Document Title:	Link Analysis Survey Status Update – January 2006
Author(s):	William M. Pottenger, Ph.D. ; Xiaoning Yang ; Stephen V. Zanias
Document No.:	219552
Date Received:	August 2007
Award Number:	2005-IJ-CX-K006

Opinions or points of view expressed are those of the author(s) and do not necessarily reflect the official position or policies of the U.S. Department of Justice.

Link Analysis Survey
Status Update – January 2006

William M. Pottenger, Ph.D., Xiaoning Yang, and Stephen V. Zanias
Lehigh University Computer Science and Engineering Department
{billp, xiy204, svz2}@lehigh.edu

Table of Contents:

Abstract: This update represents survey work conducted from June, 2005 through January, 2006 in the field of link analysis. The purpose of this survey is to identify the leading link analysis solutions that are being developed at institutions and corporations around the world and evaluate the applicability of their capabilities to the law enforcement community. The information has been organized and presented in such a way to benefit both practitioners (e.g., law enforcement officers) and researchers. After providing a high-level theoretical overview of the link analysis field and its terms, axes, capabilities, and algorithms, an in-depth analysis of over 30 different solutions is provided. The remainder of the report discusses the next steps of the survey effort.

Key words: link analysis, text analytics, text mining, data mining, machine learning, law enforcement

1 Introduction

As the amount of digital data used in law enforcement continues to grow, it is becoming increasingly important to maintain and coordinate this data accurately and precisely. There is no other field to which this is more important than in the governmental and law enforcement field, argues Dr.

Donald Brown, Chair of the Department of Systems and Information Engineering at the University of Virginia (Brown, 1998). Numerous government agencies have conducted studies to look into the field of "data mining" to determine how this technology can be used to combat this problem. Jeffrey Seifert states in a report to the U.S Congress, "Data mining is emerging as one of the key features of many homeland security initiatives" (Seifert, 2004). The consensus appears to be that data mining will be the direction of the future.

As a component of data mining, there is no doubt that link analysis will contribute greatly in this endeavor. With innumerable data formats and database schemas for existing data, the task of coordinating law enforcement and defense information is daunting. Often, it is not a matter so much of *collecting* clues and leads as it is a matter of *linking* and *coordinating* these leads and *transforming* them into actionable information that can be used to pursue justice. With the large number of data mining and link analysis solutions available to officers, the need for insight into these solutions, their capabilities, and their applicability is evident.

This update presents our survey findings of the link analysis solutions that are currently available through academic or commercial venues. The terms used and the axes along which these solutions may be viewed are presented in Section 2, while a summary of each solution surveyed is given in Section 3. Section 4 presents the conclusion and our future steps in the completion of our survey work are articulated in Section 5.

1.1 Topic Overview

As presented in our proposal paper (Pottenger and Zanias, 2005b), the government has taken great interest in the field of data mining in general, including link analysis. From May 2003 through April 2004, the GAO conducted a comprehensive survey of the data mining tools that were either currently being used or currently in the planning stages by various departments of the government. The results, published in the GAO report "Data Mining: Federal Efforts Cover a Wide Range of Uses" (GAO, May 2004), are summarized below:

- Over 40% of the federal departments (52 of 128) were either using or are planning to use data mining;
- Of the 199 data mining efforts reported, nearly 35% (68) were planned efforts;
- The data mining efforts were organized into several categories, with two of the six most frequent categories pertaining specifically to law enforcement (*detecting criminal activities or patterns*, and *analyzing intelligence and detecting terrorist activities*);
- The Department of Defense reported the largest number of data mining efforts (over 20% with 41 of the 199), and also had the highest number of efforts to analyze intelligence and detect terrorist activities (Note: the Central Intelligence Agency (CIA), the National Security Agency (NSA), and the Department of Army did not participate in this study);
- The efforts for detecting criminal activities or patterns were spread throughout the agencies, with no department claiming a clear majority of the efforts.

This report provides a glimpse into the enormity of the data mining efforts underway at the federal level. It is important to note that these numbers include neither efforts initiated at state or local law enforcement levels nor those endeavors undertaken in industry or academia. Heavy investment by In-Q-Tel, the not-for-profit extension of the Central Intelligence Agency (CIA), in data mining fields such as such as "Knowledge Management" and "Search and Discovery" also lends credence to the importance of government efforts in this arena (Kanellos, 2005). As Dr. Colleen McCue, an expert in the field and a former officer with the Richmond, VA Police Department states, "Data mining, when applied to tactical crime analysis, is a knowledge discovery tool that can be used to review extremely large datasets and incorporate a vast array of variables, far beyond what a single analyst, or even an analytical team or task force, can accurately review" (McCue, 2003).

Due to the sheer volume of information available to law enforcement coupled with the issues dealing with numerous formats, data distribution, and data quality, the task of understanding law enforcement data would seem to be intractable. However, link analysis is proving to be a vital technology to address many of the issues currently facing law enforcement. As previously mentioned, issues in law enforcement data are often not so much a matter of *collecting* clues and leads as it is a matter of *linking* and *coordinating* these leads and *transforming* them into actionable information that can be used in the pursuit of justice.

There are a number of link analysis *solutions* that address this problem. (We have used to term *solution* in this survey to incorporate any hardware component, software package, or any other type of technological product that is used to provide some sort of link analysis task and to not limit ourselves to a particular type of product.) As much of this information exists in textual form, one approach is to utilize information extraction technologies in order to transform this data into named entities that can easily be transformed into structured, searchable data. Our survey work in this domain is presented in (Pottenger et al., 2006a). Link analysis solutions, on the other hand, allow such information to be joined together or linked to transform textual data into actionable information.

There are many examples of solutions used in law enforcement that have produced impressive results. In addition to crime mapping tools (e.g., Brown (1998), Gorr (2004)), neural networks (Graham-Rowe, 2004), and forecasting and patterning technologies, link analysis solutions are being used to provide valuable insight for officers. For example, the Richmond, VA Police Department, under the direction of Dr. Colleen McCue, has been implementing many data mining techniques and applications. Working with SPSS and RTI International, the department has used the tools to predict random gunfire occurrences and helped to reduce the city's New Year's Eve 2003 gunfire incidents by 47% over the previous year (Leon, 2005). Using decision tree analyses and other techniques helps the officers to more quickly respond to the situation within the critical 48-hour time window (Leon, 2005). In another effort in Bethlehem, Pennsylvania, Dr. William M. Pottenger of Lehigh University is developing a solution entitled D-HOTM, an acronym for Distributed Higher-Order Text Mining, which enables free text conversion, semantic and link analysis in a distributed law enforcement system (Wu and Pottenger, 2005a) (Li, et al., 2005). A component of their system, which enables automatic conversion of unstructured textual data into a structured database, is currently being tested at the Bethlehem Police Department in their investigations unit. Additionally, the Florida Integrated Network for Data Exchange and Retrieval (FINDER) system developed by the University of Central Florida has enabled over 120 different agencies within the state of Florida to coordinate and share information ranging from pawn purchases to crime and vehicle data (FINDER).

One of the most well known law enforcement data mining solutions is CopLink®, which bridges the academic and commercial worlds (NLECTC, 1999). Developed at the University of Arizona's Artificial Intelligence Laboratory under the direction of Dr. Hsinchun Chen, the program received national exposure during the Washington sniper shootings of 2002. Applied after the incidents, the program was able to identify patterns in the evidence from the case that could have led to a faster apprehension of the criminals (Mnookin, 2003). Given the program's applicability, Knowledge Computing Corporation (KCC) has been formed to market and distribute the CopLink® system to police departments (KCC).

Although not from the law enforcement field, one good example of the utility of this technology is presented in Shachtman (2005). The article discusses how Whirlpool utilized link analysis software to learn about a product deficiency. The company, which receives over 400,000 customer service calls each month, was in the midst of a microwave oven recall. A team of people first went through the documents, searching for relevant key words to locate 18,500 matching records. Then, six people spent an entire weekend reading through the results to narrow down the information to 700 calls potentially related to the problem. Yet, when the company tried to identify the problem using Attensity's technology, 542 detailed, specific results were returned in approximately 10 seconds.

1.2 Survey Method

As described in our proposal, Pottenger and Zanias (2005b), there is a great need to understand the link analysis solutions that are currently available. These solutions have particular importance to the law enforcement community, as they coordinate information that allows officers to serve justice more quickly. Our goal is to not only identify the leading technologies and solutions, but also to determine an efficient and more meaningful means of evaluating these types of tools. This includes not only helping to develop valuable metrics and compiling representative datasets, but it also involves a seven step process for evaluation developed in Pottenger and Zanias (2005b) to determine whether the seven step process itself is an efficient means of carrying out this type of evaluation.

It is our hope to bring coordination and organization to the intersection of law enforcement and data mining applications. By identifying the appropriate solutions and leading technologies, these solutions can then be used to improve and verify the metrics and datasets produced. It is our sincere desire that this work will aid officers in their law enforcement efforts.

As presented in our proposal, the following is the seven-step plan that we have developed to accomplish this goal:

1. Survey the link analysis field and organize the solutions into categories;
2. Identify/develop suitable metrics/standards for comparing solution performance (e.g., precision, recall, f-beta, support for GJXDM, interoperability with other solutions, etc.);
3. Identify/compile 'ground truth' datasets for use in the evaluation of the solutions;
4. Select representative solutions from each category, and evaluate those solutions based on the ground truth datasets using the selected metrics/standards;
5. Propose the use of the selected metrics/standards, ground truth datasets and methodology of evaluation for widespread use by law enforcement agencies in evaluating other/future solutions;
6. Perform a leading edge technology analysis that identifies research directions needed to improve the utility of data mining technologies for use in law enforcement – research directions that are also suitable for funding by federal, state and local agencies;
7. Prepare a demo of and report on the various solutions evaluated, metrics identified, datasets developed and methodologies employed, as well as on the future directions needed to advance the field in terms of the application of data mining technologies in law enforcement, criminal justice and homeland defense.

This status reports presents our work up to the present date. As of this paper, we have completed our preliminary survey results and identified several *axes* or categorizations by which these solutions can be identified. This work has been conducted through the use of information and internet search gathering, as well as communication with industry experts and solution developers. We have also consulted with law enforcement personnel to learn more about their needs and requirements as well. Although these categorizations have proved to be difficult, it is our belief that utilizing these axes will aid in the development of more efficient and meaningful metrics. These axes and the future steps for the project are discussed in greater detail in the remainder of this report.

1.3 Outline and Audience Scope

Often, academic research papers serve to further the purposes of other researchers; one research work begets the next in a never-ending process. However, we believe strongly that the information contained in this work can benefit not only the interested researcher, but – equally important – be of direct aid and assistance to the law enforcement practitioner. We have, therefore, presented the information in such a way that both parties can quickly and easily glean from this survey the information they desire.

Section 2 presents a high-level theoretical overview of link analysis. In addition to defining terms that are used throughout the survey, we also present a division of link analysis solutions into categories that give insight into their capabilities. We also highlight the prominent algorithms in use in various link analysis solutions.

Section 3 presents the heart of our research work to date – summaries of the various solutions we were able to identify. Each institution and their solution(s) are presented in order, organized first into a high-level categorization of academic solutions (those coming from research institutions, universities, colleges, and the like) and commercial solutions (those solutions currently offered as part of a business venture or available from the government). Within these groupings, the solutions are arranged alphabetically by the developing institution.

Within each solution summary the information pertaining to each solution is presented under one of several headings. The first headings (Company Introduction and Domain Scope, Output/Results, Application to Law Enforcement, Evaluation, Financial, Inputs Required, and Software) contain information that is more general in nature and present material that we feel would be more pertinent to law enforcement deployment. We feel that these are the more pressing issues that a law enforcement practitioner would be interested in when looking to identify a suitable link analysis technology, and, therefore, these sections are primarily directed towards the law enforcement practitioner. The latter part of the solution summary (Link Analysis Algorithm and Knowledge Engineering Cost) contain more detailed information about the solution's process and technical details of the implementation of the solution. Therefore, these sections are directed towards the researcher.

A final component of each solution summary is a summary table. This table serves both the practitioner and the researcher in providing a condensed version of our summary of the solution and is meant to provide the reader with an easy and convenient means of learning about the solutions presented in this report. Additionally, in order to better index and organize these results, several summary terms and groupings have been utilized. A description of these terms is presented in Section 2.2.

2 Link Analysis Overview

2.1 Link Analysis Terms

Words and their context provide a great deal of insight into the structure (lexical and syntactic) and meaning (semantics) of natural language. This is also true for the link analysis field. Often, the terms used in this field provide various nuances in meaning. In order to avoid confusion over the terms used in this survey, we have provided an explanation for each of the terms used extensively throughout this report.

It is fitting that we should begin with a definition of link analysis itself. For the context of this survey, ***link analysis*** is the active pursuit of identifying relationships and connections (links) between values, entities, and objects. This definition will be further expounded upon as we develop our definitions of other related terms in what follows.

Named entities are also vital to an understanding of this field. ***Named entities*** refer to values that contain both a "value" and the associated "type" or "category" to which they belong. In other words, named entities are <type, value> pairs which are extracted from a document source. (These are also known as attribute-value pairs or items (Witten and Frank, 2000).) NIST simply defines a ***named entity*** as "a named object of interest such as a person, organization, or location" (NIST, 2001). This concept may be illustrated best through the use of a simple example. If "Pennsylvania" is read from any given medium, it may seem obvious to the human reader that this refers to a state in the United States. Therefore, the pair <state, "Pennsylvania"> represents a named entity because the value (i.e., "Pennsylvania") is assigned to a particular category (i.e., "state").

NIST defines ***information extraction*** as "the extraction or pulling out of pertinent information from large volumes of text" (NIST, 2001). Basically, this term refers to the learning of information that occurs by converting textual data into discernable, searchable information. Our companion survey specifically focuses on a portion of the Information Extraction (IE) space that is known as ***Named Entity Extraction*** (NEE), which results in the extraction of named entities as the output of the IE process (Pottenger et al., 2006a).

Another portion of the IE space includes the extraction of relationships. ***Relationship extraction*** results in the discovery of connections between values that occur within textual data. While these relationships are often learned as part of the information extraction process, we have chosen to consider relationship extraction to be part of the link analysis process. This is because extracted relationships establish links between or among various entities, and, as such, should be considered a link analysis activity.

This is an important observation about the IE field: in short, terminology used in this field is, as noted, imprecise. For instance, (Feldman, 2002) describes *entity recognition* as the process that "extracts proper names and classifies them according to a predefined set of categories, such as Company, Person, Location, and so forth" while in *information extraction*, "key concepts (facts or events concerning entities or relationships between entities discussed in the text) are defined in advance and then the text is searched for concrete evidence for the existence of such concepts." Therefore, our information extraction definition coincides with Feldman's entity recognition definition and some combination of our information extraction and link analysis definitions coincide with their information extraction definition. Because of these semantic differences, we have provided this section to clearly state the differences in terms that we wish to describe. By separating the data extraction (IE) and data linkage (LA) phases of the process, we hope to provide a framework within which both processes are easier to understand.

The last issue crucial to understanding this survey has to deal with scope. Often, data mining schemes can learn values and relationships from a variety of input. We have termed these inputs ***sources***. Often, data will occur in reports, proposals, emails, websites, or other such sources of information which could be generalized into a "documents" categorization. However, as many data mining techniques incorporate database data as well, using the term "documents" does not provide a clear representation. Therefore, to incorporate the use of database records and other such information in our survey, we have selected the term *sources.* Given this, a ***source*** refers to any one individual piece of information. An email message, a database record, a company report, a MS Word document, and a webpage each constitute an individual source.

2.2 Link Analysis Axes

As with nearly every issue, there are multiple vantage points from which to classify, organize, and divide. The field of link analysis is no different, and choosing an optimal axis is not a simple task. Not only should the classification divide the solutions along easily-differentiable attributes, but such divisions must also provide as much information in the categorization as possible. While there are many similarities to information extraction axes (Pottenger et al., 2006a), link analysis presents its own unique set of axes.

One possible way to divide link analysis technologies is based on their *sophistication* (the degree of complexity of the process) or their *practicality* (how useful the system would be to the law enforcement officer). For instance, an algorithm that simply links all words in a sentence which begin with a capital letter and concludes that they are related via the sentence's verb would not be an example of a "sophisticated" technology. But how should complexity be measured? Similarly, practicality is an abstract concept that is almost entirely dependent upon the context in which the solution is used.

Another axis along which to view link analysis solutions could be *link order*, or the number of "hops" that is required to reach a certain conclusion. For instance, the link between a parent and a child would be a first-order relationship as there is a direct relationship between the two individuals. A link between a child and a grandparent would be considered a second-order link, as the relationship between the two is only made possible because of the parent; each has a direct first order relationship to the parent and, through the parent, the grandparent and child have a relationship. This concept can also be applied to law enforcement link analysis applications. For instance, Robert may own a car (first order) which is used to commit a bank robbery. Therefore, the relationship between Robert and the bank robbery is a second order link (via the car).

Given the complexity of choosing appropriate axes, we came to the conclusion that we should leverage the aforementioned survey results in information extraction (Pottenger et al., 2006a) and categorize solutions by the nature of the entities linked. In particular, we divided solutions into those that perform link analysis based on named entities, and those that perform link analysis on some other type of value. Given our desire to target practitioners, dividing solutions based on *what* they link (named entities or some other value) provides a reasonable starting point for identifying solutions that can leverage the named entity extraction capabilities surveyed in Pottenger et al. (2006a).

The second division is based on the *scope level* of the link analysis. By *scope level*, we refer to the range of sources of information on which the links are formed: *intrasource* or *intersource*. Using the definitions provided in Section 2.1, an *intrasource* solution refers to a solution that identifies links that occur within a single source. *Intersource*, on the other hand, refers to links that are formed between values that exist across multiple sources. The scope level is of interest to researchers working in the field, especially those engaged in intersource link analysis research. Nonetheless, practitioners will also benefit due to the fact that data security and privacy are important concerns in the law enforcement community, and the scope at which the solution operates can distinguish the degree of information sharing intrinsic to the information system.

These axes are utilized extensively throughout the solution summaries. Additionally, a combination of "attributes" help in analyzing the solutions and give insight into the algorithms employed in the processes. As with the axes, these attributes are presented in the solution summaries and appear as fields in the summary tables provided with each solution analyzed. In an effort to provide further categorization, we have qualitatively created nominal values associated with each of these attributes and describe these values in what follows.

The *Domain Scope* attribute refers to the specific application domain (if any) that the information extraction solution is targeted at. Although this is a general category, we believe that it will provide some insight into how the solution should be used. The domain scope attribute values are not limited to any particular subset. A second attribute, *Application Type*, states whether the solution utilizes information extraction and/or link analysis capabilities.

A third attribute, *Knowledge Engineering Cost*, groups solutions into one of three general classes: *high*, *medium*, and *low*. Knowledge Engineering Cost (KEC) refers to the amount of effort and preparation that is required to transform raw data into actionable information usable by the solution. A *high* KEC refers to a procedure or algorithm that requires substantial effort to transform data. An example of a high KEC process would be one where a human domain expert is required to manually craft the rules needed to extract information from a given domain. If, for instance, a date entry needed to be extracted from text, there could be several ways to do this. A fully manual approach would involve a domain expert in the creation of a set of rules that could be used to extract a date feature. A rule to recognize a numerical format (i.e., MM/DD/YYYY, DD/MM/YYYY, etc.), or a textual date (i.e., January 1, 2000) could be created by the user to recognize the pattern and extract it.

A *medium* KEC solution would implement information extraction through the means of a combination of human and technological processes. While some human interaction would be required, the solution also would partially automate the process. For example, if a user labels a series of text samples that the solution then uses iteratively to formulate an information extraction rule, its KEC is medium. (Note that this is different from the solution that provides a GUI "workbench" that guides the user through a process to manually create their own rule; in the medium KEC case, the solution provides a degree of automation by analyzing the samples and formulating the rule.) Continuing the date example from above, a medium KEC approach could have the user label textual features within a text source and then have the solution create the rules from the data.

A *low* KEC rating would be given when the technology requires practically no user interaction, but is able to perform the tasks automatically. A technology where the raw data can simply be entered and information automatically extracted for the user would be the ultimate example of a low KEC technology. If the solution automatically recognized date attributes (continuing the example) without any need for training by a human user, then it would have a low KEC.

Continuing with the summary table attributes, *Financial Cost* represents the dollar cost required to obtain the solution based on the information available from the manufacturer[1]. The attribute *Input Requirements/Preparation Required* describes any special cases for the input data, such as expected data types, formats required, etc.

In order to better categorize and group the solutions based on the techniques, algorithms and processes used, we employ well-known terminology from the machine learning field such as *Labeling*, *Model generation*, and *Supervision*. *Labeling* refers to the process whereby example entities are named during the training process; the output is a set of labeled training data. If a user tediously labels the entities manually, the learning process is referred to as *manual*. If the user provides input to assist the solution in carrying out labeling, it is *active* learning. In an *automatic* approach, the solution generates all of the labels, while a *hybrid* approach uses a combination of the above. Note that not every approach will require labeling of data; for instance, a manual rule crafting approach does not utilize a labeling process. In cases where no labeling is performed, the solution is categorized with labeling class *n/a* (not applicable). If there are multiple approaches used by the system, it is considered *various.*

Model Generation produces a model which can classify unlabeled data. As with *labeling*, this attribute has five values that refer to the various levels of human interaction (*manual*, *active*, *automatic*, *hybrid*, *n/a*, and *various*).

Supervision refers to the "guidance" that is required in order to construct or develop the model. More precisely, it is the level to which labeled training data is used to construct the model for information extraction. Supervision is applicable to both the labeling and model generation processes. *Labeling Supervision* seeks to answer the question, *"Do we have to label raw data in order to bootstrap the process of labeling the training data?"* By this, we use the term *supervision* to imply that the labeling process requires some degree of labeled data to execute its algorithm. For instance, if the labeling process requires no labeled data on which to train, then the process is *unsupervised. Semi-supervised* and *supervised* labeling require increasing levels of labeled data to learn the labeling technique. For instance, labeling sentences would be considered a semi-supervised approach as opposed to labeling individual words (a supervised approach).

Similarly, *Model Generation Supervision* asks, *"Do we have to label the raw data in order to learn/discover the model?" Unsupervised* would mean that no labeled data is required to produce the model, while *supervised* would require fully-labeled data to create the model. A *semi-supervised* approach would lie between these extremes.

[1] This of course implies the solution is marketed commercially; academic solutions normally do not have a purchase price.

The *Solution Output* attribute specifies the manner in which the output is produced, including such issues as visualization or data format. The yes/no responses to "*Is performance evaluation available?*" and "*Solution/demo available?*" seek to provide quick responses as to whether testing and performance assessments have been conducted and whether the solution provider is willing to provide examples of their solution's capabilities on a readily available basis.

We also estimate the level of applicability to law enforcement we believe the solution provides in the *Application to Law Enforcement* attribute. This attribute has been qualitatively divided into one of three categories: *extensive*, *moderate*, and *limited*. By *extensive*, we mean that the solution has a high applicability to law enforcement in terms of its capabilities, domain scope, scope level, and overall performance and/or is already being actively used in law enforcement activities. A *limited* rating means that, while the solution has information extraction capabilities, we do not necessarily feel that it could be easily used or deployed in a law enforcement setting. A *moderate* rating is assigned to a solution that has applicability in law enforcement, but based on our survey is not currently being used in this domain.

Given this background on the various axes and attributes, we now go into more detail with regards to our two main axes: *named entities* and *intra-/intersource*.

2.2.1 Named Entities in Link Analysis

The first general category discussed refers to whether or not the link analysis solution uses *named entities* to identify links and relationships. In many solutions, the process of intrasource link analysis is intrinsically tied to the extraction process of entities. In *named entity link analysis*, links can only be discovered between named entities. For more detailed information on named entity information extraction, see Pottenger et al. (2006a).

2.2.2 Intrasource Relationship Formulation

Intrasource relationships represent those links formed within the lower *scope level* class of a link analysis solution. In other words, these relationships are between or among values that occur within a single source. Not only do these relationships represent a more simple kind of relationship, but also these types of relationships are often explicitly stated within the text or can be learned through the use of such techniques such as anaphora resolution. In addition, the information scope from which data can be learned is more narrowly focused and does not encompass data sharing or time frame issues commonly found (such as who the current president is or who won last night's baseball game).

Intrasource relationships are output in a variety of ways, including textual and database output, and often are produced visually or through search techniques. The common factor is that a relationship between values must be identified. For an example, consider a textual sample of a biography of Albert Einstein.

…Albert Einstein was born on March 14, 1879 in Ulm, Germany….

An intrasource link analysis could learn the following intrasource relationships:

- <person-birthdate> Albert Einstein; March 14, 1879 </person-birthdate>
- <person-birthlocation > Albert Einstein; Ulm, Germany </person-birthlocation >

As already discussed, these relationships include relationships that are stated explicitly in the text, without reference to time frame or other coordination issues.

2.2.3 Intersource Relationship Formulation

Intersource relationships take link analysis one step further and go beyond the information that is presented within a single source. By using information from multiple sources, higher order links can be learned in addition to understanding first-order links. Because of this, intersource link analysis can be considerably more powerful, discovering relationships that cannot be learned from the analysis of a single source.

Intersource relationships are similar to intrasource relationships in that they require a "field" that specifies the relationship type. While not a concern with intrasource relationships (as they all originate from the same single source), a more well-developed intersource technology will also include the path traveled or sources combined to identify the link.

As with intrasource relationships, output can also be produced in a variety of ways, including textual output, database output, visual interfaces, and/or search techniques. Continuing with the above example, if we learned the following relationships from a second source:

- <person-birthdate> Paul Ehrlich; March 14, 1854 </person-birthdate>
- <person-birthlocation > Paul Ehrlich; Upper Silesia, Germany </person-birthlocation >

Then using an intersource link analysis solution, we could discover that Mr. Ehrlich and Mr. Einstein were both born in Germany on the same date. This information cannot be learned from either source individually, but, through the use of intersource link analysis, the connection could be learned.

Similarly, links between links or relationships between relationships can be discovered using intersource link analysis. After a solution learns a relationship, it is usually stored in such a manner that it becomes its own "source" (e.g., a record in a database). In comparing such relationships to each other, intersource link analysis is being performed. For instance, if our intrasource example also included the fictitious sentence, "Alfred Einstein, Albert's brother, was born on January 1, 1860 in Ulm, Germany", we would learn the following intrasource relationships:

- <person-birthlocation > Alfred Einstein; Ulm, Germany </person-birthlocation>
- <person-brother> Alfred Einstein, Albert Einstein </person-brother>

From these, the intersource relationship that the brothers were born in the same city could be learned.

