syria

Conclusion

Information Extraction technology has been pursued by the Government for over a decade as a tool which proponents were certain could be useful both to Government and commercial applications. Its capabilities and applicability to many tasks are gaining increasing recognition in both Government and commercial spheres. It is a technology ready to revolutionize the analysis of information from text. Its revolutionary impact results from allowing the analyst access to far greater volumes of data than in the past, and from providing the analyst, using IE in conjunction with display tools, with multiple views of the same data. The access to multiple views of the data facilitates the analyst’s ability to thoroughly understand the data, supports his discovery of the unexpected, and aids his effective presentation of that data to others. Information Extraction, in Lockheed

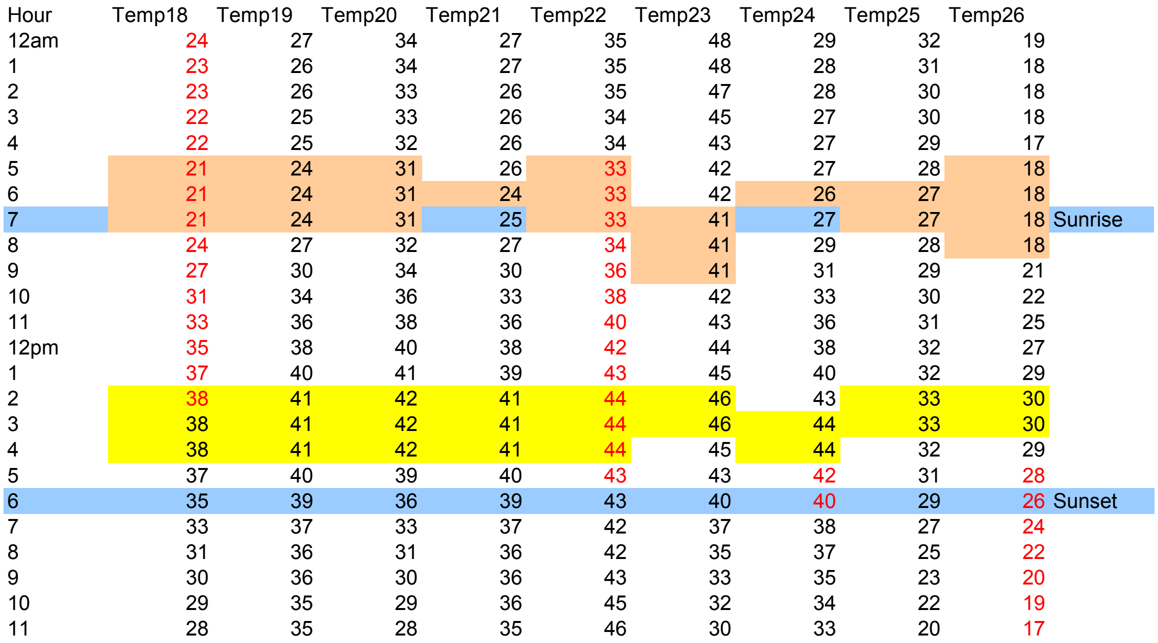
Martin and customer operational systems, vastly speeds the analyst's location of key entities in text, enables the analyst to investigate and discover trend data from text, and provides the capability to use link analysis on large volumes of text data for discovery and for final product. In all these uses, we have observed decreases in the analytic time required, and improvements in the quality of the product.

Information Extraction is an emerging technology. It is characteristic of emerging technologies that their best applications are not always understood immediately, whether by developers or by users. The development and evolution of the important uses for emerging technologies, and their acceptance, require considerable support by technologists, willingness to experiment on their part, and collaboration between user and technologist. Lockheed Martin experience with Information Extraction has demonstrated substantial IE utility in the roles outlined in this paper. These capabilities are but one step and described here as one contribution to the necessary, on-going dialogue between technologists and users concerning the ways this technology can continue to most benefit the timeliness and depth of analysis.