2.3 Link Analysis Algorithms/Techniques

This section contains an overview of some common approaches to link analysis as defined in Section 2.1 above. These algorithms can be divided into two general categories: *intrasource* and *intersource* techniques. Examples of intrasource techniques include association rule mining (ARM), Co-occurrence analysis and relation/event extraction. Intersource techniques include higher-order co-occurrence, D-HOTM, Literature-based Discovery (LBD), and Sequence Mining. This overview is not intended to be comprehensive; rather, it touches on a few representative examples of common approaches.

2.3.1 Intrasource techniques

2.3.1.1 Association Rule Mining (ARM)

Association rule mining generates association rules from transactions or records in a database. An *association rule* is expressed as X➔Y, where X is a set of items in the antecedent that implies a set of items Y in the consequent. For example, in a supermarket transaction, milk➔bread, implies that when a customer buys milk, he or she will also buy bread. Even from this simple example, it can be ascertained that association rules will not hold for all transactions. To address this issue support and confidence metrics are normally used to measure the performance of rules. Given a dataset and a threshold, the task of ARM is thus to discover rules whose support and confidence is greater than a specified threshold. Relue et al. (2001), Das et al. (2001), Denwattana and Getta (2001), and Xu et al. (2005) all discuss ARM capabilities.

2.3.1.2 Co-occurrence Analysis

In the context of named entity information extraction, two entities are said to *co-occur* when they occur together in the same source, whether it be a record, document, title, etc. Co-occurrence within a single source (i.e., intrasource) is termed first-order co-occurrence, as compared to higher order co-occurrence (see Section 2.3.2). Co-occurrence is a simple relationship and normally lexical analysis suffices to learn this relationship. To identify relationships, a degree of syntactic and/or semantic analysis is also required. Co-occurrence analysis has been widely used in numerous applications – Hasegawa et al. (2004) and Sundresan (2000) are just a couple of examples.

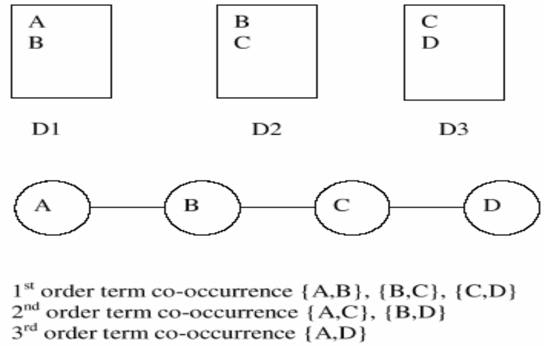
2.3.1.3 Relation/event extraction

Relation and event extraction are also intrasource link analysis technologies. Like methods based on co-occurrence, these approaches require syntactic and/or semantic analysis of the source. An example of *relation extraction* is the extraction of [author, title] pairs from Web pages. Zelenko and Aone (2002), Sundresan and Li (2000), Zhu et al. (2005) and Habegger and Quafafou (2002) are all examples of solutions that perform relation extraction.

Event extraction, on the other hand, aims to extract a predefined event. For example, a user may be interested in a "market change event", such as a change in stock price or a currency devaluation. Surdeanu et al. (2003) is an example of an event extraction solution, and Aone and Ramos-Santacruz (2000) perform both relation and event extraction.

2.3.2 Intersource techniques

2.3.2.1 Higher Order Co-occurrence


A
B
B
C
C
D
D1
D2
D3
A
B
C
D
1st order term co-occurrence {A,B}, {B,C}, {C,D}
2nd order term co-occurrence {A,C}, {B,D}
3rd order term co-occurrence {A,D}

Higher order co-occurrence refers to the occurrence of entities that can be linked through other entities. For example, in the adjacent figure (Kontostathis and Pottenger, 2006), *A* and *B* co-occur in source D1, *B* and *C* co-occur in D2 and *C* and *D* co-occur in D3. This means that *A* and *C* have a second-order co-occurrence relationship through *B* and *A* and *D* have a third-order co-occurrence relationship through *B* and *C*. These co-occurrence relationships can also be represented in a linked graph G = <V, E>, where V are named entities and E are edges. Direct links between vertices in G represent first-order co-occurrence between entities. Paths in G that involve more than two vertices represent higher-order co-occurrence between entities. Several link analysis solutions employ higher order co-occurrence include Li et al. (2005) and Weeber et al. (2001).

2.3.2.2 Sequence Mining

A *sequential pattern* can be defined as a subsequence that appears frequently in a sequence database. Many data sources are inherently sequenced, such as satellite images, DNA sequence data, website usage logs, and supermarket transactions. The purpose of sequence mining is to discover useful sequential knowledge. Patterns discovered vary based on the application. In a supermarket application, an example sequential pattern is "40% of customers who purchase a TV later purchase a VCR" (Hingston, 2001). Once sequence patterns are discovered they can be used to generate association rules. Sequence mining has been used in many different domains, and is an important link analysis technique. Hingston (2001), Kum et al. (2003), Ayres et al. (2002), Pei et al. (2001), and Tumasonis and Dzemyda (2004) are all examples of this approach.

2.4 Overview Conclusion

As can be seen from the preceding sections, the link analysis field is complex. As noted, after considering several categories, we have identified two specific axes (the use of named entities, and the scope level of the document (intra-/intersource)) for use in the evaluation of solutions. The next section utilizes the terms, concepts, and axes presented in this section to analyze a sampling of the solutions available in the link analysis field.

3 Link Analysis Solutions

3.1 Index of Solutions

Below is a list of the solutions surveyed. They have been divided into one of two groups: *Academic solutions* (those which have been or are being developed in colleges, universities, or academic institutions) and *commercial solutions* (those currently available from vendors, companies, or the government). Within these two categories, the solutions are then organized alphabetically to allow for simple searching. The following is an index of the solutions detailed in sections 3.2 and 3.3.

3.2 Academic Solutions

3.2.1 Bar-Ilan University: TEG-A Hybrid Approach to Information Extraction

Solution Introduction and Domain Scope

This solution was developed by researchers at the Bar-Ilan University in Ramat Gan, Israel. It aims to extract named entities and relations from textual data. It is suitable to be used in general domains. We categorize this solution as IE and LA, since, beside named entities, it also extracts relationships among entities.

Output/Results

The outputs are named entities and relationships. For example, person name, organization name, and location name are types of named entities that can be extracted by the solution. If a person is the manager of the company, there is some "ROLE" relationship between this person and the company that would be identified by the solution, as well.

Application to Law Enforcement

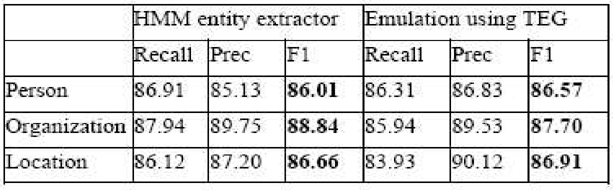
Moderate. This solution is not specially designed for law enforcement applications. The solution cannot be directly used as the named entities extracted in this solution are not comprehensive with respect to the law enforcement domain. However, as with many other IE solutions, it could be used in law enforcement since named entity extraction and relationship extraction are needed to convert narrative reports into structured data.

Evaluation

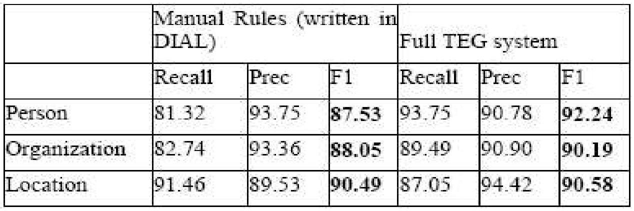
The performance of both named entity extraction and relationship extraction is evaluated in this solution. Named entity extraction is evaluated on MUC-7 data, and the relation extraction is evaluated on ACE-2 data.

The MUC-7 corpus is composed of a set of news articles related to aircraft accidents. It contains 200,000 words and four types of named entities: *person, organization, location,* and *other.* The performance is evaluated against the following entity extractors: the regular HMM, its emulation using TEG, a set of manual rules termed a Trainable Extraction Grammar, and the full TEG system. The performance results are presented in the upper table of the adjacent figure (Rosenfeld et al., 2004).

Table 1. Accuracy Results for MUC 7



	HMM entity extractor			Emulation using TEG		
	Recall	Prec	F1	Recall	Prec	F1
Person	86.91	85.13	**86.01**	86.31	86.83	**86.57**
Organization	87.94	89.75	**88.84**	85.94	89.53	**87.70**
Location	86.12	87.20	**86.66**	83.93	90.12	**86.91**



	Manual Rules (written in DIAL)			Full TEG system		
	Recall	Prec	F1	Recall	Prec	F1
Person	81.32	93.75	**87.53**	93.75	90.78	**92.24**
Organization	82.74	93.36	**88.05**	89.49	90.90	**90.19**
Location	91.46	89.53	**90.49**	87.05	94.42	**90.58**

The relationship extraction capabilities are evaluated on ACE-2 data and the *"ROLE"* relation was chosen to be evaluated. As part of the process, three named entities are also extracted: person, organization, and GPE. The lower table in the adjacent figure shows the performance results.

Software

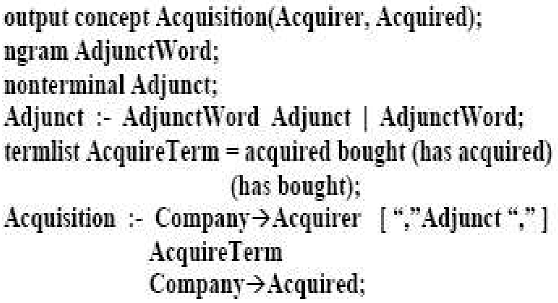
n/a

Inputs Required

Textual data

Link Analysis Algorithm

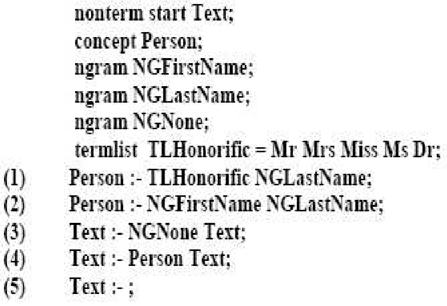
This solution is a hybrid statistical and knowledge-based IE and LA model, and it requires less manual crafting of rules and a smaller amount of training data than other approaches. The solution employs a SCFG (stochastic context-free grammar). Similar to a regular grammar, a string is accepted by a SCFG if the string can be produced from the starting symbol S. The non-terminals in a SCFG are different from the regular grammar. For example, non-terminals could be noun phrases (NP), verb phrases (VP), etc. and the rules define the syntax of the language. For example, S→NP VP. The knowledge engineer writes SCFG rules manually, and then the SCFG rules are trained on the available data. An example of a TEG grammar is provided in the figure below (Rosenfeld et al., 2004):



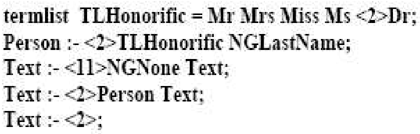
```
output concept Acquisition(Acquirer, Acquired);
ngram AdjunctWord;
nonterminal Adjunct;
Adjunct :- AdjunctWord Adjunct | AdjunctWord;
termlist AcquireTerm = acquired bought (has acquired)
                      (has bought);
Acquisition :- Company→Acquirer [ “,”Adjunct “,” ]
             AcquireTerm
             Company→Acquired;
```

This grammar can be explained as follows: the first line defines a relation “*Acquisition*”, which has two attributes, ***Acquirer*** and ***Acquired***. Next, an ngram ***AdjunctWord*** is defined, which is followed by a non-terminal ***Adjunct***. The ***Adjunct*** has two rules, which are separated by “|”, which means the ***Adjunct*** construct is defined as a sequence of one or more ***AdjunctWord***s. A term list ***AcquireTerm*** is also defined and contains the main verb phrase for acquisition. Finally, the single rule for the ***Acquisition*** concept is defined as a ***company***, which is followed by optional ***Adjunct*** delimited by commas, followed by ***AcquireTerm*** and a second ***Company***.

After the grammar/rules have been created, the resulting TEG is trained. Currently, there are three different trainable parameters in a TEG rulebook: “the probabilities of rules of non-terminals, the probabilities of different expansions of n-grams, and the probabilities of terms in a word class” (Rosenfeld et al., 2004). The initial untrained frequencies of all elements are set to “1” by default; after training, these different element frequencies will be updated to correspond to their actual value. For example, the adjacent figure (Rosenfeld et al., 2004) is a basic TEG grammar to discover simple person names. The rulebook of this grammar would then be trained on a training set containing a single sentence: “*Yesterday, <Person>Dr. Simmons</Person>, the distinguished scientist, presented the discovery.”* After the training process is completed, the result will be a rulebook as presented in the figure below (Rosenfeld et al., 2004).



```
        nonterm start Text;
        concept Person;
        ngram NGFirstName;
        ngram NGLastName;
        ngram NGNone;
        termlist  TLHonorific = Mr Mrs Miss Ms Dr;
(1)     Person :- TLHonorific NGLastName;
(2)     Person :- NGFirstName NGLastName;
(3)     Text :- NGNone Text;
(4)     Text :- Person Text;
(5)     Text :- ;
```



```
termlist  TLHonorific = Mr Mrs Miss Ms <2>Dr;
Person :- <2>TLHonorific NGLastName;
Text :- <11>NGNone Text;
Text :- <2>Person Text;
Text :- <2>;
```

As already stated, the SCFG rules are manually crafted, while the probabilities for each rule are generated from the training data. The approach balances between labeling data and writing rules. For example, more rules generally lead to less labeled data. One advantage of this solution, compared with

HMM, is that relationships among entities can also be determined. HMM is not suitable for finding the relations that exist between entities. Another advantage of this solution is that it can be adapted to any domain by developing SCFG rules and training them.

Knowledge Engineering Cost

The KEC for this solution is high. This solution requires not only labeled training data, but also manually crafted rules. Although the rules are simple, neat, easy to create and a smaller amount of training data is required compared with other pure statistical learning algorithms (e.g., HMM), the process still demands a significant degree of knowledge engineering.

Summary Table

Category: Academic	
University Name: Bar-IIan University **Company URL**: http://www.biu.ac.il/	**Location**: Ramat Gan, Israel
Solution Name: TEG-A Hybrid Approach to Information Extraction	
Domain Scope: general	**Application Type**: IE and LA
Knowledge Engineering Cost: high	**Financial Cost**: n/a
Input Requirements/Preparation Required: Textual data	
Information Extraction **Algorithm Name/Group**: a hybrid of statistical and knowledge-based model **Labeling**: manual **Labeling Supervision**: n/a **Model Generation**: hybrid **Model Generation Supervision**: supervised **Process Description**: This solution is a hybrid statistical and knowledge-based IE and LA model. The SCFG rules are manually crafted, while the probabilities for each rule are learned from training data. The resulting rules are used to extract named entities and relations.	
Solution Output: named entities and predefined relations	
Application to Law Enforcement: moderate	
Is performance evaluation available? yes	**Solution/demo available**? no

Sources

Rosenfeld, Benjamin; Feldman, Ronen; Fresko, Moshe; Schler, Jonathan; and Aumann, Yonatan (2004). "TEG – A Hybrid Approach to Information Extraction." *CIKM'04 Conference (Washington, DC, USA*) November 8-13, 2004, Online. http://delivery.acm.org/10.1145/1040000/1031280/p589-rosenfeld.pdf?key1=1031280&key2=8291408311&coll=GUIDE&dl=GUIDE&CFID=66467799&CFTOKEN=25735454. Accessed January 19, 2006.

3.2.2 Cornell University: Sequential Pattern Mining using a Bitmap Representation

Solution Introduction and Domain Scope

This solution was developed by researchers at Cornell University in Ithaca, NY for mining very long sequential patterns. This solution has applicability in many domains, e.g., business analysis and web mining, as long as the data in the domain is sequential. This solution has been categorized as an intersource link analysis solution, since it aims to discover common sequence patterns among various records/sequences. In other words, it attempts to discover relations from different records, instead of concentrating on relations within a single record.

Output/Results

The output of this solution is sequential patterns.

Application to Law Enforcement

Moderate. In criminal events, many sequential patterns exist, e.g., suspects also perform action B after performing action A. If police officers can recognize this pattern, then an occurrence of action A could aid in preventing an occurrence of action B. However, modifications would be needed to utilize this solution in a law enforcement environment, since the criminal event data would need to be stored in a sequential database format.

Evaluation

Due to the bitmap representation of the data, the solution performs well for large datasets, outperforming SPADE and PrefixSpan (Ayres et al., 2002) by over an order of magnitude. The bitmap representation is efficient for counting, which is a critical component of the analysis. For small datasets, however, PrefixSpan runs faster than this solution since the initial overhead needed to set up and use the bitmap representation outweighs the benefits of PrefixSpan's fast counting.

Aside from the execution time complexity, this solution has a larger space complexity than SPADE. One important assumption made by this solution is that the entire database fit completely into main memory.

Software

n/a

Inputs Required

The input to this solution is a lexicographic tree, which represents all the sequences in a sequence database. A normal sequence database can be transformed into a lexicographic tree in a straightforward manner.

Link Analysis Algorithm

A database D is a set of tuples (customerID, transactionID, itemset). The goal of sequential patterns mining is to discover all frequent sequential patterns in D, given a support threshold (minSup). In this solution, sequences are as noted first represented by a lexicographic tree, in which each node is a sequence. Each sequence within the sequence tree can be considered as either a sequence-extension (S_n) or an itemset-extension (I_n) from its parent (node n). For example, if we have a sequence $s= (\{a\}, \{b\})$, then $(\{a\}, \{b\}, \{a\})$ is a sequence-extension from s, and $(\{a\},\{b, a\})$ is an itemset-extension from s. A depth-first tree traversal algorithm is used to discover frequent patterns in this tree. If the support of a generated sequence $s \geq minSup$, the sequence is stored and the depth-first search (DFS) is repeated recursively on s. If the support of $s < minSup$, it is not necessary to repeat DFS on s, since no child sequence generated from s will be frequent. If all the children of a given node are infrequent, the node is labeled as a leaf and the algorithm continues to check other nodes until all the nodes in the tree have been visited.

This DFS algorithm has a large search space, so this solution employs an Apriori-based pruning algorithm to prune the candidate sequence-extensions and itemset-extensions of each node in the tree. The final aim is to minimize S_n and I_n for each node n, and, at the same time, guarantee that all frequent sequences (nodes) are visited. Counting is a significant fraction of the execution time since it is performed at each recursive step. In order to perform efficient counting, a vertical bitmap representation of the data is used. "A vertical bitmap is created for each item in the dataset, and each bitmap has a bit corresponding to each transaction in the dataset. If item i appears in transaction j, then the bit corresponding to transaction *j* of the bitmap for item *i* is set to one; otherwise, the bit is set to zero" (Ayres et al., 2002).

Knowledge Engineering Cost

The KEC for this solution is low since it needs neither labeled data nor manually crafted rules. The sequence pattern mining algorithm is automatic, supporting the conclusion that the KEC is low. In general, the KEC of unsupervised learning approaches is low; the caveat is that additional human effort may be required to interpret the model.

Summary Table

Category: Academic	
Hierarchy: not NE	**Source Scope**: Inter
Institution Name: Cornell University **Institution URL**: http://www.cornell.edu/	**Location**: Ithaca, NY, USA
Solution Name: Sequential Pattern Mining using a Bitmap Representation (SPAM)	
Domain Scope: general	**Application Type**: LA
Knowledge Engineering Cost: low	**Financial Cost**: n/a
Input Requirements/Preparation Required: Any sequence data	
Link Analysis **Algorithm Name/Group**: First Tree Traversal algorithm refined by DFS-Pruning algorithm and bitmap representation of data **Labeling**: n/a **Labeling Supervision**: n/a **Model Generation**: automatic **Model Generation Supervision**: unsupervised **Process Description**: Given a lexicographic tree, a DFS tree traversal algorithm is used to discover the frequent patterns. Pruning reduces the large search space.	
Solution Output: sequential patterns	
Application to Law Enforcement: moderate	
Is performance evaluation available? Yes	**Solution/demo available**? no

Sources

Ayres, Jay; Flannick, Jason; Gehrke, Johannes and Yiu, Tomi (2002). "Sequential Pattern Mining Using a Bitmap Representation." *Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining.* Online. http://www.cs.cornell.edu/johannes/papers/2002/kdd2002-spam.pdf. Accessed January 11, 2006.

3.2.3 Massachusetts Institute of Technology and Cornell University: The Link Prediction Problem for Social Networks

Solution Introduction and Domain Scope

This solution is a mini-survey. Several link prediction methods were evaluated, and their performance compared. This work was completed by researchers at the Massachusetts Institute of Technology in Cambridge, Massachusetts and at Cornell University in Ithaca, NY. This solution has applicability in many domains. For example it could be applied in criminal suspect prediction.

Output/Results

The output is social relationships that are predicted to occur.

Application to Law Enforcement

Moderate. Although apparently not applied in law enforcement, this solution could be used because events could be predicted from a social network comprised of crime events, organizations, and personal names and relationships.

Evaluation

The test data used in this solution is a co-authorship network *G* which is constructed from papers in five sections of the physics e-Print arXiv. Several factors are compared to determine the improvement over random predictions. Some of the methods significantly outperformed random prediction, which supports the conclusion that the social network topology contains useful information. Further detail is available in the references listed below in the Sources section.

Software

n/a

Inputs Required

The input is a snapshot of social network data structures, where nodes represent people or other entities embedded in a social context and edges represent interaction, collaboration, or influence among the entities.

Link Analysis Algorithm

Each of the link prediction methods reviewed calculates the proximity or “similarity” between pairs of nodes. The following are the main approaches used for link prediction:

Methods based on node neighborhoods: The basic idea is that two nodes are more likely to form a link if they share the same neighborhoods. Newman (2001), Jin and Newman (2001) and Adamic and Adar (2003) leverage this approach.

Methods based on the ensemble of all paths: This approach refines a shortest-path method, which is a simple way to measure the similarity between two nodes. Since shortest paths often result in short chains between nodes, this method incorporates all paths between nodes with a shortest-path approach. Jeh and SimRank (2002) and Katz (1953) employ this approach.

Meta-level approaches: Certain approaches can be combined; for example clustering can be used to delete the superfluous edges in social network formed using one of the above methods.

Knowledge Engineering Cost

The KEC for link prediction differs depending on the method employed. Nonetheless, social relationship link prediction as reviewed herein does not require labeled data, and the algorithms are automatic. Thus, we rate the KEC as low.

Summary Table

Category: Academic	
Hierarchy: NE	**Source Scope**: Inter
Institution Name: Massachusetts Institute of Technology (MIT); Cornell University **Institution URL**: http://web.mit.edu/ http://www.cornell.edu/	**Location**: Cambridge, Massachusetts, USA; Ithaca, New York, USA
Solution Name: The Link Prediction Problem for Social Networks	
Domain Scope: general	**Application Type**: LA
Knowledge Engineering Cost: low	**Financial Cost**: n/a
Input Requirements/Preparation Required: social network	

(include formats supported – if important)	
Link Analysis **Algorithm Name/Group**: differs depending on the method used. **Labeling**: n/a **Labeling Supervision**: n/a **Model Generation**: automatic **Model Generation Supervision**: unsupervised **Process Description**: Calculates the proximity or "similarity" between pairs of nodes. A variety of different methods are presented.	
Solution Output: relationships which will happen in the near future	
Application to Law Enforcement: moderate	
Is performance evaluation available? yes	**Solution/demo available**? no

Sources

Newman, M. (2001). "Clustering and Preferential Attachment in Growing Networks." *Phys. Rev. E.* 64(025102).

Jin, E; Girvan, M. and Newman, M. (2001). "The Structure of Growing Social Networks." *Phys. Rev. E.* 64(046132).

Adamic, L. and Adar, E. (2003). "Friends and Neighbors on the Web." *Soc. Networks*. 25(3).

Jeh, G. and Widom, J. (2002). "SimRank: A Measure of Structural-Context Similarity." *In KDD 2002.*

Katz, L. (1953). "A New Status Index Derived from Sociometric Analysis." *Psychometrika*. 18(1), March 1953.

3.2.4 Institute of Mathematics and Informatics: ProMFS

Solution Introduction and Domain Scope

In this solution a new probabilistic algorithm for mining frequent sequences (ProMFS) is proposed. It was developed by researchers at the Institute of Mathematics and Informatics, Vilnius, Lithuania. It uses probabilistic-statistical characteristics of sequence databases to build new, much shorter sequences and later mines frequent sequences from them. This solution is applicable in the same domains as other sequence mining solutions. It has been categorized as an intersource link analysis solution since it aims to mine the frequent patterns over all sequences.

Output/Results

Frequent sequences

Application to Law Enforcement

Moderate. In criminal events, many sequential patterns exist, e.g., suspects also perform action B after performing action A. If police officers can recognize this pattern, then an occurrence of action A could aid in preventing an occurrence of action B. However, modifications would be needed to utilize this solution in a law enforcement environment, since the criminal event data would need to be stored in a sequential database format.

Evaluation

This probabilistic algorithm was compared with the Generated Sequence Pattern (GSP) algorithm. If the minimum support (minSup) was small (below 2500), GSP discovered many more frequent sequences than ProMFS. When minSup was in the range of [2500, 6000], the number of frequent sequences discovered by GSP and ProMFS was similar. As the minimum support grew, the number of frequent sequences discovered by both algorithms became identical.

In comparing the time complexity of these two completing algorithms, it was concluded that ProMFS executes significantly faster. When minSup was in the range of [2500, 6000], ProMFS needed approximately 20 times less compute time and achieved similar results.

Software

n/a

Inputs Required

A large volume of sequences, which could be a sequential database or texts which contain string sequences.

Link Analysis Algorithm

If L= $\{\boldsymbol{i}_{1,} \boldsymbol{i}_{2,\ldots,} \boldsymbol{i}_{\mathrm{m}}\}$ is the set consisting of m distinct elements, then a sequence is formed from the elements of the set L. The basic idea of this algorithm is to use a much shorter sequence (termed the *model sequence*) to represent the entire input, and then apply GSP or a similar algorithm to mine frequent sequences on the model sequence. Model sequences are generated using probability information calculated from the input sequences.

In particular, to generate a model sequence the input sequences are first evaluated and the following characteristics determined:

1. *The probability of occurrence of element* $\mathbf{i}_j$ *in the input sequences (j=1…..m),*
2. *The probability of appearance of element* $\mathbf{i}_v$ *after element* $\mathbf{i}_j$ *(v, j=1…..m),*
3. *The distance between elements* $\boldsymbol{i}_j$ *after* $\boldsymbol{i}_v$ *element (j=1…..m), and*
4. *The matrix of average distances.*

These characteristics can be obtained during one pass through the input sequences, and a significantly shorter model sequence is generated based on these probabilities.

Knowledge Engineering Cost

The KEC is low as with other sequence mining solutions. No labeling work is needed to generate the model sequence, and the frequent sequence mining algorithm itself is automatic.

Summary Table

Category: Academic	
Hierarchy: not NE	**Source Scope**: Inter
Institution Name: the Institute of Mathematics and Informatics **Institution URL**: http://www.math.bas.bg/	**Location**: Vilnius, Lithuania
Solution Name: A Probabilistic Algorithm for Mining Frequent Sequences	
Domain Scope: general	**Application Type**: LA
Knowledge Engineering Cost: low	**Financial Cost**: n/a
Input Requirements/Preparation Required: a set of sequences	

Link Analysis **Algorithm Name/Group**: probabilistic-statistical information with GSP algorithm **Labeling**: n/a **Labeling Supervision**: n/a **Model Generation**: automatic **Model Generation Supervision**: unsupervised **Process Description**: First a *model sequence* is generated to represent the input, and then the GSP algorithm, or similar sequence mining algorithm, is used to mine frequent sequences based on the model sequence.	
Solution Output: frequent sequences	
Application to Law Enforcement: moderate	
Is performance evaluation available? yes	**Solution/demo available**? no

Sources

Tumasonis, Romanas and Dzemyda, Gintautas (2004). "A Probabilistic Algorithm for Mining Frequent Sequences." *Eighth East-European Conference on Advances in Databases and Information Systems*. Budapest, Hungary. Online. http://www.sztaki.hu/conferences/ADBIS/8-Tumasonis.pdf Accessed January 21, 2006.

3.2.5 Lehigh University: D-HOTM

Solution Introduction and Domain Scope

The Distributed Higher-Order Text Mining (D-HOTM) solution was developed by researchers at Lehigh University, Bethlehem, PA. It aims to discover higher-order association rules within a distributed environment. This solution can be applied in several domains, as long as the aim is to discover higher order association rules in distributed databases. For example, it can be used in law enforcement to coordinate various departments' data in order to counter criminal activities. This is both an IE and LA solution, since LA is used on named entities extracted from textual data, and the named entities are extracted using an IE tool discussed in Pottenger et al. (2006a). D-HOTM performs intersource LA, since it aims to link entities that belong to records in distributed databases.

Output/Results

The output of this solution is higher order association rules.

Application to Law Enforcement

Extensive. This solution is useful in law enforcement, since a challenging problem for the law enforcement community is the disorganized and fragmented nature of data. As data is not (and cannot easily be) centralized and unified, different police jurisdictions contain different information about a given suspect. D-HOTM coordinates this information and discovers links between named entities.

Evaluation

This solution is in the early stages of development both theoretical and algorithm in nature, and as such evaluation has not yet been completed. This solution employs an Information Extraction (IE) solution which is analyzed in Pottenger et al. (2006a). The solution does however present new metrics for evaluation in a distributed environment.

Inputs Required

The input to this solution is distributed databases. Each database may have its own schema, but they share a set of common items.

Link Analysis Algorithm

D-HOTM is an approach which combines information extraction and distributed data mining together. First a semi-supervised information extraction algorithm is used to extract linguistic features/entities from textual data. Then a distributed higher-order association rule mining (DiHO ARM) algorithm based on Apriori is used to mine the association rules.

Compared to current algorithms which mine distributed data, an advantage of D-HOTM is that it does not require knowledge of a global schema across all databases. Since the features/entities extracted from different distributes databases may differ, it is impractical to rely on knowledge of a global schema. D-HOTM also addresses data fragmentation issues in which records are neither horizontally nor vertically distributed; current Distributed Association Rule Mining (D-ARM) algorithms assume that databases are either horizontally or vertically fragmented.

The DiHO ARM algorithm comprises five steps:

1. *Select linkage items*
2. *Assign a globally unique ID to each record/object*
3. *Identify linkable records using Apriori on global IDs*
4. *Exchange information about linkable records*
5. *For each site*
    a. *Apply the Apriori algorithm locally*

An important theoretical aspect of the DiHO ARM algorithm is the detection of linkable records. In this solution, two types of links are defined: first-order links and higher-order links. A first-order link is a direct link; for example, if there is a record which contains information about a stolen vehicle, and VIN (Vehicle Identification Number) and Owner are two items in this record, then the link between VIN and Owner is a first-order link. A higher-order link is an indirect link that involves more than one record and uses a particular item to link records. For example, two different records may have the same VIN, so the two records are linkable through the VIN. D-HOTM mines higher order links instead of just first-order links. The framework is based on a proof of the fact that "the maximum frequent item sets generated using Apriori on the subset of items used to link records is sufficient to identify all linkable records" (Li et al., 2005).

Knowledge Engineering Cost

The IE phase of this solution employs a semi-supervised algorithm, which learns regular expression rules from labeled segments. This semi-supervised learning algorithm reduces the KEC compared with other IE solutions, since it only requires that segments be labeled, not individual features. The extracted named entities become input to the DihO ARM link analysis algorithm. DiHO ARM is unsupervised and automatic, so its KEC is low. The total KEC for this solution is thus low to medium.

Summary Table

Category: Academic	
Hierarchy: NE	**Source Scope**: Inter
Institution Name: Lehigh University **Institution URL**: http://www.cse.lehigh.edu/~billp/	**Location**: Bethlehem, PA, USA
Solution Name: Distributed Higher Order Association Rule Mining Using Information Extracted from Textual Data	
Domain Scope: general	**Application Type**: IE and LA

Knowledge Engineering Cost: low/medium	**Financial Cost**: n/a
Input Requirements/Preparation Required: distributed databases, which share a set of items.	
Link Analysis **Algorithm Name/Group**: DiHO: a higher-order association rule mining algorithm based on Apriori. **Labeling**: manual **Labeling Supervision**: n/a **Model Generation**: automatic **Model Generation Supervision**: unsupervised **Process Description**: An IE solution extracts information from textual data on distributed sites and populates (distributed) databases. Linkage items are selected and a globally unique ID is assigned to each record. Apriori is used to identify linkable records. Each site exchanges information about linkable records and applies the Apriori algorithm locally to discover rules.	
Solution Output: higher order association rules	
Application to Law Enforcement: extensive	
Is performance evaluation available? no	**Solution/demo available**? no

Sources

Li, S.; Wu, T and Pottenger, W. M. (2005) "Distributed Higher Order Association Rule Mining Using Information Extracted from Textual Data." *SIGKDD Explorations*. Volume 7, Issue 1, June 2005. Online. http://www.cse.lehigh.edu/~billp/pubs/SIGKDDExplorations.pdf Accessed January 21, 2006.

3.2.6 Middlesex University: Improving Knowledge Discovery by Combining Text-Mining and Link-Analysis Techniques

Solution Introduction and Domain Scope

This solution was developed by researchers at the School of Computing Science at Middlesex University, London and other two other universities. It aims to discover new knowledge from features extracted from online text documents. An example application is the extraction of knowledge from online news, although the solution could be used in other domains.

Output/Results

The output is new knowledge. More specifically the solution discovers the links between entities extracted from a textual data source.

Application to Law Enforcement

Moderate. This solution is suitable for use in law enforcement since entities could be extracted from police reports, then links mined from the extracted entities. The solution would however need to be adapted to this domain since the entity types differ.

Evaluation

The solution was implemented and evaluated using the ClearForest suite (see Section 3.3.4). Data was collected from four news sites: CNN, BBC, CBS, and Yahoo and two methods for link analysis were compared, one based on co-occurrences and a second based on semantics. Questions such as "*How many meetings did Arial Sharon and Colin Powel have?*" were posed. The results are summarized in the tables below (Moty Ben-Dov et al., 2004), in which the method based on co-occurrence achieved high recall and low precision while that based on deeper semantics resulted in high precision and low recall.

Table 1. The Q1 results

	Co-occurrence links	Semantic links
Correct links	20	8
Total Correct links in the database	22	22
Total links	71	9
Precision	28.17%	88.89%
Recall	90.91%	36.36%

Table 2. The Q2 results

	Co-occurrence links	Semantic links
Correct links	14	5
Total Correct links in the database	15	15
Total links	94	6
Precision	14.89%	83.33%
Recall	93.33%	33.33%

Table 3. The Q3 results

	Co-occurrence links	Semantic links
Correct links	8	5
Total Correct links in the database	11	11
Total links	9	5
Precision	88.89%	100%
Recall	72.73%	45.45%

Inputs Required

The input is textual data in the form of online documents or local documents. After IE, the output is entities (features) extracted from the texts. These entities then become the input to the link analysis process.

Link Analysis Algorithm

This solution employs two approaches to link analysis, one based on co-occurrence and another based on deeper semantics. This solution identifies a co-occurrence relationship when two features occur in the same sentence. The co-occurrence-based link can be discovered by identifying lexical features in the sentence. No syntactic or semantic analysis is required. The adjacent figure (Moty Ben-Dov et al., 2004) is an example of a relation map of co-occurrence between persons. The darker the line between two persons, the more co-occurrence links.

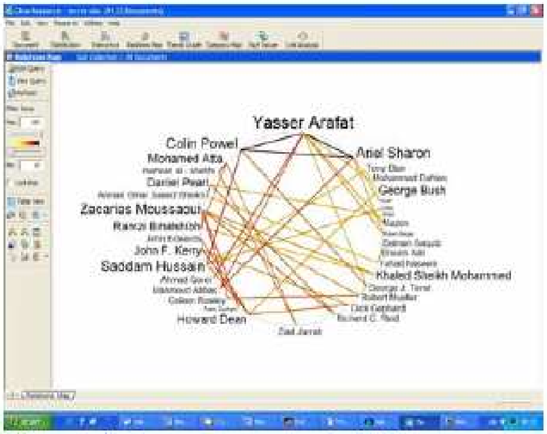


Figure 1. Co-occurrence relations between persons.

Semantic links are created using noun phrase and verb identification as well as linguistic and semantic constraints. Predefined semantic relations are extracted by performing a deep syntactic and semantic analysis of the sentence. A strategy with five layers is used:

- Layer 0 – Part-of-Speech Tagger
- Layer 1 – Grouping Noun and Verb Phrases
- Layer 2 – Extracting Verb and Noun Patterns
- Layer 3 – Recognizing Named Entities
- Layer 4 – Relationship Extraction

Implementation of this semantics-based approach was done using ClearForest's DIAL (Declarative Information Analysis Language). The adjacent figure (Moty Ben-Dov, 2004) gives an example of the semantic relations between persons. In this graph, the threshold for displaying a semantic relation between two persons is three links.

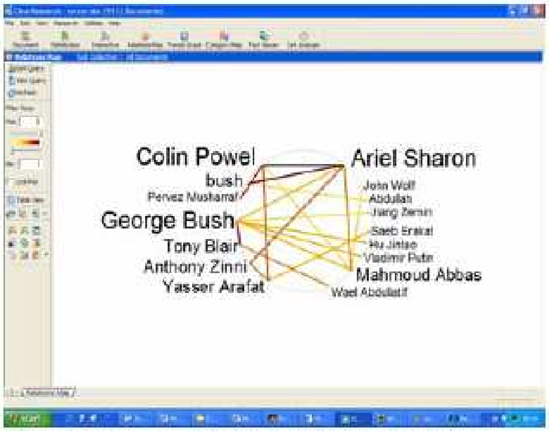


Figure 2. Semantic relations between persons.

Knowledge Engineering Cost

Both of the approaches developed involve the development of manually crafted rules; in the first case, for named entity extraction, and in the second, for semantic analysis. Although the co-occurrence analysis has a low KEC, due to the rule development cost the overall KEC is high.

Summary Table

Category: Academic and Commercial	
Hierarchy: NE	**Source Scope**: Intra and Inter
Institution Name: Middlesex University **Institution URL**: http://www.mdx.ac.uk/	**Location**: Landon, UK
Solution Name: Improving Knowledge Discovery by Combining Text-Mining And Link-Analysis Techniques	
Domain Scope: many domains	**Application Type**: LA
Knowledge Engineering Cost: high	**Financial Cost**: n/a
Input Requirements/Preparation Required: textual data	
Link Analysis **Algorithm Name/Group**: statistical method based on co-occurrence, or semantic analysis based on noun- and verb-phrases **Labeling**: manual **Labeling Supervision**: n/a **Model Generation**: manual **Model Generation Supervision**: n/a **Process Description**: In the co-occurrence-based approach IE rules are crafted, terms are filtered and indexed and co-occurrence relationships among terms are captured. In the semantics-based approach, LA rules are crafted and applied to the data.	
Solution Output: links between features	
Application to Law Enforcement: moderate	
Is performance evaluation available? yes	**Solution/demo available**? no

Sources

Ben-Dov, M.; Wu, W. and Feldman, R. (2004). "Improving Knowledge Discovery by Combining Text-Mining and Link Analysis Techniques." *SIAM Int. Conf. on Data Mining*. Online. http://www.uclic.ucl.ac.uk/paul/research/Moty1.pdf. Accessed January 25, 2006.

3.2.7 Open University and Federal University of Santa Catarina: CORDER

Solution Introduction and Domain Scope

The COmmunity Relation Discovery through Entity Recognition (CORDER) solution was developed by researchers at the Open University, UK and the Federal University of Santa Catarina, Brazil. It is able to discover social networks from an organization's documents. Based on co-occurrence relationships, CORDER finds relations between a target named entity and other named entities. As CORDER is suitable for general domain applications to mine social networks (e.g., corporate organization structure, work relationships (who works for whom)), it is suitable for use in many domains. At this point, its focus is limited to Web pages, instead of all narrative data. CORDER has been categorized as an intrasource LA, since it uses co-occurrence as the relationship between entities, and two named entities are considered to co-occur if they appear in the same web page. No higher order co-occurrence is used, so no intersource LA occurs.

Output/Results

The output is a social network. For example, given a certain person, it is able to find the most important persons related to the given person.

Application to Law Enforcement

Moderate. The solution uses ESpotter to recognize named entities from organizations' Web pages, and then set up social network based on the named entities extracted. However, this solution could also use some other information extraction tool to extract named entities to expand its capabilities. This would allow it to mine social network information based on any extracted named entities. A final obstacle to the use of this solution in the law enforcement community is that it considers *page relevance* as a factor for relation strength, which normally doesn't exist for narrative police reports. This factor would need to be removed or modified if CORDER were to be used.

Evaluation

Experiments were performed on the authors' department web site, and developers claim that CORDER can discover relations with high precision, recall and ranking accuracy. However, no exact performance results reported.

Software

A demo is available at http://kmi.open.ac.uk/people/jianhan/CORDER/.

Inputs Required

Web pages.

Link Analysis Algorithm

CORDER discovers social networks from web pages belonging to an organization. It uses ESpotter (Zhu et al., 2005) to perform named entity recognition, and is able to recognize people, project, organization and research area names, etc. The extracted named entities (NEs) are clustered to remove duplicates. CORDER also leverages co-occurrence information combined with other factors to calculate the relationship strength, e.g., distance, frequency, and page relevance.

The process of CORDER comprises the following steps:

1. *Data selection* – A web spider is used to obtain web pages from a website. Noisy data and irrelevant information are removed.
2. *Named Entity Recognition* – ESpotter is used to perform named entity recognition. After the entities are extracted, a clustering method is used to group similar NEs together, which aims to find the variants and align them.
3. *Relation Strength and Ranking* – For each given target NE, the relationship strength between this target NE and other entities is calculated by taking into account four different aspects:

a. *Co-occurrence* – Two NEs are considered to *co-occur* if they appear in the same web page. The more the number of co-occurrences, the more related the entities are.

b. *Distance* – If two NEs are highly related they tend to occur in close proximity to each other in the web pages. For example if NE_1 and NE_2 both only occur once, then the distance (NE_1, NE_2) = the offsets of NE_1 and NE_2. For other situations, please see Zhu et al. (2005) for more details.

c. *Frequency* – An NE is considered to be more important if it occurs many times on a Web page.

d. *Page relevance* – Given a target NE (NE_1), the weight of each page is assigned to indicate its relevance in association to other NEs co-occurring with NE_1 on the page. For example, a high relevance weight for a person might be set to their homepage and a low relevance weight to their blog page.

Given the target NE (NE_1) and another NE (NE_2), the relationship strength R (NE_1, NE_2) between NE_1 and NE_2 can be calculated as (Zhu et al., 2005):

$$R(E1, E2) = \hat{p}(E1, E2) \times \sum_i \left(\frac{w_i \times f(Freq_i(E1)) \times f(Freq_i(E2))}{\bar{d}_i(E1, E2)} \right)$$

After the relationship strength between NEs is calculated, they are ranked. Then, for each type (person, organization, etc.), a sub-list is created.

Knowledge Engineering Cost

If only the construction of the social network from named entities is considered, the KEC is low, since the relation strength can be automatically calculated and ranked. However, since the social network analysis is based on extracted named entities, the KEC should also take into account the NE extraction algorithm. Since the solution uses ESpotter (whose KEC is high), the total KEC of CORDER is high.

Summary Table

Category: Academic	
Hierarchy: NE	**Source Scope**: Intra
Institution Name: the Open University; Federal University of Santa Catarina **Institution URL**: http://www.open.ac.uk/ http://www.ufsc.br/	**Location**: Milton Keynes, UK; Santa Catarina, Brazil
Solution Name: CORDER: COmmunity Relation Discovery by named Entity Recognition	
Domain Scope: general (web pages)	**Application Type**: LA
Knowledge Engineering Cost: low/high	**Financial Cost**: n/a
Input Requirements/Preparation Required: Web pages	
Link Analysis **Algorithm Name/Group**: co-occurrence and distance to calculate entity relationships **Labeling**: n/a **Labeling Supervision**: n/a **Model Generation**: automatic **Model Generation Supervision**: unsupervised **Process Description**: First, ESpotter is used to recognize named entities from Web pages. Then, by using co-occurrence and a distance metric, the relationship strength between NEs is calculated. The NEs are then ranked.	
Solution Output: social network	
Application to Law Enforcement: moderate	
Is performance evaluation available? no	**Solution/demo available**? Yes

Sources

Zhua, J.; Goncalves, A.; Uren, V.; Motta, E. and Pachecob, R.. (2005). "CORDER: COmmunity Relation Discovery by Named Entity Recognition." *K-CAP'05.* October 2–5, 2005. Banff, Alberta, Canada. Online. http://kmi.open.ac.uk/people/jianhan/s32-zhu.pdf Accessed January 21, 2006.

3.2.8 Simon Fraser University and Hewlett Packard Labs: PrefixSpan

Solution Introduction and Domain Scope

PrefixSpan (Mining Sequential Patterns Efficiently by Prefix-Projected Pattern) was developed by researchers at Simon Fraser University, Canada and Hewlett-Packard Labs, Palo Alto, CA. The solution utilizes a novel method for sequential pattern mining. This solution is suitable in many domains. It has been categorized as an intersource link analysis solution as it is a sequence mining solution.

Output/Results

Frequent sequences

Application to Law Enforcement

Moderate. In criminal events, many sequential patterns exist, e.g., suspects also perform action B after performing action A. If police officers can recognize this pattern, then an occurrence of action A could aid in preventing an occurrence of action B. However, modifications would be needed to utilize this solution in a law enforcement environment, since the criminal event data would need to be stored in a sequential database format.

Evaluation

The data used to evaluate the performance of this algorithm is artificial synthetic customer transaction data generated by the data generation program described in Pei et al. (2001). Four methods were compared: GSP, FreeSpan, PrefixSpan-1 (with level by level projection), and PrefixSpan-2 (with bi-level projection). All were tested on various datasets, and the solutions' performances proved to be consistent. Their results show that PrefixSpan is more efficient and scalable than both FreeSpan and GSP when the support threshold is low and there are many long patterns. PrefixSpan-2 was even more efficient than PrefixSpan-1 in large databases with a low support threshold since PrefixSpan-2 uses bi-level projection to dramatically reduce the number of projections. The experimental results are consistent with the theoretical analysis.

Software

n/a

Inputs Required

A set of sequences is the only requirement.

Link Analysis Algorithm

Many current sequence mining mechanisms, such as GSP and FreeSpan, use Aprioi to reduce the number of candidate frequent sequences. Apriori however uses a multiple-pass, candidate-generation-and-test approach, which can result in a large candidate sequence space and multiple scans of the database. This solution develops a method which leverages the strengths of Apriori but substantially reduces the cost of the candidate-generation-and-test phase.

PrefixSpan was developed partially on the basis of FreeScan, which is a sequential mining method developed by the Institute of Mathematics and Informatics (see Section 3.2.4). The basic idea of FreeScan is to recursively project the sequence database using frequent items. In each projected database subsequence, fragments can be grown. The dataset and frequent patterns which need to be tested are partitioned, and each candidate pattern is tested only on the projected database, which is much smaller than the original database.

The major cost of using FreeSpan had to deal with projected databases. PrefixSpan was designed to lower this cost. “Its general idea is to examine only the prefix subsequences and project only their corresponding postfix subsequences into projected databases. In each projected database, sequential patterns are grown by exploring only local frequent patterns. To further improve mining efficiency, two kinds of database projections are explored: level-by-level and bi-level projection” (Pei et al., 2001).

Knowledge Engineering Cost

The KEC for this solution is low like other sequence mining solutions. The sequence pattern mining algorithm is unsupervised, so neither labeled data nor manually crafted rules are needed.

Summary Table

Category: Academic and Commercial	
Hierarchy: not NE	**Source Scope**: Inter
Institution Name: the Simon Fraser University; Hewlett-Packard Labs **Institution URL**: http://www.sfu.ca/ http://www.hpl.hp.com/	**Location**: Burnaby, British Columbia, Canada; Palo Alto, CA, USA
Solution Name: PrefixSpan: Mining Sequential Patterns Efficiently by Prefix-Projected Pattern Growth	
Domain Scope: general	**Application Type**: LA
Knowledge Engineering Cost: low	**Financial Cost**: n/a
Input Requirements/Preparation Required: a set of sequences	
Link Analysis **Algorithm Name/Group**: Refined Apriori **Labeling**: n/a **Labeling Supervision**: n/a **Model Generation**: automatic **Model Generation Supervision**: unsupervised **Process Description**: The process only examines the prefix subsequences and projects their corresponding postfix subsequences into projected databases. In each projected database, sequential patterns are grown by exploring only locally frequent patterns.	
Solution Output: sequences patterns	
Application to Law Enforcement: moderate	
Is performance evaluation available? yes	**Solution/demo available**? no

Sources

Pei, Jian; Han, Jiawei; Mortazavi-Asl, Behzad; Pinto, Helen; Chen, Qiming; Dayal, Umeshwar and Hsu, Mei-Chun (2001). “PrefixSpan: Mining Sequential Patterns Efficiently by Prefix-Projected Pattern Growth.” *In Proc. 2001 Int. Conf. Data Engineering (ICDE’01).* Heidelberg, Germany. Pages 215-224. Online. http://www-sal.cs.uiuc.edu/~hanj/pdf/span01.pdf Accessed January 21, 2006.

3.2.9 University of Arizona: CopLink®

Solution Introduction and Domain Scope

CopLink® was developed by researchers at the University of Arizona, Tucson, AZ. The CopLink system uses concept space techniques to discover the relationships between data. The paper “Using CopLink to Analyze Criminal-Justice Data” introduces how to construct a concept space and use the CopLink solution (Hauck et. al., 2002). CopLink is primarily designed for the law enforcement field. Its technology could be expanded to other domains, but the CopLink website only mentioned its usage for criminal investigations. Therefore, it would need to be adapted to other domains. We categorize CopLink as both intra- and intersource LA, as the basic relationship it discovers is the co-occurrence relationship between entities (intra), but global statistics are gathered about co-occurrence in a collection (inter). This solution also provides criminal network visualization capabilities.

Output/Results

The output is a concept space (criminal network). Given a specific concept, the other concepts related to it can be discovered.

Application to Law Enforcement

Extensive. CopLink was especially designed for law enforcement field. It aims to accelerate criminal investigations and enhance law enforcement efforts. It represents a combination of the expertise of the University of Arizona's Artificial Intelligence Lab with law enforcement domain knowledge from the Tucson Police Department.

Evaluation

CopLink has been successfully deployed at the Tucson Police Department, where its concept space completed 86% of 965 searches in less than three seconds. In addition, users can become proficient with the system within minutes of training. Perhaps most importantly, the CopLink concept space has been shown to be effective in the investigation of real crimes.

Inputs Required

The input is meaningful non-phrase terms originating from either unstructured or structured (database) data. These terms are also referred to as concepts or objects, and the CopLink solution has developed five categories of such concepts: Person, Location, Organization, Crime and Vehicle.

Link Analysis Algorithm

The underlying structure of CopLink is the *concept space*, which is a statistics-based algorithmic technique that discovers relationships between objects. A concept space could also be considered a network where nodes are the objects of interest and edges are the associations between objects. Edges have weights to represent the strength of relationships between objects. In CopLink the objects come from the criminal reports: i.e., the meaningful terms in the reports. CopLink aims to find the relationships between objects in the entire database, instead of being limited to one criminal report.

Generally, there are three steps required to build a domain-specific concept space. First, the sources from which the terms/objects/concepts will be derived must be located. Second, terms must be filtered and indexed, and some analysis used to capture the co-occurrence relationships among terms. Finally, the resulting concept space is inserted into a database for easy manipulation with an appropriate algorithm.

Theoretically, a concept space may contain as many term types as desired. In practice, considering the size of database and the query time, the number must be limited. The CopLink concept space contains five main categories: Person, Organization, Location, Crime and Vehicle. After identifying terms, the term frequency and document frequency for each term in a document is computed, basing on the methodology mentioned in Hauck et al. (2002). CopLink also provides a user interface that allows users to search from any of four search forms: Person, Location, Organization and Vehicle.

Knowledge Engineering Cost

The KEC for link analysis is low, since the technique used for building the concept space is a co-occurrence-based statistical approach, which is both automatic and does not need manually crafted rules or labeled data. There is however a cost associated with entity extraction from unstructured data which increases the KEC to high (see Pottenger et al., 2006a).

Summary Table

Category: Academic and Commercial	
Hierarchy: not NE (noun-phrase)	**Source Scope**: Intra and Inter
Institution Name: University of Arizona **Institution URL**: http://ai.bpa.arizona.edu/index.html	**Location**: Tucson, AZ, USA
Solution Name: CopLink	
Domain Scope: law enforcement	**Application Type**: LA
Knowledge Engineering Cost: low/high	**Financial Cost**: n/a
Input Requirements/Preparation Required: Non-phrase meaningful terms	
Link Analysis **Algorithm Name/Group**: statistical calculation based on co-occurrence relationship **Labeling**: n/a **Labeling Supervision**: n/a **Model Generation**: automatic **Model Generation Supervision**: unsupervised **Process Description**: First terms are filtered and indexed and co-occurrence relationships are captured. The resulting concept space is inserted into a database for easy manipulation with an appropriate algorithm.	
Solution Output: concept space	
Application to Law Enforcement: extensive	
Is performance evaluation available? yes	**Solution/demo available**? no

Sources

Hauck, Roslin V.; Atabakhsh, Homa; Ongvasith, Pichai; Gupta, Harch and Chen, Hsinchun (2002). "Using CopLink to Analyze Criminal-Justice Data." *Computer* . Volume 35, Issue 3. March 2002. Pages: 30-37. Online. http://ai.bpa.arizona.edu/go/intranet/papers/CopLinkAnalyzeCriminalData.pdf Accessed January 25, 2006.