Appendix – The data and examples used in this paper

Presenting Data section

The data for the weather example used in the section titled “Presenting Data” was developed by copying the hourly temperature forecasts listed for the days cited from the www.weather.com site for the 22046 zip code. These are forecast temperatures, not actual, and illustrate the workings of the weather model, rather than the actual events of the days concerned. The sunrise and sunset times are from the *Washington Post* web page (www.washingtonpost.com), the weather section. There is some extrapolated data (in red) to cover periods where I did not collect all the data. This data is for illustrative purposes only. Below is the formatted spreadsheet with temperatures listed in columns for each day, the hours of sunrise and sunset highlighted, and the lowest and highest temperatures of the day also highlighted.



Hour	Temp18	Temp19	Temp20	Temp21	Temp22	Temp23	Temp24	Temp25	Temp26	
12am	24	27	34	27	35	48	29	32	19	
1	23	26	34	27	35	48	28	31	18	
2	23	26	33	26	35	47	28	30	18	
3	22	25	33	26	34	45	27	30	18	
4	22	25	32	26	34	43	27	29	17	
5	21	24	31	26	33	42	27	28	18	
6	21	24	31	24	33	42	26	27	18	
7	21	24	31	25	33	41	27	27	18	Sunrise
8	24	27	32	27	34	41	29	28	18	
9	27	30	34	30	36	41	31	29	21	
10	31	34	36	33	38	42	33	30	22	
11	33	36	38	36	40	43	36	31	25	
12pm	35	38	40	38	42	44	38	32	27	
1	37	40	41	39	43	45	40	32	29	
2	38	41	42	41	44	46	43	33	30	
3	38	41	42	41	44	46	44	33	30	
4	38	41	42	41	44	45	44	32	29	
5	37	40	39	40	43	43	42	31	28	
6	35	39	36	39	43	40	40	29	26	Sunset
7	33	37	33	37	42	37	38	27	24	
8	31	36	31	36	42	35	37	25	22	
9	30	36	30	36	43	33	35	23	20	
10	29	35	29	36	45	32	34	22	19	
11	28	35	28	35	46	30	33	20	17	

Figure 9: Daily Temperatures, Lowest and Highest Temperatures, and sunrise and sunset times - a different presentation.

Presentation of hypotheses for review, Exploring trends in large data sets, Finding links, and Finding the unexpected sections

In order to provide a consistent set of “live” examples to illustrate the themes in this paper, I developed a modest sized set of data based on a web search, using Google, with the query “tariq aziz” iraq. I then scanned the first 500 references and downloaded articles that looked useful. My exclusion criteria were to avoid obvious duplicates; to avoid compendiums of dates or lists of short citations; to avoid languages other than English; to avoid anything that looked wildly polemical or was obviously humor or satire. The resulting document set is 250 articles. We then ran these articles through our Information Extraction software (AeroText), out-of-the-box, without any tuning of the extraction patterns, and loaded the resulting extracted entities into an Excel spreadsheet, where I reviewed and cleaned up the data. This clean-up process was associated entirely with errors made by the output process, which loaded the information into the spreadsheet. There were a number of cases where dates, for example, rather than being dumped into the date column, were misread by the output program and placed into the spreadsheet as off-sets instead. Once the data was in consistent columns I transferred it to an Access database, where the analysis was performed which provided the data for the examples cited in this paper.

The text of the 250 web documents amounts to 5.3 MB of text data. From these documents, we extracted entities in 5 categories: Person names, Facility names, Organization names, Times (dates and hours in several formats), and Place names. There were a total of 24,202 entity instances extracted, for an average of 96.8 per document. In Excel the extraction records occupied about 2.5 MB of spreadsheet. The extractions are loaded in a single table in Excel, and then in another single table in Access, with the fields Type (Person, Time, Place etc), Value (the extracted entity name), Source (the file name), Left Off-set (character location of the first character of the extraction, counted from the beginning of the document), Right Off-set (character location of the last character of the extraction, counted from the beginning of the document). Additionally, I manually developed a list of short source titles for a field in the Access database called ShortSource. Since the Source field uses the rather long and often meaningless file name of each of the articles as the name of the source, the ShortSource field is a convenience; it allowed me to easily type in source names for queries and remember source names, if I needed to do so. The entity extractions break down as follows:

Type	Number	Unique (approximate)
Person	5,792	1,325
Facility	56	41
Organization	4,845	1,045
Time	3,365	N/A
Place	10,143	583

Table 5: Numbers of extractions of each type - the Example document set

The number is the total number of entity instances identified by the extraction. So, for example, every time "Aziz" or "Tariq Aziz" is located in the document, it contributes to this count. The Unique count is only approximate: I simply used the "distinct" query in the database, and so "Aziz" and "Tariq Aziz" are counted as separate entities by this method. The level of redundancy is reassuring. It tends to suggest that while information in one instance may be missed by the automated extraction, there is a considerable chance it will be picked up in one of its other instances.