3.2.10 University of Arizona: Fighting organized crimes

Solution Introduction and Domain Scope

The solution presented in Xu and Chen (2004) was developed by researchers at the University of Arizona, Tucson, AZ. This solution uses link analysis techniques to aid law enforcement and intelligence agencies in the fight against organized crimes including terrorism and kidnapping. The link analysis technique developed in this solution uses shortest-path algorithms, priority-first-search (PFS) and two-tree PFS to identify the most important association paths between given entities in a criminal network. Although this solution is designed for use in the law enforcement field, it could also be used in other domains, such as in the bibliographic area to discover relationships between authors. We categorize this solution as intersource LA, since it extracts relationships between entities which may belong to different sources.

Output/Results

The output is the important association paths between entities. For example, given two entities, Person A and Person B, the returned result could be, A-C-B, which means A and B may be related through person C.

Application to Law Enforcement

Extensive. This solution is especially designed as a law enforcement application.

Evaluation

The dataset used in the evaluation was a year's worth of crime reports provided by the Phoenix Police Department, totaling to 1 GB of data. Two organized crimes were selected as test bed samples: kidnapping (4.5 MB of data) and narcotics (38 MB). Reports with less than five lines were removed from the datasets, since the length of a report can affect the co-occurrence weights of the concepts it contains. The paths found by shortest-path algorithms were compared with the paths found by a modified breadth first search (BFS) algorithm. Precision was used to evaluate the performance, where precision is defined as the number of useful paths selected by experts divided by the total number of paths returned by the algorithm. The results show that, on average, the shortest-path algorithms returned more useful paths than the modified BFS algorithm. The precision was 70% for kidnapping and narcotics networks, while BFS only achieved a 30% precision for the kidnapping network and 16.7% precision for the narcotics network. For the two shortest-path algorithms (modified PFS and two-tree PFS), their efficiency (average execution time) was also compared. For the kidnapping algorithm, two-tree PFS was much faster than the Modified PFS.

Based on these results, the two-tree PFS algorithm is suitable for use in small, dense networks, and the Modified PFS algorithm is suitable for larger, sparse networks.

Inputs Required

The input is a linked graph G=<V, E>, where V are the nodes and E are the edges. Each node represents an entity which is a noun phrase. Each edge represents a co-occurrence relationship between the nodes it connects. The network is also termed a concept space.

Link Analysis Algorithm

For this solution, the criminal network must first be constructed. This network is a concept space, where nodes represent some specific concepts and links represent weighted co-occurrence relationships between concepts. This solution uses an approach to construct the concept space that is also used in COPLINK Detect. This criminal network is constructed from unstructured textual data.

The second step is to mine important indirect associations between entities. Two entities are indirectly associated if there is a path between them, and the path has length greater than one.

The solution employs shortest-path algorithms. This approach is better able to discover the more important associations compared to traditional association search approaches (e.g., breadth-first search). The Dijkstra algorithm is a typical shortest-path algorithm. Many other shortest-path algorithms are based on it, including the priority-first-search (PFS) algorithm and the two-tree Dijkstra algorithm. While these algorithms can be directly used to discover important associations between two entities in the criminal network, a transformation of the network representation must be performed.

Within a criminal network, each edge has a weight which represents the association strength (importance) of the link. A higher weight represents a stronger association. If two nodes are linked by a path of length greater than one, the association strength between the two nodes is the product of weights of the edges on path. To discover the most important paths between entities is to find the path with the largest weight.

Knowledge Engineering Cost

As with other unsupervised learning algorithms, the KEC is low. Nonetheless, this does not account for the KEC of the noun-phrase extraction phase, which has a high KEC.

Summary Table

Category: Academic and Commercial	
Hierarchy: not NE (noun-phrase)	**Source Scope**: Inter
Institution Name: University of Arizona **Institution URL**: http://ai.bpa.arizona.edu/research/coplink/index.htm	**Location**: Tucson, AZ, USA
Solution Name: Fighting Organized Crimes: Using Shortest-path Algorithms to Identify Associations in Criminal Networks	
Domain Scope: law enforcement	**Application Type**: LA
Knowledge Engineering Cost: low/high	**Financial Cost**: n/a
Input Requirements/Preparation Required: a linked graph G=<V, E> of noun-phrase nodes and co-occurrence relationships edges	
Link Analysis **Algorithm Name/Group**: shortest-path algorithms **Labeling**: n/a **Labeling Supervision**: n/a **Model Generation**: automatic **Model Generation Supervision**: unsupervised **Process Description**: After the criminal network has been constructed from unstructured data, the network is transformed into a new network which is suitable for use by shortest-path algorithms. Shortest-paths are used to determine the most important paths between two given nodes.	
Solution Output: important association paths between entities	
Application to Law Enforcement: extensive	
Is performance evaluation available? yes	**Solution/demo available**? no

Sources

Xu, J. and Chen, H. (2004). "Fighting Organized Crimes: Using Shortest-Path Algorithms to Identify Associations in Criminal Networks". *Decision Support Systems.* Volume 38, pages 473-487. Online. http://ai.bpa.arizona.edu/go/intranet/papers/Printed.pdf. Accessed January 4, 2006.

3.2.11 University of California, Los Angeles (UCLA) and IBM: Mining the Web for Relations

Solution Introduction and Domain Scope

This solution was developed by researchers at IBM Almaden Research Center in San Jose, California and the University of California at Los Angeles, California. It aims to mine patterns and relations from the web. This solution could be used in many domains (for example, discover the author of a particular book or discover all books written by a particular author), but in this description is limited to Web use. The solution is categorized as an information extraction and intrasource link analysis solution, since it not only extracts entities, but also mines patterns and relations within a source (i.e., a web page).

Output/Results

The output of this solution is relation pairs which may be predefined or new. For example, given authorship seeds such as an {author, title} pair, the output consists of new {author, title} pairs. The new relation pairs may also include relationships which were not predefined. For example, the {author, title} relationship might have been what the user input, but the solution may also discover new relationships such as {person, company}.

Application to Law Enforcement

Moderate. The solution is not targeted to the law enforcement domain, but could be adapted for use in it.

Evaluation

The experimental evaluation mined (acronym, expansion) relationships or AE pairs within the context of a topic-specific search engine for XML-related information (located at http://www.ibm.com/xml). The web pages used for acronym mining were gathered by a targeted crawler (Sundresan and Li, 2000) which crawled the web for information related to XML. The crawler was started off with the following AE-pairs provided as seeds:

(DCD, Document Content Description)
(CSS, Cascading Style Sheets)
(XML, eXtensible Markup Language)
Document Content Description (DCD)
Cascading Style Sheets (CSS)
eXtensible Markup Language (XML)

The crawler downloaded and analyzed 13,628 web pages, from which 2.694 unique AE-pairs and 948 unique patterns were discovered. The researchers also minded the same web pages with 10 a-priori acronym patterns without using their duality-based mining process. The results indicated that the mining by duality extracted twice as many AE-pairs as the extraction without duality.

No performance results were provided for the mining of new relationships.

Inputs Required

Web pages and a seed set are the required inputs to this solution. A seed is, in fact, a relation, such as {author, title}. The input seed determines the pattern to be discovered by the solution.

Link Analysis Algorithm

The paper discusses how a seed is used to discover other relation pairs. The solution performs supervised learning using duality. A duality problem is defined as in the following example:

1. *Begin with a small seed set of {author, title} pairs, which is provided manually.*
2. *Find all occurrences of these pairs on the web, which is automatically performed by the solution.*
3. *Identify patterns for the citation of the books from these occurrences, which is also done automatically*
4. *Search the web using these patterns to recognize additional {author, title} pairs.*
5. *Repeat the steps with the new {author, title} pairs.*

This process stops when it reaches a steady state, which means no additional pairs or patterns are found. In practice, complete coverage is unattainable, so a threshold is chosen according to coverage, time, etc. to define the steady state.

Patterns are grouped into one of two types: text patterns and structure patterns. A *text pattern* is based on the actual content of the text; this approach effectively removes the HTML from a webpage and considers web pages as plain text. *Structure patterns*, on the other hand, leverage the HTML tags. Both pattern types are used in pattern generation. As noted the solution is also able to discover new relations by treating the relations themselves as variables that can be mined and defined iteratively. Further details can be found in Sundresan and Li (2000).

Knowledge Engineering Cost

The algorithm is unsupervised aside from the need to provide a seed to bootstrap the process. In this sense the KEC is low. However, no evaluation of mining new patterns was provided, so the KEC is unclear for this aspect of the solution's capabilities.

Summary Table

Category: Academic and Commercial	
Hierarchy: not NE	**Source Scope**: Intra
Institution Name: University of California, Los Angeles; IBM Almaden Research Center **Institution URL**: http://cs.ucla.edu/ http://www.almaden.ibm.com/almaden/	**Location**: Los Angeles, California, USA; San Jose, California, USA
Solution Name: Mining the Web for Relations	
Domain Scope: general	**Application Type**: LA
Knowledge Engineering Cost: low/?	**Financial Cost**: n/a
Input Requirements/Preparation Required: Web pages and seed	
Link Analysis **Algorithm Name/Group**: duality algorithm **Labeling**: manual **Labeling Supervision**: n/a **Model Generation**: automatic **Model Generation Supervision**: supervised **Process Description**: The process beings with a small seed set of relation pairs and identifies occurrences of the pairs on the web. Patterns for the citation of those occurrences are discovered, and a search to recognize additional new pairs ensues. Processing repeats until a steady state is reached.	
Solution Output: Relation pairs	
Application to Law Enforcement: moderate	
Is performance evaluation available? yes	**Solution/demo available**? no

Sources

Sundaresan, Neel and Yi, Jeonghee Yi (2000). **"**Mining the Web for Relations." *Proceedings of the 9th International World Wide Web Conference on Computer Networks: the International Journal of Computer and Telecommunications Networking***.** Amsterdam, The Netherlands, pages: 699-711 Online. http://www9.org/w9cdrom/363/363.html Accessed January 21, 2006.

3.2.12 University of North Carolina, Chapel Hill and State University of New York, Buffalo: ApproMAP

Solution Introduction and Domain Scope

ApproMAP is a solution developed by researchers at the University of North Carolina at Chapel Hill and the State University of New York at Buffalo. This solution aims to mine the sequential patterns from sequence databases. ApproMAP has applicability in many domains, e.g., business analysis and Web mining, as long as the data in the domain is sequential. It is especially suitable for use mining large databases because of its efficient mining algorithm. ApproMAP is categorized as an intersource link analysis solution, since it aims to discover common sequence patterns among records/sequences. It discovers relationships in a group of records, instead of concentrating on the relations available in a single record.

Output/Results

The output of this solution is consensus sequence patterns. A *consensus pattern* is a long pattern which covers many short patterns and is shared by many sequences.

Application to Law Enforcement

Moderate. In criminal events, many sequential patterns exist, e.g., suspects also perform action B after performing action A. If police officers can recognize this pattern, then an occurrence of action A could aid in preventing an occurrence of action B. However, modifications would be needed to utilize this solution in a law enforcement environment, since the criminal event data would need to be stored in a sequential database format.

Evaluation

Evaluation was performed on a dataset of welfare services accumulated over a few years in the state of North Carolina. The researchers claim that the sequential patterns generated from this data prove to be general, useful, concise and understandable according to their manual analysis.

Software

n/a

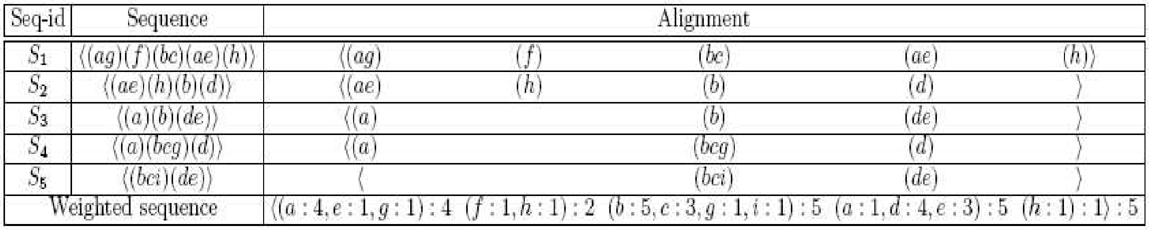
Inputs Required

The input to this solution is a sequence database (SDB). Let I = {i_1, ….., i_l} be a set of items. An itemset X= {i_{j1}, ….., i_{jk}} is a subset of I and a sequence S = <$X_{1, \ldots\ldots,}$ X_n> consists of a set of itemsets. A SDB is a multi-set of sequences. ApproxMAP is effective mining databases with long sequences and noise which pose difficulties for conventional sequential pattern mining methods.

Link Analysis Algorithm

The algorithm first divides the sequences into clusters. It then generates a consensus pattern for each cluster. A uniform kernel k-NN (Nearest Neighbor) clustering algorithm is used to perform sequence clustering. Initially, each sequence is considered a cluster, and then they are merged to produce a larger cluster. Sequences in the same cluster are similar with approximately the same patterns. The complexity for the clustering algorithm is O (kN_{seq}).

"Once sequences are clustered, the problem becomes how to summarize the general pattern in each cluster and discover the trend"(Kum et.al, 2003). First, the sequences in each cluster are aligned according to their density. A weighted sequence is then used to represent the aligned sequence, as in the table below (Kum et al., 2003).



Seq-id	Sequence	Alignment				
S_1	$\langle(ag)(f)(bc)(ae)(h)\rangle$	$\langle(ag)$	(f)	(bc)	(ae)	$(h)\rangle$
S_2	$\langle(ae)(h)(b)(d)\rangle$	$\langle(ae)$	(h)	(b)	(d)	$\rangle$
S_3	$\langle(a)(b)(de)\rangle$	$\langle(a)$		(b)	(de)	$\rangle$
S_4	$\langle(a)(bcg)(d)\rangle$	$\langle(a)$		(bcg)	(d)	$\rangle$
S_5	$\langle(bci)(de)\rangle$	$\langle$		(bci)	(de)	$\rangle$
Weighted sequence		$\langle(a:4,e:1,g:1):4$	$(f:1,h:1):2$	$(b:5,c:3,g:1,i:1):5$	$(a:1,d:4,e:3):5$	$(h:1):1\rangle:5$

Table 1: Sequences in a cluster and the complete alignment.

After discovering the weighted sequence, a consensus pattern is generated by identifying portions of the weighted sequence that are shared by most sequences in the cluster (the user can specify a strength threshold on which this is based). Given a threshold=30%, the consensus pattern for the weighted sequence in the table above is < (a) (bc) (de)>. It can be seen that this pattern is not contained in each sequence, but with the exception of S2, each sequence can fit this pattern with a single insertion. This indicates that this pattern is a general pattern underlying the data.

Knowledge Engineering Cost

The KEC for this solution is low since it does not require labeled data or manually crafted rules. The sequence pattern mining algorithm is also automatic.

Summary Table

Category: Academic	
Hierarchy: not NE	**Source Scope**: Inter
Institution Name: University of North Carolina, Chapel Hill; State University of New York, Buffalo **Institution URL**: http://www.unc.edu/ http://www.suny.edu/	**Location**: Chapel Hill, NC, USA; Buffalo, NY, USA
Solution Name: ApproxMAP: Approximate Mining of Consensus Sequential Patterns	
Domain Scope: general	**Application Type**: LA
Knowledge Engineering Cost: low	**Financial Cost**: n/a
Input Requirements/Preparation Required: Sequence database	
Link Analysis **Algorithm Name/Group**: K-NN clustering algorithm refined by multiple alignment and search methods **Labeling**: n/a **Labeling Supervision**: n/a **Model Generation**: automatic **Model Generation Supervision**: unsupervised **Process Description**: First, a K-NN clustering algorithm is used to cluster the input sequences. Then, sequences for each cluster are aligned. Finally, frequent sequence patterns are drawn from the aligned sequences for each cluster.	
Solution Output: frequent sequence patterns	
Application to Law Enforcement: moderate	
Is performance evaluation available? yes	**Solution/demo available**? no

Sources

Kum, H. C.; Pei, J.; Wang, W. and Duncan, D. (2003). "ApproxMAP: Approximate Mining of Consensus Sequential Patterns." *In International Conference on Data Mining, IEEE*. Online. http://www.siam.org/meetings/sdm03/proceedings/sdm03_36.pdf Accessed January 21, 2006.

3.2.13 University of Southern California: Unsupervised Link discovery in Multi-relational data via Rarity Analysis

Solution Introduction and Domain Scope

This solution was developed by researchers at the University of Southern California in Los Angeles, California. Its aim is to discover novel links within a link graph. This solution could be used in other domains, provided multi-relational data is available. An example of applicable domains includes social network analysis and bibliographic link analysis. This solution performs intersource link analysis, since it aims to discover new relations among sources, instead of concentrating on relationships in a single source.

Output/Results

The output of this solution is novel links defined as the most interesting path between two arbitrary nodes, the most important nodes for an arbitrary node, or the most interesting loop. For example, given a bibliographic database and a person A who is in the database, this solution discovers the most important people related to A in the database.

Application to Law Enforcement

Moderate. This solution could be adapted for use in law enforcement.

Evaluation

An evaluation was performed on the "High Energy Physics-Theory" bibliographic database, which is the experimental dataset for the KDD Cup 2003. The dataset includes five types of nodes and ten types of links. The nodes are paper IDs (29,016), author names (12,755), journal names (267), organization names (753) and publication times encoded as year/season (60). The links available are: authorship (author and his/her paper), date_published (paper and its publication date), affiliation (person and an organization he belongs to), published_in (paper and the journal it is published on), and citations (paper p and paper t, p cites t); inverse links for the above relationships are also recognized. Together, there are 42,871 nodes and 461,932 links in the graph.

First, the authors evaluated a significant node discovery method. After choosing the most significant person (A) in the dataset, their algorithm discovered the most important nodes related to A. The evaluation was manually completed by a human expert. These results provide some evidence that this approach is able to discover significant relationships without knowing the semantics of the entities or the links in the domain. Second, the authors evaluated their algorithm for discovery of the most interesting path. Their results were inconclusive.

Software

n/a

Inputs Required

The input to this solution is a linked graph G=<V, E>. V is the set of nodes in the graph, and E is the set of edges in the graph. Each node in the graph represents a named entity, which may be a person name, journal name, etc. Each edge in the graph represents some relationship between two nodes. For example a link between a person and a journal indicates an authorship relation. The linked graph can also be understood as a representation of multi-relational data.

Link Analysis Algorithm

The first challenge for link analysis in this solution is how to define "novel" or "interesting." Interest is driven largely by domain; a relation that is interesting for user A may not be for user B. For example people may think "A married B" more novel than "A wrote B a letter" since a 'marriage' event is less frequent than a 'write' event. This would not hold for a wedding registry database, however, where the reverse might be true. This solution defines the concept of *rarity* to define novelty. Three classes of NLD problems are considered:

1) *Novel path discovery*: the most interesting path between two arbitrary nodes,
2) *Significant node discovery*: the most important nodes for an arbitrary node, and
3) *Novel loop discovery*: a loops that was previously undiscoverable was found

Shortest-path algorithms may not be ideal to mine novel paths because the interestingness of a path is non-linearly related to the interestingness of individual links. In order to address these issues, *rarity* is defined as the reciprocal of the number of similar paths to a given path. Generally, an n-step path can be defined as n+1 nodes and n relations/edges (r_i) between them. The *type* of a path is defined as the ordered sequences of relations [r_0…..r_{n-1}] of that path. For example, the path "A writes a paper that cites a paper published at time t1" and the path "B writes a paper that cites a paper published at time t2" have the same type [write, cite, publish_at].

This solution developed four measurements for the similarity of two paths. Two paths are similar if they have the same:

1. *type*, *source node* and *target node;*
2. *type* and *emanate* from the same *source node;*
3. *type* and *target* at the same node;
4. *type.*

Users can select one of the four measurements to use, which provides flexibility. Novel path discovery is done by enumerating all paths between nodes X and Y and returning the one with the highest rarity value. For the second class of problems, two nodes are considered to be significantly connected with each other if they have many rare paths between them. Therefore, both the quantity and quality of paths are important. For the third class of problems (novel loop discovery), a loop is considered a special path where both the source node and the target node are the same. Given this, the same method used for novel paths can also be applied.

By having the user select a measurement for rarity and the problem he or she wants to solve, the solution proceeds automatically. No labeled data and no manually crafted rules are required by this solution.

Knowledge Engineering Cost

The KEC is low since the solution does not require labeled data or manually crafted rules. The novel relation discovery algorithm is automatic, as well.

Summary Table

Category: Academic	
Hierarchy: NE	**Source Scope**: Inter
Institution Name: University of Southern California	**Location**: Los Angeles, CA, USA

Institution URL: http://www.isi.edu/~sdlin/English/index_frame.htm	
Solution Name: Unsupervised Link Discovery in Multi-Relational Data via Rarity Analysis	
Domain Scope: general	**Application Type**: LA
Knowledge Engineering Cost: low	**Financial Cost**: n/a
Input Requirements/Preparation Required: linked graph/multi-relational data	
Link Analysis **Algorithm Name/Group**: a search algorithm with rarity as a measurement **Labeling**: n/a **Labeling Supervision**: n/a **Model Generation**: automatic **Model Generation Supervision**: unsupervised **Process Description**: First, four measurements for path similarity are defined. Then, the rarity (interestingness) of a path is calculated as the reciprocal of the number of similar paths. Finally, the most interesting path problem/most interesting node problem/most interesting loop problem can be solved based on the rarity of path.	
Solution Output: novel links	
Application to Law Enforcement: moderate	
Is performance evaluation available? yes	**Solution/demo available**? no

Sources

Lin, Shou-de and Chalupsky, Hans (2003). "Unsupervised Link Discovery in Multi-relational Data via Rarity Analysis." *In Proceedings of the Third IEEE International Conference on Data Mining (ICDM'03)*. 19-22 Nov, 2003. Pages:171-178. Online. http://www.isi.edu/~sdlin/publication/KDDExploration_V14.pdf. Accessed January 25, 2006.

3.2.14 University of Texas, Austin: TEXTRISE

Solution Introduction and Domain Scope

The TEXTRISE solution was developed by researchers at the University of Texas, Austin. It aims to discover rules that relate specific words and phrases. It has applicability in many domains such as book description corpora and patent documents. This solution is an intrasource link analysis solution, since for every rule generated the antecedents and consequents are from the same record.

Output/Results

The output is a set of rules which relate specific words and phrases classified as association rules. In a book description example, the rule could be "**title** dance ➔ **subject** romance, fiction [31.68%, 0.98]", which means if the title of a book contains the word "dance", then the subject of this book could be "romance" or "fiction." 31.68%' is the confidence, and 0.98 is the support.

Application to Law Enforcement

Moderate. Although not applied in law enforcement, this solution could be adapted to discover relationships between different fields of a police report.

Evaluation

Two domains were employed in the evaluation of TEXTRISE: book data from *Amazon.com* and patent data downloaded from *Getthepatent.com*. The book dataset is composed of six subsets, science fiction, literary fiction, mystery, romance, science, and children's books. 1,500 titles were randomly selected for each genre for a total size of 9,000. Six slots were extracted from each of the books: *titles*, *authors*, *subject terms*, *synopses*, *published reviews*, and *customer comments*.

Three thousand patent documents were collected from dynamically generated web pages returned by a keyword search for "artificial intelligence". Four slots of *titles*, *abstracts*, *claims*, and *descriptions* were extracted for each patent.

The experiments were performed on the 9,000 book descriptions using ten-fold cross validation. Learning curves for predicting the title slot are shown in the adjacent figure (Nahm and Mooney, 2001). The graph shows 95% confidence intervals for each point. The average similarities, precisions, and F-measure results were statistically evaluated using a one-tailed, paired t-test. For each training set size, two pairs of systems (TEXTRISE versus nearest-neighbor and nearest-neighbor versus nearest neighbor without information extraction) were compared to determine if their differences were statistically significant ($p < 0.05$).

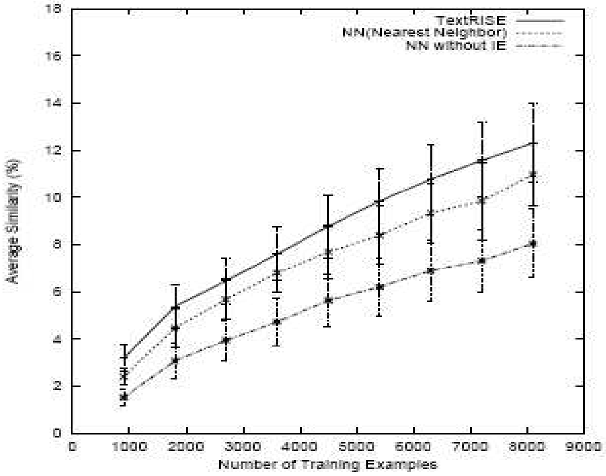


Figure 6: Average similarities for book data (`title`)

The results indicate that TEXTRISE performed best, while nearest-neighbor without IE produced the worst results. This demonstrates that TEXTRISE successfully summarized the input data in the form of prediction rules. The rule-compression rate of TEXTRISE is about 80%, which means the number of rules TEXTRISE produces is 80% of the number of examples originally stored in the initial rule base. The same experiments were conducted for other slots, and similar results were found. These results support the usefulness of soft-matching rules in prediction tasks for textual data.

Link Analysis Algorithm

Based on RISE, TEXTRISE is a hybrid algorithm which inherits characteristics from rule-based learning and instance-based learning. Each document is represented as a list of bags of words (BOW). During text processing, 524 commonly-occurring stop-words are eliminated, but stemming is not performed. A learned rule is represented as an antecedent that is a conjunction of BOWs for some subset of slots and a consequent that is a predicted BOW for another slot.

The following is an example presented in Nahm and Mooney (2001):

title nancy(1), drew(1)
synopses nancy(1)
subject children(2), fiction(2), mystery(3), detective(3), juvenile(1), espionage(1)
→
author keene(1), carolyn(1)

Rule Generation is performed using the following process. First, the rule set is initialized to the training set. After that, each rule in the rule set is iteratively generalized to cover more training instances. The iteration stops when there is no increase in *TextAccuracy* between two consecutive iterations. TextAccuracy is a criterion defined as the average cosine similarity of the predicted fillers for the examples in the instance set and the corresponding fillers predicted by a rule set. The output of TEXTRISE is an unordered set of soft matching rules. These rules are then ranked based on an interestingness metric in order to help human users focus on the most promising relationships. The *similarity-support* of rule A→C is defined as the sum of similarities between C and the consequents of the examples soft-matched by A in the dataset. *Similarity-confidence* of a rule A→C is computed by the similarity-support of that rule divided by the similarity-support of the antecedent A, where the similarity-support of A is the number of examples for which A is the closest rule in the rule set.

Knowledge Engineering Cost

Training data is needed to learn rules the training data must have feature names (slots in a template). As a result, although the KEC of the link analysis algorithm itself is low, overall the KEC is medium.

Summary Table

Category: Academic	
Hierarchy: NE	**Source Scope**: Intra
Institution Name: University of Texas **Institution URL**: http://www.utexas.edu/	**Location**: Austin, Texas, USA
Solution Name: TEXTRISE	
Domain Scope: : general	**Application Type**: LA
Knowledge Engineering Cost: medium	**Financial Cost**: n/a
Input Requirements/Preparation Required: textual data (could be online Web pages)	
Link Analysis **Algorithm Name/Group**: an algorithm extending the RISE algorithm **Labeling**: manual **Labeling Supervision**: n/a **Model Generation**: automatic **Model Generation Supervision**: supervised **Process Description**: First, the rule set is initialized to the training set, in which each document is represented as a bag of words. Then, each rule in the rule set is iteratively generalized to cover more training instances until there is no increase in TextAccuracy.	
Solution Output: The output is rules which relate specific words and phrases	
Application to Law Enforcement: moderate	
Is performance evaluation available? yes	**Solution/demo available**? no

Sources

Nahm, Un Yong and Mooney, Raymond J. (2001). "Mining Soft-Matching Rules from Textual Data." *Proceedings of the Seventeenth International Joint Conference on Artificial Intelligence(IJCAI-01).* Seattle,WA. Pages: 979-984. August, 2001. Online. http://www.cs.utexas.edu/users/ml/papers/discotex-ijcai-01.pdf. Accessed January 12, 2006.

3.2.15 University of Texas, Austin and University of Wisconsin, Madison: LD

Solution Introduction and Domain Scope

The Link Discovery (LD) solution is a solution developed by researchers at the University of Texas at Austin, TX and the University of Wisconsin at Madison, WI. It is a part of the Evidence Extraction and Link Discovery program and aims to discover links (patterns) from multi-relational data. Although LD was developed to detect and prevent terrorism, it is suitable for use in other domains, provided the domain employs multi-relational data. LD is categorized as an intersource link analysis solution, since it aims to identify patterns among records.

Output/Results

The output of this solution is relation patterns (i.e., association rules).

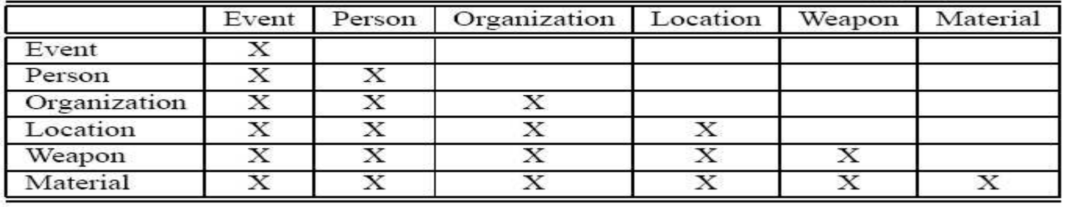
Application to Law Enforcement

Extensive. LD is suitable for use in law enforcement as a DARPA EELD-sponsored project.

Evaluation

The LD solution was tested in two domains, nuclear-smuggling and contract-killing using three different datasets. The contract-killing domain is divided into real world data (manually extracted from news sources) and artificial data (generated by a simulator). The nuclear-smuggling dataset consists of reports on Russian nuclear materials smuggling and contains 572 incidents (Mooney et al., 2002). Each incident includes entities, events and links between/among entities and events (the links are shown in the table below from Mooney et al. (2002).

Table 1. Links among Entities and Events in Nuclear-Smuggling Data



	Event	Person	Organization	Location	Weapon	Material
Event	X					
Person	X	X				
Organization	X	X	X			
Location	X	X	X	X		
Weapon	X	X	X	X	X	
Material	X	X	X	X	X	X

LD learns which events in an incident are related. For example, events A and D are related if they involve two people C and E and these people are connected to a third person through event F. This is, in fact, an association rule, so this solution is related to association rule mining. The data is organized in relational tables, such as in the above figure. The natural Contract-Killing Data is taken from Russian organized crime activities. Each incident in the chronology is based on a description from news articles (perhaps more than one source). For the task of identifying linked events in the nuclear-smuggling dataset, LD produced an average accuracy of 83% compared to a baseline of 78%. Ensemble learning increased the accuracy to 86%. The task of identifying motive in the contract-killing dataset is much more difficult, for which the accuracy of LD was 56% compared to a baseline of 50%. Ensemble learning increased this to 63%. Another approach to LD evaluated on the artificial contract-killing dataset resulted in precision and recall above 85%.

Inputs Required

The input to this solution is a linked graph G = <V, E>, where V is the set of nodes in the graph, and E is the set of edges in the graph. Each node in the graph represents a named entity, which could be a person, organization, object or action. Each edge in the graph represents a relationship between two nodes. As before, the linked graph is also a multi-relational data representation.

Link Analysis Algorithm

This solution is part of the EELD program which focuses on three sub-tasks: 1. Evidence Extraction (EE). 2. Link Discovery (LD). 3. Pattern Learning (PL).

This solution employs Inductive Logic Programming (ILP) to discovery links. ILP is the study of learning methods for data and rules that are represented in first-order predicate logic. For instance, suppose uncle(Tom, Bob) is an instance/example and uncle(x,y): brother(x,z), parent(z,y) is a rule. Given uncle and parent examples of this nature (both positive and negative), ILP is able to discover the uncle rule. The parent relation may also be given a priori, which means it is not necessary to learn/identify that portion of the rule from the data.

Three ILP systems were evaluated in this solution. One is named *Aleph*, which uses a simple greedy set covering algorithm that constructs a complete and consistent hypothesis one clause at a time. The second is named *Ensembles*, which uses bagging to balance unstable learning algorithms

(Mooney et al., 2002). The last is named *mFoil*, which also uses a greedy covering algorithm, but includes the use of constrained, general-to-specific search to learn individual rules.

Knowledge Engineering Cost

The KEC of this solution is high since its learning algorithm is a supervised covering algorithm, although it is automatic. Training data needs to be manually labeled and no effort is made to reduce the knowledge engineering cost.

Summary Table

Category: Academic	
Hierarchy: NE	**Source Scope**: Inter
Institution Name: University of Texas, Austin; University of Wisconsin **Institution URL**: http://www.cs.utexas.edu/users/ml/	**Location**: Austin, TX, USA; Madison, WI, USA
Solution Name: Relational Data Mining with Inductive Logic Programming for Link Discovery	
Domain Scope: general	**Application Type**: LA
Knowledge Engineering Cost: high	**Financial Cost**: n/a
Input Requirements/Preparation Required: linked graph/multi-relational data	
Link Analysis **Algorithm Name/Group**: Inductive Logic Programming **Labeling**: n/a **Labeling Supervision**: n/a **Model Generation**: automatic **Model Generation Supervision**: supervised **Process Description**: ILP is used to mine patterns/links from multi-relational data. Multiple ILP systems are tested, and their performance is compared.	
Solution Output: Links/Patterns	
Application to Law Enforcement: extensive	
Is performance evaluation available? yes	**Solution/demo available**? no

Sources

Mooney, R.; Melville, P.; Tang, L.; Shavlik, J.; Dutra, I.; Page, D. and Costa, V. Santos (2002). "Relational Data Mining with Inductive Logic Programming for Link Discovery." *Proceedings of the National Science Foundation Workshop on Next Generation Data Mining*. Baltimore, Maryland. Online: http://citeseer.ist.psu.edu/cache/papers/cs/32831/ftp:zSzzSzftp.cs.wisc.eduzSzmachine-learningzSzshavlik-groupzSzmooney.nsf02.pdf/mooney02relational.pdf Accessed January 10, 2006.

3.2.16 University of Washington: LitLinker

Solution Introduction and Domain Scope

LitLinker is a solution developed in Pratt et al. (2003) at the University of Washington, Seattle, Washington. It aims to discover novel links between biomedical terms. Since it requires domain specific knowledge in the medical field, it may not be as scalable as other approaches. The solution is categorized as intersource LA since it aims to discover second-order co-occurrence, a type of intersource relationship. LitLinker is a variant of Literature Based Discovery; additional detail on LBD is available in Ganiz et al. (2005).

Output/Results

The output is novel links between biomedical terms. In order to a link to be novel, it must not have previously been discovered by medical science. For example, Swanson (1988) discovered a link between *"migraines"* and *"magnesium."*

Application to Law Enforcement

Limited. LitLinker is specifically designed for use in the biomedical field and requires domain specific knowledge. Although the basic idea is suitable for use in law enforcement applications, it would require significant changes. For example, LitLinker uses the Unified Medical Language System (UMLS) to help identify concepts (biomedical terms).

Evaluation

To evaluate the solution's performance, the authors compared LitLinker to the results obtained by Swanson (1988). In Swanson (1988), 11 valid connections were discovered between migraine and magnesium. LitLinker was able to identify 118 linking concepts using "migraine" as the starting concept, but only 29 of those concepts linked migraine to magnesium. Of these 29, only five were one of the 11 valid connections discovered by Swanson; LitLinker missed six connections. However, four of them were not discovered because the terms were too infrequent in the migraine literature to satisfy their support threshold. An additional link was not found because the term was not identified as a biomedical term by MetaMap, and the final term never co-occurred with "magnesium" in a title. It is important to note that LitLinker did find an additional 24 useful linking terms. In addition, perhaps most importantly, magnesium was clearly identified as a target concept for migraine.

Inputs Required

The input is various biomedical sources, including literature (e.g., biomedical articles) and databases (e.g., MEDLINE).

Link Analysis Algorithm

LitLinker is a system which combines knowledge-based methodologies, Natural Language Processing (NLP), and a data-mining algorithm to mine the biomedical literature for novel links between biomedical terms. A high level view of the overall process for LitLinker is illustrated in the figure below. The process begins with a given *starting concept*, which is the concept the user wants to investigate. Next is a text-mining process whereby terms that are correlated to the starting concept are identified. These correlated terms are called *linking concepts*. For each of the linking concepts, the same text-mining process is used to identify a set of terms that are correlated with the linking concept. These final terms are called *target concepts*. Finally LitLinker groups and ranks the target

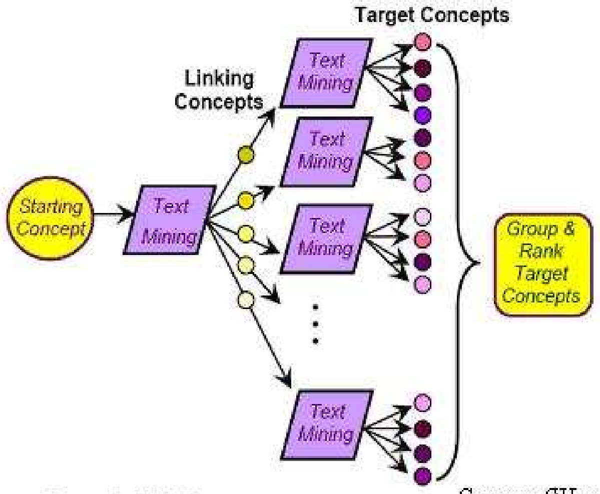


Figure – The Discovery Process in LitLinker. Source: (Wanda and Meliha, 2003) concepts by the number of linking concepts that connect the target to the starting concept.

The process is exemplified by following steps:

1. *Searching the Literature*: In the biomedical literature, for example MEDLINE, LitLinker searches for all citations that contain the start concept in the title. LitLinker chooses *"migraine"* as the starting concept,

2. *Identifying the Literature Concepts*: This step is part of a text mining process. Identifying a biomedical term is a very important part of the solution's performance, and LitLinker uses a knowledge-based natural language processing approach. The solution uses UMLS as the knowledge base, which is a large, publicly available knowledge base containing 875,000 biomedical concepts as well as 2.1 million concept names. It also uses the National Library of Medicine tool MetaMap to map from the textual data to the biomedical concepts in UMLS. LitLinker saves each concept identified by MetaMap, as well. The solution then merges all the synonymous terms, and assigns a preferred name to each such group.

3. *Pruning Concepts*: In step 2, many biomedical concepts are identified, but not all of them are useful. There are three issues: a. many terms are too general, e.g. *test*, *problem*; b. some terms are too closely linked to the starting concept, e.g. *headache*, *retinal migraine*; c. some terms do not have reasonable connections with the starting concept as judged by a human expert. Terms which are not useful are removed.

4. *Finding Correlations*: This is also a critical step, as it identifies the associated or correlated concepts. In this process the linking concepts and target concepts are identified. The Apriori association rule mining algorithm is used to identify the correlated concepts.

5. *Assembling Target Concepts*: First, LitLinker merges the target concepts from each of the linking concepts. Then, previously known connections are pruned, since only novel connections are desired. The process is completed by listing the target concepts together with the linking concepts.

Knowledge Engineering Cost

Although association rule mining is an unsupervised learning algorithm, the KEC is high for this solution because of its dependence on sophisticated knowledge bases such as MetaMap and UMLS.

Summary Table

Category: Academic	**Hierarchy**: NE	**Source Scope**: Inter
Institution Name: University of Washington **Institution URL**: http://litlinker.ischool.washington.edu/index.jsp	**Location**: Seattle, Washington, USA	
Solution Name: LitLinker		
Domain Scope: biomedical	**Application Type**: LA	
Knowledge Engineering Cost: high	**Financial Cost**: n/a	
Input Requirements/Preparation Required: Non-phrase meaningful terms		
Link Analysis **Algorithm Name/Group**: Apriori combined with domain specific knowledge and NLP **Labeling**: n/a		

Labeling Supervision: n/a **Model Generation**: automatic **Model Generation Supervision**: unsupervised **Process Description**: Given a starting concept, terms correlated with the concept are identified, called linking concepts. For each linking concept, text-mining is used to identify a set of terms which are correlated with the linking concept. These final terms are called target concepts. LitLinker groups and ranks the target concepts by the number of linking concepts that connect the target to the starting concept.	
Solution Output: novel links between biomedical terms	
Application to Law Enforcement: limited	
Is performance evaluation available? yes	**Solution/demo available**? yes

Sources

Ganiz, Murat C., Pottenger, William M. and Janneck, Christoper D. (2005) "Recent Advances in Literature Based Discovery." *Technical Report*. Online. http://www.cse.lehigh.edu/~billp/pubs/JASISTLBD.pdf Accessed January 10, 2006.

Pratt, W. and Yetisgen-Yildiz, M. (2003). "LitLinker: Capturing Connections across the Biomedical Literature." *Proceedings of the International Conference on Knowledge Capture (K-Cap'03)*. Florida. October, 2003. Online. http://www.ischool.washington.edu/wpratt/Publications/KCap-p032-pratt.pdf Accessed January 21, 2006.

3.3 Commercial Solutions

3.3.1 Autonomy Corporation plc

Company Introduction and Domain Scope

Autonomy Corporation plc is one of many leading companies identified in this survey. Headquartered in both Cambridge, UK and San Francisco, California, the company was founded in 1996 and has experienced "a meteoric rise" in becoming a leader in the field of handling and processing unstructured information from emails to video content (Autonomy). According to the company's website, Autonomy has been acknowledged by Delphi "as the fastest growing of the publicly traded companies in the [unstructured information] space" (Autonomy) and is also recognized a market leader by such organizations as Gartner Group and Forrester Research. With the acquisition of Verity (see below), the company employs more than 800 people.

The company's technology is based on research conducted at Cambridge University and the company has over 16,000 customers in both the public and private sectors, including such organizations as Ford, Reuters, Deutsche Bank, BAE Systems, Sun Microsystems, Coca Cola, BBC, Motorola, General Electric, the US Department of Defense, NASA, and the U.K. Houses of Parliament. It is also important to point out that Autonomy solutions have been "adopted as the organization standard at the US Department of Homeland Security" (Autonomy) and Autonomy serves as the primary organization responsible for coordinating the 22 agencies incorporated within this government entity (Franklin, 2002). Corporate partners include Lexis-Nexis, Moreover.com, NewsEdge, Oracle, OpenMarket, and Factiva. Autonomy's technology is also marketed under specialist brands including Aungate, etalk, Virage, and Cardiff (Autonomy).

Autonomy has also been the recipient of numerous awards including distinctions as a "company to watch" by both EContent and KMWorld in 2005 in addition to an "Effective IT Award 2005" from Information Age. Other awards and honors can be found on the company's website at http://www.autonomy.com/content/Autonomy/Awards.html.

A final important note is to point out the purchase of Verity by Autonomy in December 2005 for approximately $500 million. This merger created the largest search business at approximately $200 million annualized revenue. In comparison, the number two company is Google's $60 million enterprise search products business. Verity search products will be integrated into Autonomy's IDOL architecture (CNNMoney, 2006).

The company's solutions provide both information extraction and link analysis technologies. Unstructured, semi-structured, and structured data can all be used in the solution and require the use of information extraction to handle the earlier two data forms. Intrasource link analysis is performed as the solution attempts to understand the context of extracted concepts, and extracted information and sources are compared and linked through the use of categorization and search capabilities.

Output/Results

Autonomy Content Infrastructure™ (ACI™) is the technology specification and standardized format that the company uses to organize and structure the unstructured data sources. Other modules can communicate over this infrastructure or through the use of the Simple Object Access Protocol (SOAP). As much of the link analysis process of the system utilizes search-like technologies, the results can be returned as categorizations and summarizations as well as lists of sources or hyperlinks. Sources can be arranged in order of contextual relevance/distance.

Application to Law Enforcement

Extensive. In addition to coordinating the agencies of the U.S. Department of Homeland Security, Autonomy continues to work in the law enforcement arena. In 2002, the company teamed up with Unisys to develop HOLMES II, a system which coordinated the databases of 56 British police forces. The system "allows officers in different departments to search one another's crime databases and uses artificial intelligence technology to recognize the meaning of words from their context and make links between similar clues that may have been entered differently by different people" (Franklin, 2002).

Autonomy's technology could be a great asset to a police department by coordinating data and information from a wide variety of textual, audio, and video inputs with the company's search capabilities. According to a report specifically about Autonomy and its Homeland Security applications, "Technologies like Autonomy's increases the likelihood that accidents of discovery will take place, and therefore organizations, that deploy it in a sufficiently rich information environment will be better equipped to identify potentially hazardous situations before the occur" (Rasmus, 2002).

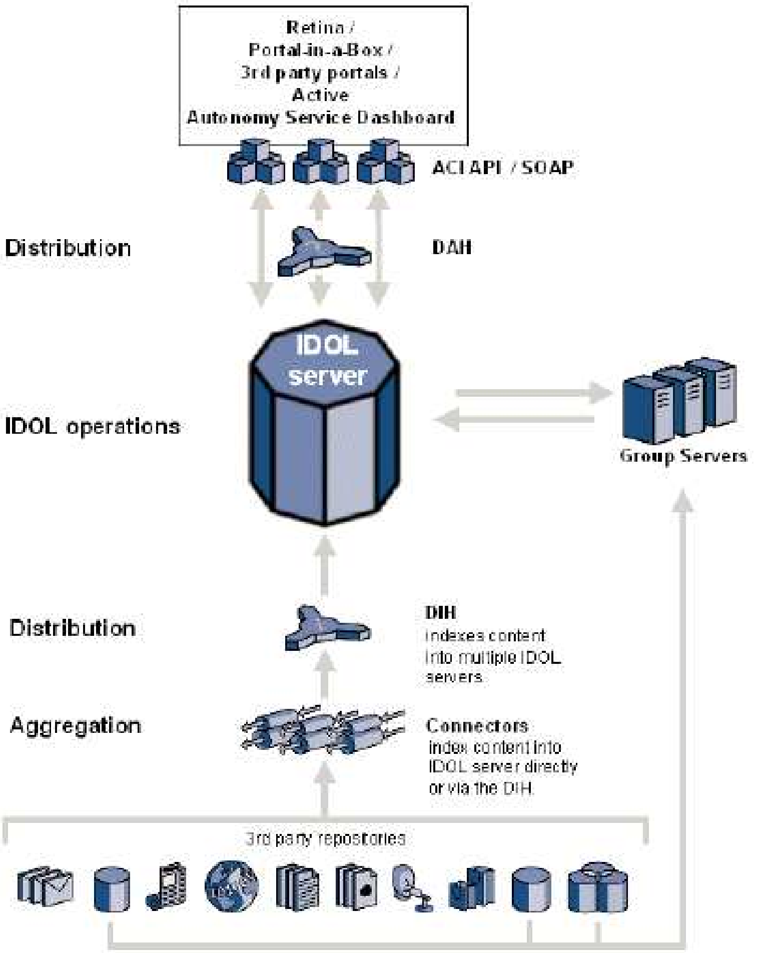
Evaluation

Some detail is provided in terms of the solution's retrieval speeds, but little of substance in terms of metrics such as precision and recall. One gigabyte of corporate data in HTML, Lotus Notes, MS Office, and PDF formats can be retrieved in 20 milliseconds while 3 gigabytes of real-time news on a fully distributed system can be processed in 40 milliseconds (Autonomy). In terms of categorization, approximately 4 million documents can be categorized in 24 CPU hours, working out to one document every 25 milliseconds. Other details speed and performance results can be found at (http://www.autonomy.com/content/Technology/Technology_Benefits/SpeedAndPerformance.html) and in (Autonomy, 2003b).

Financial

No details of the financial cost of Autonomy's solutions were available.

Software



Autonomy's core product offering and "flagship product" is the Intelligent Data Operation Layer™ (IDOL) Server, which serves as the heart of the company's software infrastructure. Details of the IDOL server are discussed in the Algorithm section. However, it is important to note that, with the acquisition of Verity, IDOL Federator and IDOL K2 versions are also available which make use of and integrate Verity's technology.

Different product offerings such as *Aungate* (real-time enterprise governance), *Cardiff* (business process management (BPM)), *etalk* (customer service applications), *Virage* (rich media management), and *softsound* (audio processing and speech search) are available to allow for easier information access and coordination (Autonomy). *Autonomy Retrieval* "offers a wide range of retrieval methods, from simple legacy keyword search to highly sophisticated conceptual querying" (Autonomy). *Portal in a Box* extends the power of the solution to be accessed via an Information Portal Infrastructure while *IDOL Enterprise Desktop Search* customizes a user's search experience by constructing a search history and profile. *Autonomy Answer* provides responses to common customer questions as an automated CRM system. *Collaboration and Expertise Networks* (*CEN*) keeps track of user queries and searches to profile users and "foster a collaborative network" (Autonomy). *Retina* extends Autonomy's retrieval methods as a web interface application. A high level diagram is presented in the figure above (Autonomy, 2005a; Autonomy).

GUIs allow for easy access and modification and custom-built applications in C, Java, COM, and COM+ are available to communicate with the ACI API over HTTP. Security is also included within the solutions, "allowing fully mapped and unmapped models with document level and user level entitlement, as well as secure communication between servers" (Autonomy). The Intellectual Asset Protection Service (IAS) also provides security on many levels, including asset and group membership scalability, at least 128-bit encryption, as well as authentication and entitlement.

No demos or evaluation copies of the solution are available.

Inputs Required

A wide variety of data inputs can be used within Autonomy solutions, include both textual data (emails, documents, spreadsheets, ASCII text, emails, repositories, etc.) and video data. Over 300 different repositories and over 250 data formats (http://www.autonomy.com/content/Technology/Technology_Benefits/SupportedFormats.htm) are supported. The IDOL server can integrate "unstructured, semi-structured, and structured information from multiple repositories through an understanding of the content" (Autonomy, 2005a).

The company uses the phrase *piece of content* to refer to various inputs into the system. Sentences, paragraphs, pages of text, email bodies, records of human-readable information, and derived contextual information of an audio or speech extract are all examples of pieces of content (Autonomy, 2003a). In terms of link analysis, searches and queries can process these pieces of content and produce results.

Link Analysis Algorithm

According to the company's website, "Autonomy's strength lies in a unique combination of technologies that employ advanced pattern-matching techniques (non-linear adaptive digital signal processing), utilizing Bayesian Inference and Claude Shannon's Principles of Information" (Autonomy, 2003a). (See (Autonomy, 2003a) for descriptions of these two approaches). The technology "identifies the patterns that naturally occur in text, based on the usage and frequency of words or terms that correspond to specific ideas concepts" (Autonomy, 2003a). "Based on the preponderance of one pattern over another in a piece of unstructured information, Autonomy enables computers to understand that there is X% probability that a document in question is about a specific subject. In this way, Autonomy is able to extract a document's digital essence, encode the unique 'signature' of the concepts, then enable a host of operations to be performed on the text, automatically" (Autonomy, 2003a).

Over 65 languages are supported, and the IDOL Server engine can be trained on any language's pattern, such as German, Spanish, Portuguese, Arabic, Italian, French, Japanese, Chinese, Norwegian, etc. Auto-detection of languages is also provided with the solution. This Dynamic Reasoning Engine™ "is based on advanced pattern-matching technology (non-linear adaptive digital signal processing) that exploits high-performance probabilistic modeling techniques to extract a document's digital essence and determine the characteristics that give the text meaning. As this technology is based on probabilistic modeling, it does not use any form of language dependent parsing or dictionaries" (Autonomy).

The solution does not use keyword searching or Boolean query, but matches concepts by taking into consideration the context of the data. It does use collaborative filtering or social agents, but automatically generates user profiles "by extracting key ideas from the actual information the user reads" (Autonomy, 2003a). Parsing and NLP are also avoided as Autonomy uses a pattern-matching technology which "uses predictable statistical word patterns to represent concepts and functions independently of any given language" (Autonomy, 2003a). Manual tagging has also been replaced by an additional "layer of intelligence to the management of XML" (Autonomy, 2003a).

Autonomy's technology extracts "concepts" and utilizes metadata and XML tags to enhance and automate this process through the use of the EDUCE module. "Autonomy IDOLServer™'s conceptual understanding enables it to automatically insert XML tags and links into documents, based on the concepts contained in the information. This eliminates all manual cost…IDOL server [also] enables XML applications to understand conceptual information, independent of variations in tagging schemas or the variety of applications in use. This means, for example, that legacy data from disparate sources, tagged using different schemas, can be automatically reconciled and operated upon" (Autonomy, 2003c). Weighting (positive and negative) as well as stop words and stemming are also used to enhance linking.

Searches are performed using a wide range of technologies, from conceptual queries (example, keyword, Soundex algorithm, etc.) to Boolean, parametric, and field searches. The solution also performs taxonomy categorization. Automatic learning and clustering on approximately 10 to 20 document sources (the seed) can be used to form the taxonomies or they can be manually created. Keywords, relationships and weighting, and Bayesian Inference can all be utilized. Dynamic linking of sources returned from searches is also an important component of the solution.

However, little detail into the exact processes in terms of information extraction and link analysis were provided in the literature. While it is apparent that the approach is highly mathematical and probabilistic in nature, few details are available.