Two caveats must be made with respect to these illustrative examples. First, no-one should draw any conclusions about Iraq or Tariq Aziz on the basis of these examples. They are simply illustrative of the kinds of things that can be done with Information

Extraction, and have not been double checked or properly sourced for this paper, nor did I develop the data set in the way I would have if I was doing rigorous analysis. Second, although the data set is smaller than what we use in analytic operations, and although I used only a database as an analysis tool, nonetheless the examples presented here are all exactly analogous to tasks that we have performed in operational setting, where we use not only databases, but tools to help facilitate the analysis of the data in those databases such as link analysis and geographic information systems. But the underlying keys to the analysis are the extractions and the database query capabilities, which are the basis of the examples used in this paper.

Table 6: Raw (unedited results) – source of Table 2 (p. 13)

98	99	00	01	02	03
Afghanistan	Amman	Amman	Afghanistan	America	Afghanistan
America	Baghdad	Assisi	America	Ankara	America
Amman	Beirut	Baghdad	Baghdad	Ankara, Turkey	Ankara
Assisi	China	Beirut	Britain	Assisi	Assisi
Baghdad	Egypt	China	China	ASSISI, Italy	ASSISI, Italy
Bahrain	Hassan	Damascus	i1025	Baghdad	Baghdad
Britain	i1025	i1025	Iraq	Bahrain	Bahrain
China	Iran	Iraq	Israel	Britain	Britain
Earth	Iraq	Israel	Kuwait	China	Cairo
France	Italy	Jordan	Middle East	City of Doom	China
Geneva	Jordan	Kuwait	Moscow	Damascus	City of Peace
Hassan	Lebanon	Lebanon	New York	Earth	Damascus
i1025	Middle East	Middle East	Pakistan	Egypt	Earth
Iran	South Africa	Mideast	Russia	France	Egypt
Iraq	Syria	Moscow	South Africa	Germany	Europe
Israel	U.S.	Prague	Syria	Hollywood	France
Italy	Umbrian hill city	Rome	U.S.	Iran	Germany
Johannesburg	United Arab Emirates	Russia	United States	Iraq	Hollywood
JOHANNESBURG, South Africa	US - British	Syria	US	Israel	Iran
Jordan	WASHINGTON	US	USA	Israel's Dead Sea coast	Iraq
Kuwait		WASHINGTON		Italy	Israel
Lebanon				Johannesburg	Israel's Dead Sea coast
Middle East				JOHANNESBURG, South Africa	Italy
Moscow				Jordan	Johannesburg

Mosul
Mosul, Iraq
New York
Pakistan
Paris
Russia
Saudi Arabia
Switzerland
Syria
Turkey
U.S.
United States
United States of America
US
US - British
USA
Vatican
Vienna
WASHINGTON

Karbala	JOHANNESBURG, South Africa
Kuwait	Jordan
Lebanon	Korea
Marrakesh	Kuwait
Middle East	Lebanon
Mideast	Marrakesh
Morocco	Middle East
Moscow	Milan
New York	Morocco
Northampton	Moscow
Northern Alliance	Mosul
Northern Alliance in Afghanistan	New York
Paris	north of Baghdad
Rome	Northampton
Russia	Ottawa, Canada
Saudi Arabia	Pakistan
South Africa	Paris
southwest of Baghdad	Persian Gulf
Syria	Rome
Turkey	ROME, Italy
U.S.	Russia
UK	Saudi Arabia
Umbrian hill city	South Africa
United Arab Emirates	Syria
United States	Telkaz
United States of America	Turkey
US	U.S.
USA	Umbrian hill city
Vatican	United States
Vienna	US
W.Va.	USA
WASHINGTON	Vatican
	WASHINGTON