Knowledge Engineering Cost

Autonomy solutions allow a wide range of user input. Taxonomies, categories, and key words can all be either manually crafted or identified or can also be automatically learned. This allows great flexibility within the system. However, as the approach provides a high degree of automation through the use of mathematical approaches such as Bayesian Inference and Shannon's Information Theory, the solution has been classified as having a medium KEC.

Summary Table

Category: Commercial	
Company Name: Autonomy Corporation plc **Company URL**: http://www.autonomy.com/	**Location**: Cambridge, UK and San Francisco, CA, USA
Solution Name: Intelligent Data Operation Layer™ (IDOL) Server; IDOL Federator; IDOL K2; Aungate; Cardiff; etalk; Virage; softsound; Autonomy Retrieval; Portal in a Box; IDOL Enterprise Desktop Search; Autonomy Answer; Collaboration and Expertise Networks (CEN)	
Domain Scope: general	**Application Type**: IE and LA
Knowledge Engineering Cost: medium	**Financial Cost**: unknown
Input Requirements/Preparation Required: Unstructured, semi-structured, and unstructured data source including both textual data (emails, documents, spreadsheets, ASCII text, emails, repositories, etc.) and video data can be used. Over 300 different repositories and over 250 data formats are supported.	
Link Analysis **Algorithm Name/Group**: concept extraction through the use of Shannon's Information Theory (entropy) and Bayesian Inference (probabilistic models) **Labeling**: n/a **Labeling Supervision**: n/a **Model Generation**: hybrid **Model Generation Supervision**: supervised **Process Description**: Categorizations and taxonomies are automatically generated but can also be created or modified through the use of a human expert. Patterns can automatically be identified, while manual searches can also be performed.	
Solution Output: Search results can be returned as categorizations and summarizations as well as lists of sources or hyperlinks.	
Application to Law Enforcement: extensive	
Is performance evaluation available? yes	**Solution/demo available**? no

Sources

Autonomy. Available: http://www.autonomy.com/. Accessed January 16, 2006.

Autonomy (2003a). *Autonomy Technology White Paper*. 2003. Online.http://www.autonomy.com/downloads/Marketing/Autonomy%20White%20Papers/Autonomy%20Technology%20WP%2020040105.pdf. Accessed January 16, 2006.

Autonomy (2003b). *Performance & Scalability White Paper*. August, 2003. Online. http://www.autonomy.com/downloads/Marketing/Autonomy%20White%20Papers/Performance%20and%20Scalability%20WP%2020050811.pdf. Accessed January 16, 2006.

Autonomy (2003c). *XML White Paper*. Online. http://www.autonomy.com/downloads/Marketing/Autonomy%20White%20Papers/Autonomy%20XML%20WP%2020031003.pdf. Accessed October 10, 2005.

Autonomy (2005a). *Autonomy IDOL Server™ 5 Technical Brief.* Online. http://www.autonomy.com/downloads/Technical%20Briefs/Servers/TB%20IDOL%20server%205%200305.pdf. Accessed October 10, 2005.

Autonomy (2005b) *Document Management Technical Brief.* Online. http://www.autonomy.com/downloads/Technical%20Briefs/Servers/TB%20Document%20Management%20Server%200205.pdf. Accessed October 10, 2005.

CNNMoney (2006). "Google Gets More Personal." *CNNMoney.com.* January 12, 2006. Online. http://money.cnn.com/2006/01/12/technology/google_enterprise.reut/index.htm. Accessed January 22, 2006.

Franklin, Daniel (2002). "Data Miners: New Software Connects Key Bits of Data that Once Eluded Teams of Researchers." *Time: Online Edition*. December 23, 2002. Online. http://ai.bpa.arizona.edu/go/intranet/papers/GlobalBusiness.pdf. Accessed June 2, 2005.

3.3.2 AeroText™ (Lockheed Martin)

Company Introduction and Domain Scope

AeroText™ is a solution developed at the Integrated Systems and Solutions division of Lockheed Martin Corporation, a leading U.S. Defense contractor. Originally developed for the U.S. intelligence community (Department of Defense), the solution has become one of the leading solutions available and is often integrated into other solutions. For instance, Entrieva, one of the company's partners, has integrated AeroText's technology into their product line, and their SemioTagger solution has been used by the U.S. Army (KMWorld, 2003). Evidenced Based Research, Inc. has also heavily utilized AeroText capabilities within their own "information fusion" solution development (Nobel, a) (Nobel, b) as it is considered a "state of the art text extractor" (Nobel, a) for single-sentence analysis within its system. NetMap Analytics also incorporates the technology into their solution, which allows analysts to "visualize vast volumes of data and apply unique algorithms to reveal the hidden patterns and relationships within" (Hill, 2005).

AeroText solutions provide both information extraction and link analysis capabilities.

Output/Results

AeroText output is normalized and stored within the solution's cache as templates (see Algorithm). However, the information can be output in a variety of ways using the Run Time Integration Toolkit (RIT) to integrate the output into existing systems through the use of RIT modules. Wrappers for XML and the DARPA Agent Markup Language (DAML) and also provided.

Application to Law Enforcement

Extensive. As already mentioned, the solution was originally developed for intelligence applications and has been deployed in the field, as well. However, the solution is also flexible enough to be utilized in other domains. For instance, the solution was presented to the National Institute of Health's Biomedical Computing Interest Group (BCIG) in April of 2002 and demonstrated excellent applicability to the biomedical domain. "AeroText is data-independent, which means it does not rely on or have a bias towards a particular domain, document type, document source, or natural language" (Haser and Childs, 2002). Sample target applications include automatic database generation, document routing, browsing, summarization, enhanced full text search, and targeted document search in addition to link analysis.

The solution's multilingual utility is also a strength. The technology is also flexible enough to be able to support format standards, such as DAML (Kogut and Holmes), which aid in law enforcement activities.

Evaluation

No specific evaluation results were found. However, the company claims to identify and extract information "with an accuracy that matches or exceeds a human's ability to do so" (Mordoff, 2005). It also can process at "high speed (100 – 1,000 Mbytes/hr)" and leaves a small hardware footprint (AeroText).

Financial

While no specific information was found, (Noble, b) reports that "[d]eveloping rules for a new domain can be labor intensive, sometimes requiring more than a month of effort from experienced AeroText™ users."

Software

AeroText, which released its most recent solution version (4.0) in April, 2005, exists as a set of various components that are used to carry out integration and data mining tasks. The *Integrated Development Environment (IDE)* is, perhaps, the most important component as it provides the rule development, modification, and coordination capabilities – "a complete environment to build, test, and analyze linguistic knowledge bases" (Kogut and Holmes). This graphical interface includes not only object oriented editors and rules wizards, but is also allows visual tools for analyzing extracted data, debugging linguistic data, and analyzing performance (AeroText). As a result, customized logic domains are available.

The *Instance Based Run-Time Engine* actually carries out the extraction on input documents by applying a Knowledge Base (see below). According to the company, "an Instance is defined as the creation of a single Document Object in the AeroText Application Program Interface (API)." The engine is available in Java, C, or COM API's and has wrappers for XML and DAML. The *Run Time Integration Toolkit (RIT)* helps to deploy AeroText by minimizing the need for integration code and provides for the integration of AeroText output into existing systems through the use of RIT modules. The *Corpus Analyzer* clusters documents based on entity and conceptual similarities between documents. The *Answer Key Editor* creates an information store for scoring by assigning "an Answer Key that corresponds to a specific collection of documents" (AeroText). This Key helps to determine the accuracy of the extraction process.

Much of the solution's technology is provided within the company's *Knowledge Bases* (KBs). English serves as the key core KB and provides linguistic-driven rules which contain over 50 entity types uses to extract text. KBs are also available for the Arabic, Chinese (simplified and traditional), Spanish, and Bahasa Indonesia (including Melagu) languages. A KB Compiler is used to convert "linguistic data files into an efficient run-time knowledge base" (Kogut and Holmes).

AeroText's solution components are available separately or as one of two product bundles. The Standard bundle includes the IDE, Instance-based Run-Time Engine, Core English Knowledge Base, and the Customization Tool. The Professional bundle includes the Standard components as well as the Corpus Analyzer and the Answer Key Editor). (AeroText).

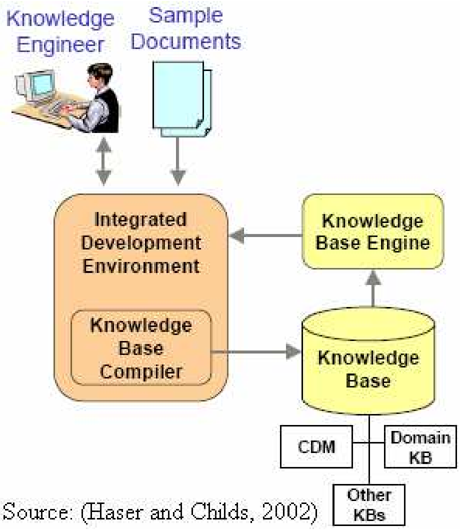
A small demo of AeroText's capabilities on a few sample documents (and compared with METIS and NetOwl) is provided on the web at http://im-dev-1.industrialmedium.com/xp/IC__working/AeroText/SMLA/040505_SMLA_IRAN.xml.

Inputs Required

AeroText can handle any textual input, as the Instance Based Run-Time Engine supports both ASCII and Unicode text.

Link Analysis Algorithm

AeroText's main focus is on "information extraction," which includes both named entity extraction and intrasource link analysis. "AeroText information extraction technology is designed for natural language text" (AeroText, 2003). The company has organized its capabilities into several groupings. Specifically for information extraction, *entities* (persons, organizations, places, etc.), *key phrases* (time expressions, money amounts, etc.), and *grammatical phrases* (verb phrases, etc.) can all be extracted. In terms of link analysis, the solution provides *entity coreference* (resolution of multiple mentions of the same entity), *entity associations* (identify relationships), *event extraction* (who, what, when, where), *topic categorization* (subject matter determinations), temporal resolution (resolution of time expressions, etc.), and *location resolution* (identification of a particular place which can be tied to GIS). Additionally, the company's BlockFinder™ can be used to understand textual tables. (Haser and Childs, 2002).



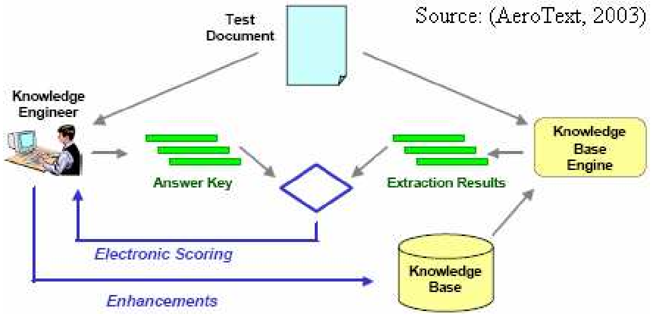
Source: (Haser and Childs, 2002)

The solution gains its flexibility and broad range of applicability from the fact that the system is based on the use of manually crafted rules. These rules are used to perform both entity extraction and intrasource link analysis. While different modules developed will be extensively subject-matter specific, the solution can be easily modified to handle the requirements of a different domain. Therefore, in order to use the solution, "an AeroText specialist must generate a set of extraction rules. These rules describe for AeroText how to identify and structure the information to be extracted. In effect, they create fairly abstract templates that describe all the different ways a concept can be expressed in the target language" (Noble, b). These rules not only extract the information from the text, but also specify how the information should be structured within event records (Nobel, a).

(Haser and Childs) explains that the fundamental components of the solution include features, elements, templates, packages, rulebases, and caches. (These terms are explained using the following example: "Feb. 28, 2002 AAA Corporation will acquire Tampa-based ZZZ Inc. within 60 days.")

- A *feature* is "a list of terms that represents a common idea based on meaning or grammar," e.g., 'inc.' and 'corp.' are business designations {*CorpDesignator*}.
- An *element* is "a set of regular expressions that allow binding of information to matched text"; for instance, "FEB" and "February" both refer to the second month (month = "2").
- A *template* is "a frame with slots used to hold extracted text and sometimes related information." A time template, for example, would include a "text" field as well as "StartDate" and "EndDate" fields.
- A *package* is "a set of rules, similar to elements, but with associated actions that fill template slots with extracted information." The example above would have Time, Organization, and Location templates into which extracted information could be organized.

- A *rulebase* is "a collection of packages that are activated at the appropriate time during a processing sequence." This example would have the Time and Organization templates feed into an Acquisition template.
- A *cache* provides "a virtual bin for storing extracted information." An *entities cache* stores times, organizations, and other such information, while an *events cache* can store event information, such as acquisitions.



Source: (AeroText, 2003)

A high-level overview of how the solution is set up is provided by the adjacent figure. Given a test document, a knowledge engineer produces the answer key of supposed output while the knowledge base engine uses pre-packaged and user-developed rules to extract the entities and relationships from the text. These two outputs are compared and scored. If changes need to be made, the knowledge engineer creates additional rules or makes other enhancements to the knowledge base (which in turn updates the knowledge base engine).

A more detailed analysis of the solution is provided in (Wu and Pottenger, 2005b). According to this source, the first step of AeroText's process is to *segment* the text; this is done using sentence boundaries (e.g., ".", "!", "?"). The solution then *tokenizes* the text into words, numbers, and punctuation. The third step requires the use of "either a pre-defined or custom designed database schema to represent various patterns as Features, Elements and Support Patterns to guide AeroText in rule generation" (Wu and Pottenger, 2005b). Each training dataset instance requires a domain expert to identify the sub string that exactly matches an expression of a given attribute; an example of this could be a *Date*. The system would then display each token's feature (e.g., "year", "month", "day") to the domain expert knowledge engineer to have them select the portion of each feature that is to be used in the pattern (rule). AeroText applies the pattern to all instances in the training set to remove the instances that are covered by the pattern. It then selects another instance to find another pattern. The process stops when all instances have been covered by generated rules (Wu and Pottenger, 2005b).

Each rule generated is assigned a weight to express the knowledge engineer's confidence in the rule (a larger weight indicates a higher confidence). AeroText "also includes support for negative patterns that are used to remove useless instances from other patterns' results to purify the results. Negative weights are assigned to negative patterns" (Wu and Pottenger, 2005b).

Within the system, slots in a template are used to express patterns. "A *slot* is akin to an attribute in a database schema, or can be an entire pattern. A technique using dynamic binding is employed to decide the content of a given slot in a template. This method allows complex patterns to be identified and expressed. Furthermore, it can be used to find relationships between patterns. For example, if a *Date* is related to a person's *Name*, it often is a person's birthday" (Wu and Pottenger, 2005b). Wu and Pottenger (2005b) conclude that "AeroText is a manual covering algorithm," requiring the tagging of exact features.

Knowledge Engineering Cost

As the system requires manual rule generation and tagging of exact features, the system involves a high knowledge engineering cost. While the solution produces excellent results and works in many domains, it relies to a large extent on human interaction to generate the rule sets and iterate through the process until all of the training instances are covered.

Summary Table

Category: Commercial	
Company Name: Lockheed Marin Corporation **Company URL**: http://www.aerotext.com/	**Location**: Gaithersburg, Maryland, USA
Solution Name: AeroText™	
Domain Scope: general	**Application Type**: IE and LA
Knowledge Engineering Cost: high	**Financial Cost**: requires approximately 1 month for each domain rule set to be developed
Input Requirements/Preparation Required: Any textual input (ASCII, Unicode)	
Link Analysis **Algorithm Name/Group**: covering **Labeling**: n/a **Labeling Supervision**: n/a, **Model Generation**: manual **Model Generation Supervision**: supervised **Process Description**: Test text is segmented and tokenized before a user is directed through a covering approach to ensure all instances are covered by a manually crafted rule	
Solution Output: Normalized and stored within the solution's cache as templates. Can be output in any format via RIT, as well as XML and DAML	
Application to Law Enforcement: extensive	
Is performance evaluation available? no	**Solution/demo available**? yes

Sources

AeroText. Available: http://www.aerotext.com/. Accessed August 5, 2005.

AeroText (2003). *AeroText Products: Extracting Intelligence from Text*. May, 2003. Online. http://www.lockheedmartin.com/data/assets/3497.pdf. Accessed January 9, 2006.

Entrieva (2003). "Retrieving Information." *KMWorld*. Vol. 12, Issue 8. September, 2003. Online. http://www.kmworld.com/Articles/ReadArticle.aspx?ArticleID=8558. Accessed January 9, 2006.

Haser, Tom and Childs, Lois (2002). "Drug Discovery through Information Extraction Technology." Presentation at *NIH BCIG*. April 18, 2002. Online. http://www.altum.com/bcig/events/seminars/2002_04.pdf and http://www.altum.com/bcig/events/seminars/2002_04.htm. Accessed January 9, 2006.

Hill, Ryan (2005). *Lockheed Martin Signs NetMap Analytics as Authorized Distributor of AeroText™ Information Extraction Software*. August 3, 2005. Online. http://www.netmapanalytics.com/press/AeroText.pdf. Accessed January 9, 2006.\

KMWorld. *KMWorld Buyers Guide: Lockheed Martin Corporation*. Online. http://www.kmworld.com/buyersGuide/ReadCompany.aspx?CategoryID=77&CompanyID=17. Accessed January 9, 2006.

Kogut, Paul and Holmes, William. *AeroDAML: Applying Information Extraction to Generate DAML Annotations from Web Pages*. Online. http://semannot2001.aifb.uni-karlsruhe.de/positionpapers/AeroDAML3.pdf. Accessed January 9, 2006.

Mordoff, Keith (2004). *Lockheed Martin's NEW AeroText™ Version 4.0 Helps Users Tackle Data Overload, Pinpoint Critical Information.* April 14, 2005. Online. http://www.lockheedmartin.com/data/assets/10586.pdf. Accessed August 9, 2005.

Noble, David (a). *Fusion of Open Source Information.* Online. http://www.ebrinc.com/files/Noble_Fusion.pdf. Accessed January 9, 2006.

Noble, David (b). *Structuring Open Source Information to Support Intelligence Analysis.* Online. http://www.ebrinc.com/files/Noble_Structuring.pdf. Accessed January 9, 2006.

Roberts, Gregory (2003). *AeroText™ Products: Executive Summary Information.* Online. http://www.lockheedmartin.com/data/assets/3504.pdf. Accessed January 9, 2006.

Taylor, Sarah M. (2004). "Information Extraction Tools: Deciphering Human Language." *IT Professional.* Vol. 06, no. 6, pages: 28-34. November/December, 2004. Online. http://ieeexplore.ieee.org/iel5/6294/30282/01390870.pdf?tp=&arnumber=1390870&isnumber=30282. Accessed January 9, 2006.

Wu, Tianhao and Pottenger, William M. (2005b). "A Very Brief Comparison of AeroText with Lehigh University's Approach to Information Extraction." Private communication from authors received on August 15, 2005.

3.3.3 Attensity Corporation

Company Introduction and Domain Scope

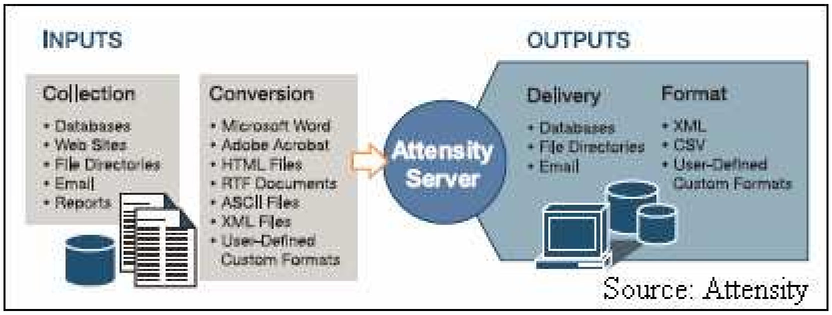
Palo Alto, California-based Attensity has developed powerful information extraction technology that is "the culmination of over a decade of research in computation linguistics at the University of Utah" (Attensity). They already have five patents, with twenty additional patents pending. The company's primary client is the government (60% of Attensity business (Shachtman, 2005)), but the company's client base also includes many leading companies such as Whirlpool, John Deere, Honeywell, and General Motors. "Attensity also maintains ongoing relationships with leading systems integrators and consultants including Booz Allen Hamilton, EDS and SAIC, and business and technology partnerships with such vendors as IBM, Ascential and Teradata" (Attensity).

The company is also the recipient of many awards. Attensity's solution was recognized as a KMWorld Trend Setting Product in both 2004 and 2005, a finalist in Red Herring's list of 100 Private North American companies, and one of Fortune magazine's "Breakout Companies" of 2005. It also received a "Most Likely to Succeed" award at Silicon Valley Venture Capital Event (HBD Network).

Attensity's solution performs both information extraction and link analysis tasks.

Output/Results

Attensity solutions convert unstructured text into structured tables or databases. The entities (which answer such questions as *who*, *what*, *when*, *where*, and *why*) are then "output in XML and in a structured relational data format that is fused with existing structured data" (Attensity). Using additional tools (including Attensity Discover and Attensity Analytics (see Software)), the data can then be analyzed.



Source: Attensity

Application to Law Enforcement

Extensive. As already mentioned, the majority of the company's business is with the government, including such organizations as the Federal Bureau of Investigation, the National Security Agency, and the Defense Intelligence Agency (Shachtman, 2005). Given this, and the fact that the Central Intelligence Agency's venture capital arm, In-Q-Tel, served as the company's original investor, it seems apparent that the software has extensive use in the law enforcement community.

The solution is designed to be as simple as possible for the user and requires no data mining expertise in order to use the system. The solution employs the company's Directed Learning approach (see Algorithm). While the primary focus has been on providing extraction and link analysis tasks for the English language, the company has been expanding its capabilities to handle any European, Latin American and select Asian languages.

Evaluation

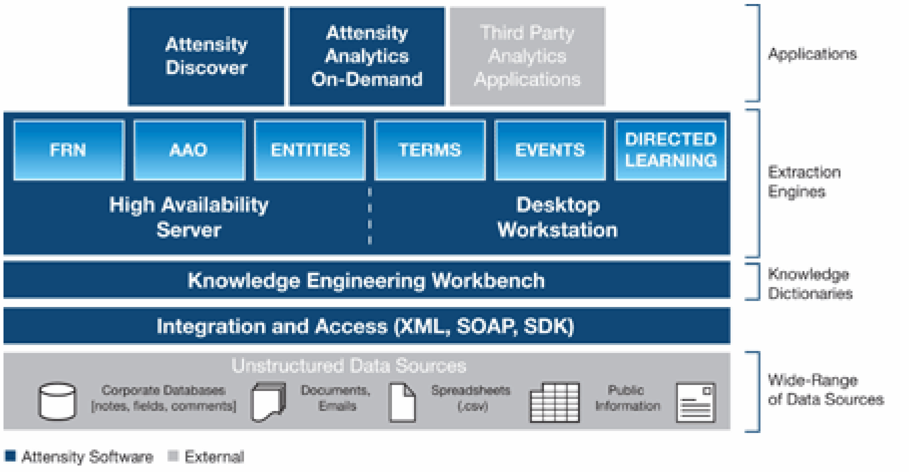
In a company white paper (Attensity 2005b), Attensity claimed that they "made a fundamental breakthrough in converting unstructured text into structured tables with 95% or better accuracy (precision + recall)." Using a 1GHz Intel CPU, the solution can process high raw text at a rate of 5MB/minute. The core Natural Language Processing engine's performance has a linear relationship with the amount of input text (Attensity, 2005b). Additionally, Mena (2004) states that Attensity's technology "can process nearly 100 single-spaced pages per second." Shachtman (2005) mentions that the novel *Moby Dick* took only nine and a half seconds to analyze.

Financial

Little information was available as to the costs of obtaining Attensity's solution. However, Shachtman (2005) states that Whirlpool is spending $250,000 annually for "Attensity's expertise."

Software

Attensity carries several products that are available in its Text Analytics Suite, such as Attensity Server, Attensity Workstation, Attensity Discover, Attensity Analytics (On-Demand), and engineering and integration tools. Attensity incorporates both information extraction and link analysis capabilities by automatically extracting valuable data from free-form text and combining it with structured data to quickly generate datasets. The company's Knowledge Libraries provide pre-packaged in-depth industry and business-based expertise to the user.



Source: Attensity

Attensity's Extraction Engines provide the key information extraction capabilities of the solution, as it converts unstructured textual data into structured information. Attensity Server brings these engines together to allow the linear scaling of the text extraction.

Attensity Discover and Attensity Analytics provide the key link analysis tools. These tools allow query and exploration of the extracted, structured data to identify relationships and drill down into details. By incorporating the newly extracted data with the existing data, Attensity is able to provide a more complete analysis. Additionally, they allow browser-based visualization capabilities.

Attensity Workstation is the company's desktop analysis tool which allows the user to easily and rapidly perform ad hoc desktop analysis of textual data. Attensity Software Development Kit allows users to create unstructured data applications to extract custom information. Finally, the company's Application Suite carries out several application functions that are of specific concern to businesses, such as Warranty, Customer Care, Risk Management, Government Intelligence, Government Law Enforcement, and Government Logistics.

The solution is available for purchase through both direct sales channels and system integrators. No demos or trial versions are available.

Inputs Required

Attensity Server (and therefore the Extraction Engines) support many formats, including XML, text, pdf, rtf, csv, and other custom data types. Attensity Analytics can integrate data from multiple sources, including the output from the company's information extraction tasks.

Link Analysis Algorithm

Mena (2004) states that "the company's text extraction technology relies on structural linguistic principles and can convert all types of unstructured content." Fortune Magazine claims that Attensity's technology should be thought of as "lightning-fast computerized sentence diagramming: Each document is distilled into a spreadsheet of who did what when, where, and to whom, making patterns, repetitions, and relationships between words easy to spot" (Hira, 2005).

Attensity's own literature provides a more detailed explanation. The company has divided natural language processing and text extraction into four complexity stages: *stemming and morphological processing*, *named entity recognition and part-of-speech tagging*, *parsing*, and *thematic role recognition and discourse processing*. The first stage provides textual transformation. Stemming "is the process of stripping prefixes and suffixes from words in an attempt to handle lexical variation and reduce the size of information retrieval indexes," while morphological analysis takes stemming one step further and requires more sophisticated processing, a dictionary, and a set of morphological transformation rules. The next stage analyzes what the textual terms refer to by labeling the entities with types and parts of speech. The third stage, parsing or syntactic analysis, works to understand the relationships that exist between the words and phrases within a sentence.

The fourth stage involves an even more complex level of analysis, and is the stage at which Attensity's technology resides. Thematic role understanding "takes the structural representation that parsing identifies and transforms it to a standardized representation of who did what to whom, when, where, and how." Discourse processing is "the ability to recognize the relationships between sentences and their constituents" (Attensity, 2005b). For instance, anaphora resolution, aka coreference resolution, falls into this category, as it involves for example identifying the object to which "he," "she," or "it" refers to in the text.

Attensity has broken down their approach to extract events and attributes into a three-step process. First, event triggers (i.e., verbs or normalized verb forms) are identified. Next, features and named entities are extracted from the text, mapping variations back to a single entity. In the final step, an "analysis of the roles of words and entities, and their relationship to each other and to event triggers" is carried out.

Using a proprietary *Directed Learning*™ approach, the user is guided through an active approach to label the data. After providing a seed (a manual process), users "tell the system what items of interest they want to extract and then direct the system through a series of sample texts. Based on the examples, the system begins performing extractions and the user interactively tells it when it is right and when it is wrong" (Atttensity, 2005a). Attensity's solutions also utilizes sentence diagramming as part of its part of speech learning to better analyze the text and handle unknown words, misspellings, and ungrammatical constructions. These extractors can then be reused. As a final step, the unstructured textual data is then converted into structured tables or databases.

According to the company's website, "[Our] technology allows users to extract and analyze facts like who, what, where, when and why and then allows users to drill down to understand people, places and events and how they are related. It then creates output in XML and in a structured relational data format that is fused with existing structured data so that it can be analyzed using Attensity's applications, Attensity Discover and Attensity Analytics, or by using business intelligence applications already installed in the enterprise" (Attensity).

Knowledge Engineering Cost

While involving a significant amount of human interaction, the solution performs its model generation with an active and supervised approach that utilizes a seed and then guides the user through the rule-generation process – as opposed to having the user develop the rule solely on their own. The solution also automates many link analysis tasks. Therefore, we would consider Attensity's approach to have a medium knowledge engineering cost.

Summary Table

Category: Commercial	
Company Name: Attensity Corporation **Company URL**: http://www.attensity.com/www/	**Location**: Palo Alto, California, USA
Solution Name: Text Analytics Suite (Attensity Discover, Attensity Analytics (On-Demand), Attensity Server, Attensity Workstation, Attensity Integration, Attensity Knowledge Engineering)	
Domain Scope: general	**Application Type**: IE and LA
Knowledge Engineering Cost: medium	**Financial Cost**: approximately 40 man/days to create rules for a new domain
Input Requirements/Preparation Required: Data is labeled using proprietary *Directed Learning*™ approach, a walk-through with sample texts.	
Link Analysis **Algorithm Name/Group**: proprietary **Labeling**: n/a **Labeling Supervision**: n/a **Model Generation**: manual **Model Generation Supervision**: n/a **Process Description:** Links are formed from the extraction process (facts), as well as through the use of search and visual analysis.	
Solution Output: The entities are converted into structured tables or databases.	
Application to Law Enforcement: extensive.	
Is performance evaluation available? yes	**Solution/demo available**? no

Sources

Attensity. Available: http://www.attensity.com/ Accessed January 16, 2006.

Attensity (2005a). *Attensity Text Analytics Suite: Overview*. Online. http://www.attensity.com/www/pdf/AttenWorkstation_4_13_05.pdf. Accessed January 26, 2006.

Attensity (2005b). *Natural Language Processing and Text Extraction*, October 2005. Obtained via email correspondence. Received October 21, 2005.

Hira, Nadira A. (2005). "25 Breakout Companies 2005." *Fortune*. May 16, 2005. Online. http://www.fortune.com/fortune/breakout/snapshot/0,23871,21,00.html. Accessed August 11, 2005.

Mena, Jesus (2004). "Homeland Security as Catalyst." *Intelligent Enterprise*. July 1, 2004. Online. http://www.intelligententerprise.com/showArticle.jhtml?articleID=22102265. Accessed June 2, 2005.

Shachtman, Noah (2005). "With Terror in Mind, a Formulaic Way to Parse Sentences." *New York Times*. New York, NY. March 3, 2005. Online. http://www.nytimes.com/2005/03/03/technology/circuits/03next.html?ex=1135141200&en=b7e59924788a2cdb&ei=5070. Accessed August 11, 2005.

3.3.4 ClearForest

Company Introduction and Domain Scope

ClearForest is another leading company identified in our survey effort. Located in Massachusetts and Israel, this company was founded in 1998 by Dr. Ronen Feldman (Bar-Ilan University, Israel) and has emerged as one of the industry leaders and offers an entire solution suite to its customers. Partnering with many leading companies (such as IBM, EDS, Endeca, LAS, Verity), ClearForest serves major clients such as Johnson and Johnson, J.D. Power and Associates, NASDAQ, and Dow Chemical Company in addition to many government/defense clients, such as Boeing, Sandia National Laboratories, the US Air Force, and Israeli security agencies, among many others. ClearForest's solution performs both information extraction and link analysis tasks.

Output/Results

ClearForest solutions tag the entities which can then be stored in XML, CSV, or standard DB format. While keeping the original document in its original form, data is learned, extracted, and transformed into a structured form that can then be used to effectively searched and queried.

Application to Law Enforcement

Extensive. ClearForest works heavily with governments and the defense industry. The extracted information is highly structured and can be readily used to aid in law enforcement applications. Specifically, ClearForest's factual tags (see Algorithm section for more information) allow valuable clues, relationships, facts, and events to be structured for analysis and comparison.

Evaluation

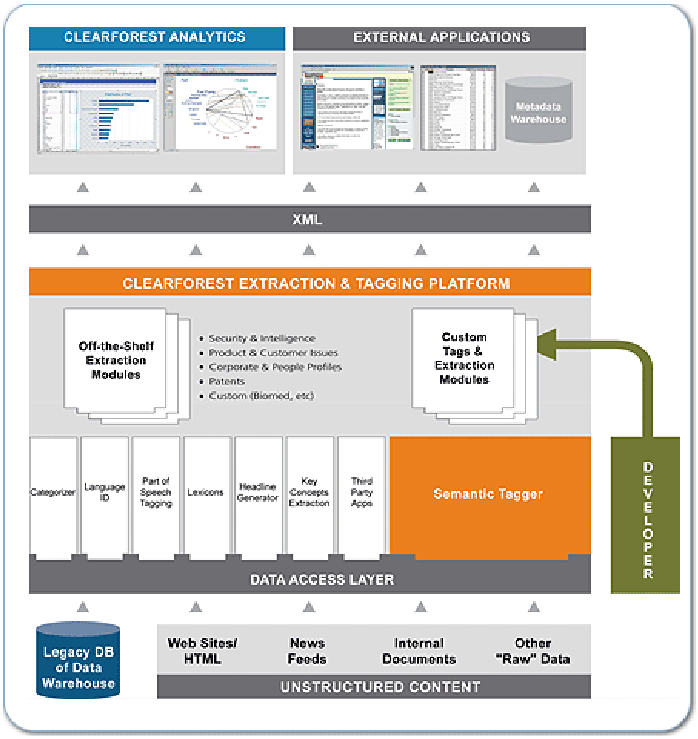
As mentioned in (Wu and Pottenger, 2005a), ClearForest participated in the 2002 KDD Challenge Cup competition in biomedical domain (Regev, et. al, 2002). During this competition, F-measure scores of 78% and 67% were achieved in the Document Curation task and the Gene Product task, respectively.

Financial

According to Bock (2002), ClearForest reported that its average deal size was approximately $450,000 "and depends on such criteria as the size of the installation, range of ClearForest capabilities implemented, the number of people accessing the application, the number of licensed CPUs, and other business considerations." These costs include both installation and set up of the system (including the creation of manual rule set if the available Extraction Modules are not sufficient). Systems can be implemented in as little as three weeks.

Software

ClearForest produces a suite of tools: Text Analytics Platform, ClearForest Tags, ClearForest Extraction Modules, ClearForest Analytics, and ClearForest Developer all perform various stages of the information generation process. The solution is available for purchase only. No demos or trial versions are available.



Inputs Required

In terms of data, there are no requirements for input other than textual data. ClearForest solutions work with ASCII text, pdf, HTML, XML, and Microsoft Office, etc. and can be configured to work with any format.

Link Analysis Algorithm

ClearForest's technology is based on an information extraction algorithm, which recognizes several distinct types of entities which are recognized and then *tagged* from the original document source. These *tags* are first organized into Document level tags, which organize the documents into categories, and *inner document* tags, which deal with the information contained within the document. For the purposes of this survey, we are more concerned with the inner document tags. This category is further organized into *descriptive* tags, *factual* tags, and *role* tags. *Descriptive tags* and *role* tags provide information extraction capabilities. (Pottenger et al., 2006a) *Factual tags* "provide information on facts and events mentioned within the text" (ClearForest, a), an intra-document link analysis technology. Factual tags are logically based and capture relationship information that exists among elements found within the document. For instance, in a corporate environment, ClearForest solutions could help to identify a subsidiary or business relationship that is static (such as a division or subsidiary) or a partnership or merger; the former are termed *facts* by ClearForest, and the latter are termed *events*. ClearForest factual tags learn both facts and events.

ClearForest uses different approaches in how they process these different tags: statistical tagging (dependent upon token frequency, etc.), semantic tagging (based on the "meaning of the underlying text" (ClearForest White Paper)), and structural tagging (based on typographic and positional characteristics). Factual tags are obtained using semantic tagging. These tags are used to extract and understand facts and events. These tags require no labeling and the rule generation approach is Manual/Active.

More specifically, the system is based on the use of Tagging and Extraction Modules, which contain the core rules that are necessary to tag and extract the entities. Through the use of DIAL (Declarative Information Analysis Language), which uses manually crafted rules to extract the entities from the data, these rule sets can be generated. However, the company goes to great lengths to make the extraction process as simple and powerful as possible. Several domain-specific extraction modules are already available off the shelf. User-defined extraction modules may be developed using ClearStudio (non-code) or ClearLab (DIAL-code creation). ClearStudio allows a non-technical individual with industry experience to walk through the creation of these modules, while ClearLab enables more technical users to write their own DIAL code.

Knowledge Engineering Cost

Given the mix of manual and automatic approaches to rule creation, ClearForest's approach has a medium to high KEC.

Summary Table

Category: Commercial	
Hierarchy: NE	**Source Scope**: Intra
Company Name: ClearForest **Company URL**: http://www.clearforest.com/	**Location**: Waltham, Massachusetts, USA
Solution Name: CF Text Analytics Platform (infrastructure platform); CF Tags, CF Extraction Modules, CF Analytics, CF Developer, ClearStudio, ClearLabs (applications)	
Domain Scope: general (dependent upon Extraction Module used)	**Application Type**: IE and LA
Knowledge Engineering Cost: high	**Financial Cost**: average deal size $450,000 (2002)
Input Requirements/Preparation Required: The primary makeup of the system is the Extraction Module, which is based on industry or domain scope. Once the Extraction module has been created, the solution is ready to begin extraction.	
Link Analysis **Algorithm Name/Group**: proprietary **Labeling**: n/a **Labeling Supervision**: n/a **Model Generation**: manual **Model Generation Supervision**: n/a **Process Description**: The system uses DIAL (Declarative Information Analysis Language), which uses manually developed rules to extract the entities from the data.	
Solution Output: Tagged entities within the context of the original source.	
Application to Law Enforcement: extensive	
Is performance evaluation available? no	**Solution/demo available?** no

Sources

Bock, Geoffrey E. "Meta Tagging and Text Analysis from ClearForest: Identifying and Organizing Unstructured Content for Dynamic Delivery through Digital Networks." *Patricia Seybold Group.* 2002. Online. http://www.instinct-soft.com/WhatsNew/Research.asp Accessed August 8, 2005.

ClearForest. Available: http://www.clearforest.com/ Accessed December 17, 2005.

ClearForest (a). *White Paper - Tagging Textual Data: Why? What? How?* Available: http://www.clearforest.com/WhatsNew/Research.asp Accessed August 8, 2005.

Feldman, Ronen; Aumann, Yonatan; Libetzon, Yair; Ankori, Kfir; Schler, Jonathan and Rosenfeld, Benjamin. (2001). "A Domain Independent Environment for Creating Information Extraction Modules." *CIKM 2001*. Pages: 586-588. Online. http://www.cs.biu.ac.il/~aumann/papers/IEInvironment.pdf. Accessed November 1, 2005.

Regev, Y., Finkelstein-Landau, M., and Feldman R. (2002). "Rule-based Extraction of Experimental Evidence in the 15 Biomedical Domain – the KDD Cup 2002 (Task 1)." *SIGKDD Exploration. Newsl.* 4, 2 Dec, 2002, pages: 90-92. Online. http://delivery.acm.org/10.1145/780000/772874/p90-regev.pdf?key1=772874&key2=8532584311&coll=GUIDE&dl=GUIDE&CFID=63236164&CFTOKEN=96493586. Accessed December 17, 2005.

Wu, T. and Pottenger, W. M. (2005). "A Semi-Supervised Active Learning Algorithm for Information Extraction from Textual Data." *Journal of the American Society for Information Science and Technology*. JASIST, Volume 56, Number 3, Pages: 258-271. Online. http://www.cse.lehigh.edu/~billp/pubs/JASISTArticle.pdf. Accessed September 1, 2005.

3.3.5 Delphes Technologies International

Company Introduction and Domain Scope

Founded in 1998, Montreal, Canada-based Delphes Technologies International "offers an intelligent knowledge service that integrates advanced information structure expertise with an innovative technology for organizing know-how" (Delphes). The company's management team consists of several linguists and IT experts from institutions such as MIT, McGill, and the University of California at Berkeley.

The company's solution is utilized in many different industries, such as government, insurance, finance, legal, manufacturing, healthcare, technology, education, professional services, and tourism by approximately 200 customers. Clients of the company include L'Oreal, CSQ, CAIJ, Bell Canada, Bombardier Inc., Quebec's finance department, and Desjardins Financial Security.

The solution provides both information extraction and link analysis capabilities. It is an intrasource link analysis solution because it performs extensive contextual analysis on the extracted entities and intersource link analysis solution because it allows users to search among sources to determine results and allows the results to be saved.

Output/Results

Ranges of characters, structured sets of morphemes, words, phrases, and text are all extracted with Delphes' technology. Indexing activity reports can be generated to display the knowledge currently available in the solution in CSV format. Summaries generated by the Summarizer can be saved in PDF, HTML, or RDF format.

Application to Law Enforcement

Moderate. While the linguistics-based processing technologies prevent a novel approach to information extraction and information retrieval, the application to the law enforcement domain is fairly limited. While a more efficient and effective means of entering queries and returning search results would benefit anyone, its application domain is not specifically targeted towards the law enforcement community. However, the solution has seen widespread use in the legal and government domains.

Evaluation

No evaluation information was available. Dr. Anna Marie Di Scuillo serves as the company's Vice President of Linguistic Strategy and has written papers on which the company's technology is based. Although these papers were not readily available, they could provide an evaluation of the methodology used by Delphes' solutions.

Financial

Pricing for licensing the company's DioWeb solution was available online (https://www.delphesintl.com/ecommerce/) and was determined based on the number of languages desired (English, French, Spanish) and the number of documents supported (up to 1,500 or up to 5,000) in the offering. A price of $10,867.50 was given for a solution with all three languages and up to 5,000 documents, which included an annual maintenance fee of $1,417.50 and provided one hour of technical support. Any solution with one language and support up to 1,000 documents cost only $1,840.00 (with $240.00 for the annual maintenance fee). Breakdowns are provided in the following table:

Number of Languages	Number of Documents	Price	Maintenance	Total
1	1,000	$1,600.00	$240.00	$1,840.00
2	1,000	$1,920.00	$288.00	$2,208
3	1,000	$2,240.00	$336.00	$2,576.00
1	5,000	$6,750.50	$1,012.00	$7,762.50
2	5,000	$8,100.00	$1,215.00	$9,315.00
3	5,000	$9,450.00	$1,417.50	$10,867.50

Software

Delphes solution is offered as one of three product offerings: *DioSMW*, *DioMillenium Series*, and the *DioWeb Series*. DioWeb works primarily in the extranet and internet domain while the intranet portal domain is covered by DioMillenium. DioSMW is the company's flagship offering and provides the most comprehensive technology the company has to offer. However, the three solutions are fairly similar with modifications in the number of technical features included.

Delphes' technology is divided into a set of modules where are each responsible for a different task. The *extraction module* allows for search results to be returned, while the *indexing module* makes sure that the sources are indexed within the system for more rapid retrieval. Indexing is based on a wide-variety of input aside from the main textual body of the source and uses such input as annotations, metatags, notes, bookmarks, and titles. Parameters are also stored to keep the document's size, date, type, and language (Delphes, 2004a).

The *statistics module* generates search statistics and analyzes the solution. This includes analyzing the actual needs of users (by compiling the search queries and analyzing user search sessions) as well as understanding the information available (by generating indexing statistics). The *Information Manager* is an optional module which expands the capabilities and allows for dynamic management of information assets and maintains the history of search and summaries generated by users (Delphes, 2004b). The *security module* allows for user login and the hiding of information from those without the necessary permissions.

The *linguistics module* is available in two versions: *enterprise* and *standard*. The standard version provides "advanced analytical capabilities to distinguish a query's related concepts" (Delphes, 2004a) and identifies morphological concepts. Grammatical and spelling errors are identified and spelling suggestions are provided. Language detection is also provided for English, French, Spanish, and German. The enterprise version includes the standard components but provides even more advanced capabilities, including normalization and syntactic information to recognize context. "Semantic search capabilities distinguish heads, names, subjects, verbs, and complements in order to extract the query's meaning and related concepts" (Delphes, 2004a). Named entities (proper nouns, compound words, acronyms, symbols, and abbreviations), locutions, and homographs are also identified and extracted. The solution performs these tasks through the use of specialized dictionaries (Delphes, 2004a).

The *customization module* allows search results (color, number of results, etc.) to be customized according to the user's tastes and preferences. Several optional modules are also available; these include a *summarizer* (which automatically generates and displays summarized information for specific subjects), *multi-server search*, *advanced search statistics* (CRM-type statistics), *advanced security* (document section-level), and *specialized dictionaries*.

The solution also includes security protections such as fail over clustering, load balancing, and Web security integration (Basic, NTLM, DPA, Cookie/Script, HTML/Form). Group, category, and file management levels are also available. The software also coincides with industry standards such as .NET, COM, and API as well as supporting C++, C, Perl, VB, C#, VB.NET, ASP, and ASP.NET.

A limited online demo of the solution comparing Delphes to Google (on Cisco's English website) and Microsoft Index Server (on CSST's French site) is available at http://209.41.142.136/demo1/home.asp.

Inputs Required

Information can be extracted from over 250 different file formats, including MS Excel, MS PowerPoint, PDF, HTML, MS Exchange, and Lotus Notes files.

Link Analysis Algorithm

Delphes' technology utilizes Diogene, a linguistics-based information extraction and retrieval technology, and Dynamic Natural Language Processing, which allows for contextual indexing, searching, and information retrieval (EMC^2, 2006).

Delphes' integrated information system works to determine the words' contextual purpose by performing "configurational analysis on all phrases in texts to determine the logical function of words" (Delphes, 2003). This process involves four main steps. The *localization* step parses the text to locate each of the individual words. Next, the *morphology* step performs a morphological analysis of words by comparing to dictionaries. Delphes' dictionaries specify not only the stem of the word among its lexical variants, but also identify the potential grammatical categories of the words. The *syntax* step disambiguates the grammatical category of the word by analyzing the context, such as part of speech, function, and meaning. This information is then formed into *constituents*. A constituent is "a structural unit of one or more linguistic elements (as morphemes, words, or phrases) that can occur as a component of a larger construction" (Delphes, 2003). By forming the text into constituents, users can maximize the usability and relevance of search results. (Delphes, 2003).

Indexing is also an important part of Delphes' technology. Both the data and metadata are *indexed* to allow for efficient retrieval of the extracted information. Indexing can occur on a regular schedule and be limited by document size, date, type, language, section, and URL. These capabilities are enabled through the use of the Universal Axiomatic Engine (UNAX™). This engine "is based on advanced principles and parameter scanning technology that models high-performance human properties" (Delphes, 2004b). Four main functions are performed by the UNAX™.

Configuration detection: This stage detects information by identifying abstract structured entities which are referred to as "configurations." These entities "range from structured sets of characters to structured sets of morphemes, to structured sets of words, to structured sets of phrases, to structured sets of texts" (Delphes, 2004b) while common practices only target single characters, morphemes, etc. "The UNAX™ mimics a fundamental feature of the human cognitive system: the ability to process information supported by natural language in terms of the manipulation of abstract configurations and categories" (Delphes, 2003).

Relation Preservation (Transformational Facilities): Using a limited set of transformations, relations between the query and the equivalent expressions are maintained. For instance, "the portrait of Mona Lisa by Da Vinci" will also include "Mona Lisa's portrait by Da Vinci," "Da Vinci's portrait of Mona Lisa," and "the portrait of Mona Lisa that Da Vinci painted" while not including incorrect expressions such as "the portrait of Da Vinci" or "Da Vinci's portrait by Mona Lisa" (Delphes, 2003).

Concept Expansion: As the configurations can contain multiple meanings, the solution seeks to determine the true concept behind the configuration by identifying the entity, property, or event which refers to the configuration in question. "UNAX™ derives conceptual expansion from the relation between a root and a derivational affix, as well as from the relation between a root and an inflectional affix" (Delphes, 2003). Compound words are also analyzed using a lexical map to determine their contextual function. "The identification of conceptual relations supported by nominal expression is central in the system, as the referent (object of a search) is supported mainly by nominal expressions in natural languages" (Delphes, 2003).

Evolved Text Search: The search capabilities of this axiomatic system include noun phrase (NP) detection and shallow parsing.

The solution also claims to incorporate the principles and parameters of universal grammar. These principles "determine both the morphological shape and the syntactic makeup of expressions in natural language" (Delphes, 2003). This allows the system to be used with other languages.

Knowledge Engineering Cost

Given the use of a dictionary, it would appear that the approach has a high KEC. The application of the solution to multiple languages and its adherence to "universal grammar" indicates that the solution is more flexible that a dictionary could provide. Therefore, it has been concluded that the KEC of this solution is medium.

Summary Table

Category: Commercial	
Hierarchy: NE	**Source Scope**: Intra and Inter
Company Name: Delphes Technologies International **Company URL**: http://www.delphes.com/	**Location**: Montreal, Canada
Solution Name: DioSMW DioMillenium Series DioWeb Series	
Domain Scope: general	**Application Type**: <LA, IE and LA>
Knowledge Engineering Cost: medium	**Financial Cost**: $1,840 - $10,867.50 (DioWeb)
Input Requirements/Preparation Required: Information can be extracted from over 250 file formats.	
Link Analysis **Algorithm Name/Group**: linguistics-based configuration and constituent analysis **Labeling**: n/a **Labeling Supervision**: n/a **Model Generation**: hybrid **Model Generation Supervision**: supervised **Process Description**: Queries entered into the solution have the information extracted from them prior to being applied against the indexed information.	
Solution Output: Ranges of characters, structured sets of morphemes, words, phrases, and text are all extracted with Delphes' technology. Reports and summaries are generated in CSV, PDF, HTML, or RDF format.	
Application to Law Enforcement: limited	
Is performance evaluation available? no	**Solution/demo available**? yes

Sources

Delphes. *Delphes Technologies International*. Available: http://www.delphes.com/. Accessed January 23, 2006.

Delphes (2003). *White Paper: Integrated Information System*. Online. http://www.delphes.com /pdf/en/white_paper.pdf. Accessed January 23, 2006.

Delphes (2004a). *Extranet and Internet Solutions.* Online. http://www.delphes.com/pdf/en/ internet.pdf. Accessed January 23, 2006.

Delphes (2004b). *Intranet Portal Solutions.* Online. http://www.delphes.com/pdf/en/intranet.pdf. Accessed January 23, 2006.

Delphes (2005). *Data Sheet – Intelligence Knowledge Service.* Online. http://www.delphes.com/pdf /en/datasheet.pdf. Accessed January 23, 2006.

Di Sciullo, Anna Maria and Fong, Sandiway (2001). "Efficient Parsing for Word Structure". *In the Proceedings of the Sixth Natural Language Processing Pacific Rim Symposium.* November 27-30, 2001. Online. http://www.afnlp.org/nlprs2001/pdf/0034-03.pdf. Accessed January 23, 2006.

EMC^2 (2006). *EMC^2 Partners: Delphes Technology International*. Online. http://www.emc. com/partnersalliances/partner_pages/delphes.jsp. Accessed January 23, 2006.

3.3.6 Eidetica

Company Introduction and Domain Scope

The Amsterdam, Netherlands-based Eidetica provides text mining software. The company was founded in 1998 by scientists of CWI, the Dutch national research Centre for Mathematics and Computer Science, and merged with Filter Control Technologies in 2002. While Eidetica works with a wide variety of customers, the company's focus primarily rests in the web-publishing domain. The company services customers from the Netherlands, Belgium, Germany, and the United States such as Trouw, Care4Cure, CWI, EULER (an EU project working to connect via Z39.50 and Dublin Core standards), LIMES, Filter Control Technologies, and PCM Uitgevers. The company's text mining solution is used by Mediargus "to process the content of all Flemish newspapers and enrich it with keywords every morning" (Eidetica) prior to transmitting it via FTP.

The company's name comes from the adjective *eidetic*, which refers to someone who has "the ability to close their eyes and imagine a previously perceived object so clearly that it is as if they are actually looking at it" (Eidetica). The company claims that this ability is reproduced in their software.

The company's technology, while primarily designed for information retrieval and search, does provide information extraction capabilities. Intra- and intersource link analysis can also be conducted through the use of the t-mining tool which establishes relationships among the extracted entities.

Output/Results

Extracted information is stored within the Eidetica database in XML format. Communication with the Eidetica's repository software is conducted through secure XML query and data upload protocols. Additionally, search results are presented to the user for manual analysis.

Application to Law Enforcement

Limited. As the company's focus lies in the publishing domain, direct application to law enforcement is not strong. The company provides only limited information retrieval and link analysis capabilities on a small number of entities. However, the application service provider portion may be an appeal to smaller law enforcement agencies if security and privacy issues could be reconciled.

Evaluation

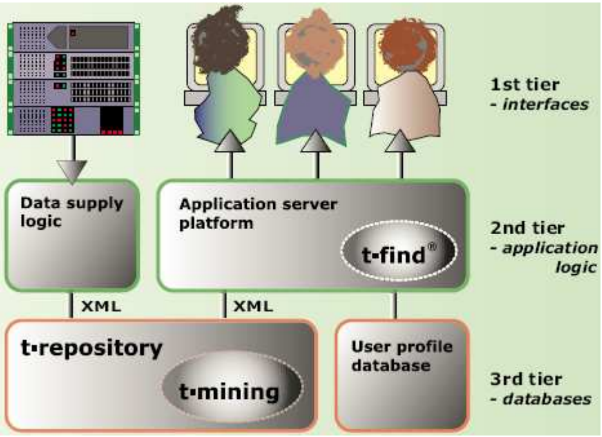
No evaluation information was available.

Financial

No financial information was available.

Software

The company is an application service provider which builds web pages for search, on-line publishing, and document categorization to integrate these components together. As an ASP, the solution is maintained by the company and is based on a central cluster of Linux application servers. Additionally, the technology has been developed as a three-tier architecture, as seen in the adjacent figure (Eidetica, a). The core technology is provided within the company's *t-repository* offering, "a cutting edge textual database and indexing system" (Eidetica). The company's search engine, *t-find*®, provides a web-interface to the index repository. This patented approach allows the system to guide the user's search through the use of options and suggestions to refine the results. Both "known-item searches" and broader "subject searches" can be performed. *t-mining* is the company's text mining solution, forming links among various types of information. More information on these offerings is available in the Algorithm section.



The company also uses a *language guesser* component which attempts to identify the language of a given text sample. The company utilizes a *web-crawler* to index web pages for storage within the Eidetica database. Access to information can also be controlled through classifications and the use of a "scrambler" module, which encrypts transmitted data.

Consulting is a primary emphasis of the Eidetica business model, and the company handles system set up and administration. The company also offers both a *protocol* and a *full* service model; the former provides the company's technology as a building block to a larger system, and the latter allows the company to fully maintain the system.

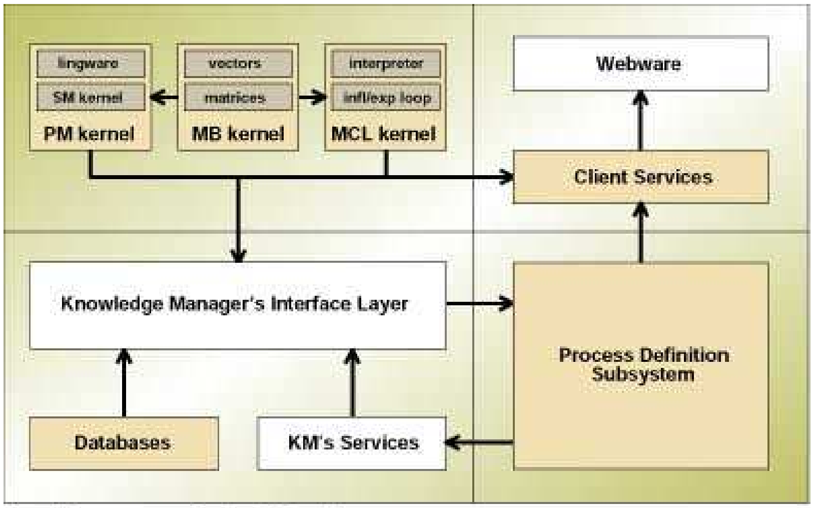
A demo of t-find® was available at http://cwi-opac.eidetica.com/ but was not active at the time of this survey. The company's language guesser has a demo at http://www.eidetica.com/services/guesser.

Inputs Required

The solution works with both structured and unstructured (free) textual sources. "As long as it's text, Eidetica solutions will be able to index it, mine it and possibly give it an extra spark of life" (Eidetica).

Link Analysis Algorithm

The company uses its *Hosted Knowledge* concept to uniquely combine "advanced and understandable search interfaces with text mining solutions" through the use of "content technology on the basis of software services" (Eidetica). "At the core of Eidetica's system is a proprietary clustering method…and advanced methods to extract subject keywords inside documents and titles" (Nieland, 1999). A high-level architecture of the solution is provided in the figure below (Nieland, 1999) and indicates that matrices and linguistic processing are also used.



Architecture of the Eidetica system.

t-repository is an XML-based indexing and mining system which filters and routes information based on criteria provided by the customer and Eidetica. Term extraction and indexing are performed as the system "actually reads the incoming text [and] filter[s] out the relevant subject terms and document features. It does not need dictionary vocabularies, precompiled thesauri or hand-made 'rules,' and yet through advanced statistical methods, is nonetheless capable of 'understanding' the content" (Eidetica). The extraction process also includes type integration to allow all elements (e.g., author, publication date, keywords, words, phrases, and character strings) to be treated uniformly. The extracted information is then indexed.

t-mining can link entities such as authors, publishers, time frames/dates, subjects, classification codes (e.g., Mathematics Subject Classification (MSC)), and terms used in text (Eidetica). The company claims that these links can be collected, filtered, clustered, connected, categorized, cleaned, enriched, and reversed. Automated classification (taxonomy-generation) is also available and is based on machine learning, language recognition and relationship discovery.

According to (Nieland, 1999), the process consists of five main steps:

1. "Merge the complete, miscellaneous document collection into a uniform format,
2. Read all documents to extract a dictionary of subjects,
3. Create various 'maps' of the collection: which documents address which subjects, what authors write about what subjects, what subjects are connected to other subjects,
4. Quality control: visualize the constructed maps and give the information manager tools to refine them, and
5. Use the subject maps to build browsing and querying interfaces that guide the user through the collection to find precisely the right information."

The technology utilized by the company includes the use of neural networks that require 200-1000 samples for training. Additionally, "human-supervised meta information" (Eidetica) can also be utilized to enhance the process and is incorporated into the system through the use of system suggestions. Fixed keyword lists or hierarchical systems are also utilized in the system, and multiple languages are able to be processed, as well.

Knowledge Engineering Cost

As the company claims that the solution does not need dictionaries or hand-made rules, the system utilizes a strong manual component. This is evident in the user feedback provided in the query process as well as in the use of manual enrichment of documents leading to findings that "well managed and enriched collections make better implementations of the t-find® search system" (Eidetica). Additionally, the information provided in the company's website indicates a strong preference for manual evaluation of search results.

While this is true, the solution is also driven primarily by the use of a proprietary clustering method. After taking these factors into consideration, it was decided that the technology should be classified as requiring a medium knowledge engineering cost.

Summary Table

Category: Commercial	
Hierarchy: NE	**Source Scope**: Intra and Inter
Company Name: Eidetica **Company URL**: http://www.eidetica.com/	**Location**: Amsterdam, the Netherlands
Solution Name: t-repository t-find® t-mining	
Domain Scope: general (emphasis on publishing domain)	**Application Type**: IE and LA
Knowledge Engineering Cost: medium	**Financial Cost**: unknown
Input Requirements/Preparation Required: any text (structured or unstructured)	
Link Analysis **Algorithm Name/Group**: proprietary clustering method; neural-type network **Labeling**: n/a **Labeling Supervision**: n/a **Model Generation**: hybrid **Model Generation Supervision**: supervised **Process Description**: Entities are clustered through the use of autonomies, machine learning, language recognition and relationship discovery, and the search results are displayed to the user.	
Solution Output: Search results are presented to the user for manual analysis.	
Application to Law Enforcement: limited	
Is performance evaluation available? no	**Solution/demo available**? yes

Sources

Eidetica. Available: http://www.eidetica.com/. Accessed January 24, 2006.

Eidetica (a). *Content Matters (Brochure).* Online. http://www.eidetica.com/content/downloads/Eidetica-brochure.pdf. Accessed January 24, 2006.

Nieland, Henk (1999). "Eidetica – A New CWI Spin-off Company." *Research and Development, ERCIM News, No. 37.* April, 1999. Online. http://www.ercim.org/publication/Ercim_News/enw37/nieland.html. Accessed January 24, 2006.

3.3.7 Endeca Technologies, Inc.

Company Introduction and Domain Scope

Cambridge, Massachusetts-based Endeca is yet another leading data mining company. With its named derived from the German word *entdecken* ("to discover"), the company was founded in 1999. Endeca's technology has been used in enterprise portals, intranets, websites, online self-service applications and within industries such as information publishers, manufacturers, financial services, and governments. The company's client base includes leading companies such as Wal-Mart, The Home Depot, Barnes and Noble, Bank of America, Putman Investments, IBM, Tesco, Texas Instruments, John Deere, and NASA. Endeca has also been the recipient of several awards and recognitions, such as a KMWorld Trend Setting Product (2004, 2005) as well as one of their "100 Companies that Matter" in Knowledge Management (2003 – 2005), an AlwaysOn Top 100 Private Company award (2004, 2005), an EContent "Matters Most" in the Digital Content industry (2002 – 2004), IndustryWeek's Technology of the Year (2004), and ComputerWorld Innovative Technology Award (2003).

While primarily a search tool, the company's solution incorporates both information extraction and link analysis technology.

Output/Results

While most of the results are presented visually to the user in response to the query, results can also be exported in MS Excel format or can be emailed to the user (text). When the query results are presented to the user via the Presentation API (as provided by InFront), XML objects are used.

Application to Law Enforcement

Extensive. While Endeca is currently being used by government intelligence agencies (such as the Defense Intelligence Agency (Solomon, 2005)), it is primarily being used by manufacturing and e-commerce companies. We believe that Endeca's solution represents an excellent technology for more extensive use in law enforcement applications.

Evaluation

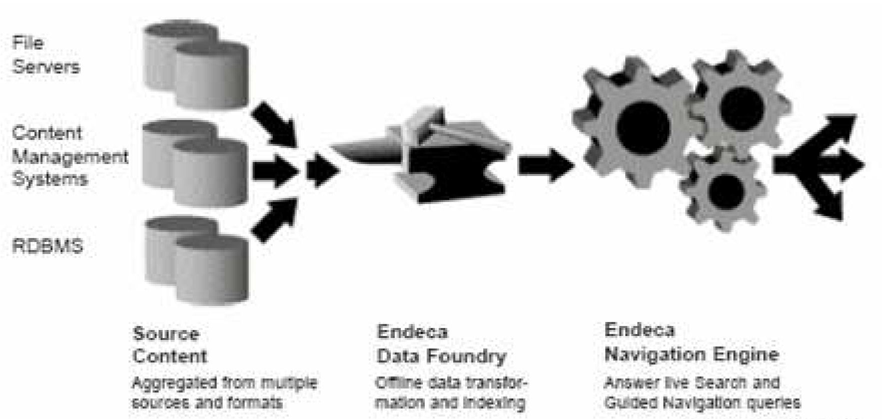
No detailed performance evaluations were found, although it is claimed that World Book experienced an increase in search speeds by a factor of 8 – 10 times (Endeca). However, "current deployments [of the Endeca Navigation Engine] scale to over a billion records, terabytes of contents, thousands of facets [dimensions], and support millions of users" (Endeca, 2005e). Combined with the large and varied client base (from Wal-Mart to IBM to NASA), Endeca's technology is robust and scalable, able to support many domains and vast quantities of data.

Financial

No information found at this time.

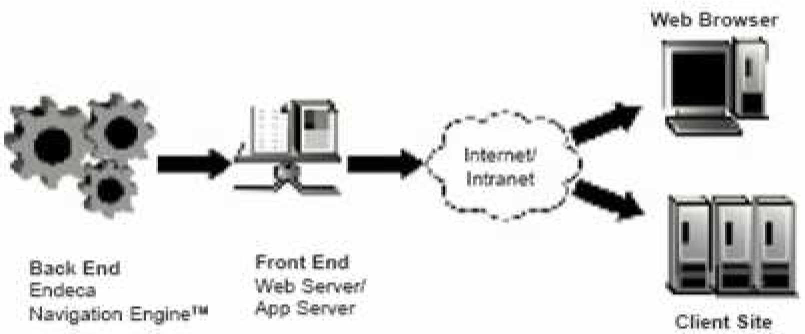
Software

The company has organized its solutions into three categories: enterprise search (ProFind), e-commerce search (InFront), and analytics (Latitude). However, driving each of these products is the company's Guided Navigation® system. At the heart of this system is the Navigation Engine ™, a two-tier architecture platform consisting of an application logic tier and a presentation logic tier. The application logic tier consists of three steps. The first step, *source data acquisition*, "extracts data from nearly any source


File Servers
Content Management Systems
RDBMS
Source Content
Aggregated from multiple sources and formats
Endeca Data Foundry
Offline data transformation and indexing
Endeca Navigation Engine
Answer live Search and Guided Navigation queries

Source: Endeca

system in nearly any language" (Endeca). Data is obtained from a variety of sources, including content management systems, enterprise resource systems, file servers, databases, and other textual content. Using the Endeca Content Acquisition System ("a full-featured crawler" (Endeca, 2005b) and other methods (data dump, FTP, ETL systems), unstructured (.doc, .ppt, .pdf, .txt, etc.), semi-structured (.xls, email, reports, etc.), and highly-structured (enterprise systems, Lotus Notes, MS Access, databases, etc.) data is entered into the Endeca Data Foundry. According to ClearForest (2003), most of Endeca's unstructured information extraction technology is performed using ClearForest's entity extraction technology. Then, the *configuration, modeling, and indexing* step occurs within the Data Foundry to perform "offline transformations that convert and standardize the source data into the form the live Endeca Navigation Engine will query" (Endeca, 2005b). Using Endeca Studio, a web-based GUI tool, search options, relevancy ranking modules, and business rules are formulated to "add editorial control to how metadata and other structured and unstructured information will be transformed into Guided Navigation" (Endeca, 2005b). This second step in the Navigation Engine also performs indexing, calculating the relationships between the source data, the data modules and configuration files by building a Meta-Relational Index (Endeca, 2005b). This index automatically discovers every valid navigation path to each record and is updated to reflect the most recently available data. With the data obtained and organized, the final step in the application logic tier is to *load and update the engine* with the indexes created in the foundry. The presentation logic tier consists of a single step, *query by end-user applications.* In this step, the user utilizes the Endeca Presentation API to query the Navigation Engine and mine the data. In summary, "data flows from original sources of all types into the Endeca Data Foundry™, where it is configured, modeled, and indexed. Then it is loaded onto the Endeca Navigation Engine for high-performance querying by end-user application through the Endeca Presentation API" (Endeca, 2005b).



Source: Endeca

Endeca ProFind® helps users to search through the information coordinated by the Navigation Engine. After the user enters their search query, ProFind "determines the meaning of each query using linguistic analysis, synonyms, and concept search" (Endeca, 2005e) and aids in the search by using phonetic and programmatic spelling correction, word stemming, wildcards, and bi-directional thesaurus (Endeca, 2005e). The system suggests search alternatives and allows phrase, fielded, Boolean, and within results searches. For sensitive information, ProFind incorporates secure sign-in to allow users to search the information content they hold permissions for (Endeca, 2005e).

Endeca InFront® utilizes the Guided Navigation and is similar to the ProFind, yet packages this technology for use in online retail and similar applications to enhance user product searches. Another variation, Endeca Product Data Navigator, allows manufacturing workers to quickly search for required materials parts and components critical to manufacturing processes by combining current inventories, content information providers, and vendor data. This has lead to millions of dollars in savings from reductions in direct materials costs, consolidation of purchases, streamlining supply chains, and improved field services.

Endeca's Latitude component is a Business Intelligence solution that utilizes Interactive Reporting. Released in December 2004, this tool extends interactive reporting to the middle of the business structural pyramid and simplifies the complicated and cumbersome process of navigating business data.

Demos are available by contacting the company and registering at http://endeca.com/register/registration_form.php.

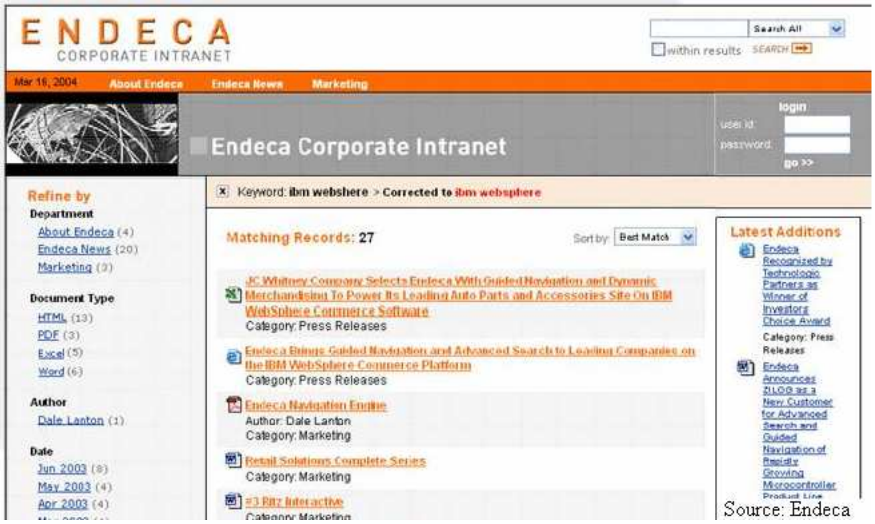
Inputs Required

The Navigation Engine can access over 370 different file formats and supports over 250 languages. While much is done automatically, the solution can also be configured to enhance and refine search options, relevancy ranking modules, and business rules in the formation of links. Configurations are performed using scripts (Perl, etc.), ODBC connections, as well as text and XML files.

Link Analysis Algorithm

As already described in the Software section, Endeca's technology is based on the Guided Navigation® system which utilizes the Endeca Navigation Engine™. Information extraction techniques are performed through the use of the solution's Endeca Content Acquisition System and the Endeca Data Foundry. According to ClearForest (2003), most of Endeca's unstructured information extraction technology is performed using ClearForest's entity extraction technology. This system joins all of the data sources, ranging from unstructured data to structured data into the Endeca Data Foundry. Here the Foundry "guides administrators to select and name fields" (Feldman, 2005) and also handles the configuration, modeling, and indexing of the data to normalize and structure the data through the use of Endeca Studio. Clients can also tune the search results returned by the solution to coincide with business goals (such as identifying "most popular" products or promoting new or special products). As a result of this process, all of the values are extracted into the Foundry are linked to each other via every possible navigation path and creates immense intersource link structures.

The strength of the Endeca solution lies in its link analysis technology, which is primarily enabled through its Navigation Engine. After the user enters their query, the query is expanded using linguistic analysis, synonyms, concept search, phonetic and programmatic spelling correction, word stemming, wildcards, and a bi-directional thesaurus. By analyzing the search results in this form, the search is then compared to the "universe of metadata" that consists of all the terms found within the dataset. The next step narrows that universe by removing all those categories that have not been tagged with the search terms. Then, within the remaining values, the categories are grouped into dimensions of related attributes. This results in not only information retrieval of the sources desired, but also creates a links to categories of sources. Using these categories, the search scope can continue to be narrowed to aid the user in the location of the desired information.



☒ Keyword: ibm webshere > Corrected to ibm websphere

Refine by

Department
About Endeca (4)
Endeca News (20)
Marketing (3)

Document Type
HTML (13)
PDF (3)
Excel (5)
Word (6)

Author
Dale Lanton (1)

Date
Jun 2003 (8)
May 2003 (4)
Apr 2003 (4)

Matching Records: 27

Sort by: Best Match

JC Whitney Company Selects Endeca With Guided Navigation and Dynamic Merchandising To Power Its Leading Auto Parts and Accessories Site On IBM WebSphere Commerce Software
Category: Press Releases

Endeca Brings Guided Navigation and Advanced Search to Leading Companies on the IBM WebSphere Commerce Platform
Category: Press Releases

Endeca Navigation Engine
Author: Dale Lanton
Category: Marketing

Retail Solutions Complete Series
Category: Marketing

#3 Ritz Interactive
Category: Marketing

Latest Additions

Endeca Recognized by Technologic Partners as Winner of Investors Choice Award
Category: Press Releases

Endeca Announces ZILOG as a New Customer for Advanced Search and Guided Navigation of Rapidly Growing Microcontroller Product Line

Source: Endeca

While "taxonomies can be imported to supply familiar terminology and categories" (Feldman, 2005), the system automatically analyzes the search terms and understands the appropriate categories. As the user updates the search and selects appropriate refining categories, the categories will be updated to represent the full depth and breadth of the search.

Knowledge Engineering Cost

From a link analysis perspective, the approach involves a high level of human interaction as the business rules, relevancy ranking modules, and search options are all configured by the user through the use of a GUI tool. While the GUI may enable a more simple process for the user, the fact that the user must explicitly configure the system and establish the links among the values leads us to categorize this solution as having a high knowledge engineering cost.

Summary Table

Category: Commercial	
Hierarchy: NE	**Source Scope**: Intra and Inter
Company Name: Endeca Technologies, Inc. **Company URL**: http://endeca.com/	**Location**: Cambridge, Massachusetts, USA
Solution Name: Endeca Search and Guided Navigation® (Endeca Content Acquisition System, Endeca Data Foundry, Endeca Studio, Endeca Navigation Engine™); Endeca ProFind®; Endeca InFront®; Endeca Latitude™	
Domain Scope: general	**Application Type**: IE and LA
Knowledge Engineering Cost: high	**Financial Cost**: unknown
Input Requirements/Preparation Required: Over 370 different file formats and over 250 languages are supported. Information can come from unstructured to structured sources.	
Link Analysis **Algorithm Name/Group**: proprietary **Labeling**: n/a **Labeling Supervision**: n/a **Model Generation**: manual **Model Generation Supervision**: supervised **Process Description**: After the search criteria is entered and expanded by the system, the user is given a network of links joining the search terms to the resulting sources. Navigating through these results refines the search and updates the remaining links.	
Solution Output: While search results are primarily presented to the user visually, results can also be saved via XML, MS Excel, or in text/email format.	
Application to Law Enforcement: extensive	
Is performance evaluation available? no	**Solution/demo available**? yes

Sources

ClearForest (2003). "Endeca and ClearForest Announce Strategic Partnership For Advanced Searching of Unstructured Data" March 31, 2003. Online. http://www.clearforest.com/whatsnew/PRs.asp?year=2003&id=34. Accessed December 2, 2005.

Endeca. Available: http://endeca.com/index.html. Accessed January 4, 2005.

Endeca (2005a). *Endeca InFront® for Online Retail*. Online. http://endeca.com/resources/pdf/Endeca_InFront_Overview.pdf. Accessed January 4, 2005.

Endeca (2005b). *The Endeca Navigation Engine*. Online. http://endeca.com/resources/pdf/Endeca_Technical_Overview.pdf. Accessed October 8, 2005.

Endeca (2005c). *Endeca Product Data Navigator*. Online. http://endeca.com/resources/pdf/ProductDataNavigator_Overview.pdf. Accessed January 4, 2005.

Endeca (2005d). *The Endeca ProFind® Platform for Search and Guided Navigation® Solutions.* Online. http://endeca.com/resources/pdf/Endeca_ProFind_Overview.pdf. Accessed October 8, 2005.

Endeca (2005e). *New Search and Discovery for the Federal Government.* Online. http://endeca.com/resources/pdf/Endeca_ProFind_Overview_Govt.pdf. Accessed January 4, 2005.

Endeca (2005f). *Product Data Information Access and Retrieval: The Missing Component of Manufacturers' PLM Strategy: Endeca Business White Paper for Manufacturers.* Online. http://endeca.com/resources/pdf/Endeca_Manufacturing_BusinessWP.pdf. Accessed January 4, 2006.

Feldman, Susan (2005). "Product Flash: Endeca's Latitude: Easy Access to Business Intelligence." *IDC #32716.* January, 2005. Online. http://endeca.com/resources/pdf/idc_bi.pdf. Accessed January 4, 2006.

Solomon, Jay (2005). "Investing in Intelligence: Spy Agencies Seek Innovation Through Venture-Capital Firm." *The Wall Street Journal* (Eastern edition). pg A.4. September 12, 2005. Online. http://endeca.com/about_endeca/news/n_091205_wsj.html Accessed January 4, 2005.

3.3.8 InferX Corporation

Company Introduction and Domain Scope

InferX, headquartered in McLean, VA, works with distributed data mining technology and was founded in 1999 as a spin-off of Datamat Systems Research, Inc. Datamat is a professional R&D services firm which was founded in 1992 to develop technology for distributed analysis of sensory data relating to airborne missile threats for the Department of Defense and the Missile Defense Agency contracts (InferX). The technology used by InferX was developed under a Small Business Innovation Research (SBIR) Grant and targets the threat detection market.

The company specializes in the homeland security, insurance, and financial industries, serving such clients as Lockheed Martin, Northrop Grumman, the Air Force Research Lab, the Missile Defense Agency, and George Mason University. Expansion into the healthcare industry is expected soon. The company currently has no partners and consists of approximately 20 employees. Jesus Mena, a respected author in the data mining and homeland security, serves as one the company's board of advisors.

Since this solution works in a distributed database environment, this solution should be classified as an intersource link analysis solution.

Output/Results

Little detail is provided other than that the results are displayed in a "user-friendly format" (InferX). It is believed that the results are primarily presented in visual format through the use of InferView (see Software section). Results to the predictive solutions may be emailed to users to identify and flag the important objects or values. For instance, specific containers on ships were able to be flagged and identified for further search by customs agents (InferX, b).

Application to Law Enforcement

Extensive. The capabilities of these solutions are appropriate to the law enforcement community. By allowing data to be analyzed in a distributed environment, the data stays with the owner, "thus eliminating privacy and timeliness issues of warehousing data" (InferX), two issues prevalent in the law enforcement community. Additionally, having a central intelligence server alleviates the cost of having a centralized server or a data warehouse and also allows near real-time analysis that can quickly adapt to changes in data.

Evaluation

No detailed results were available. The company states that the solution is "robust and scalable, being able to mine millions of records for patterns" (InferX).

Financial

According to MDA (2004), the InferAgent solution is available for $250,000 to $1,000,000. A lease-per-month option is also available.

Software

According to the company's website, "InferX sets a new standard in distributed data mining, coupled with predictive analytics and traditional expert analysis…Its patent-pending, distributed data mining technology can recognize unanticipated patterns in data stored in geographically dispersed databases" (InferX). Launching its solution in 2002, the company provides two product offerings. The InferAgent™ suite is the company's main offering and is "agent-based software for analyzing multiple networks of databases to detect threats and opportunities" (InferX, a). (More information is found in the Algorithm section.) The InferView™ solution is a companion product to the InferAgent and represents "a knowledge discovery and integrated 3D visualization tool that supports stand-alone analysis" (InferX, a). The component supports stand-alone data mining analysis and is licensed for single database use (InferX).

Inputs Required

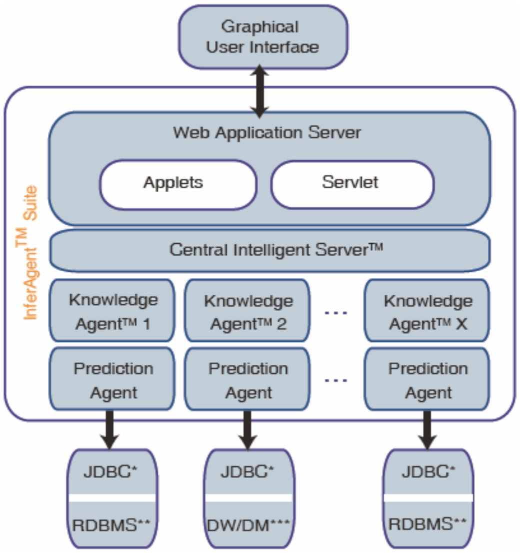
A distributed network of databases is required for the solution to work. Oracle 8*i*, SQL Server 2000, and MS Access databases are supported (InferX, 2004). Additionally, biometric (face, figure, iris scans, etc.) and demographic (age, height, education, etc.) data, as well as information about objects and processes, can be used in the solution (InferX, b).

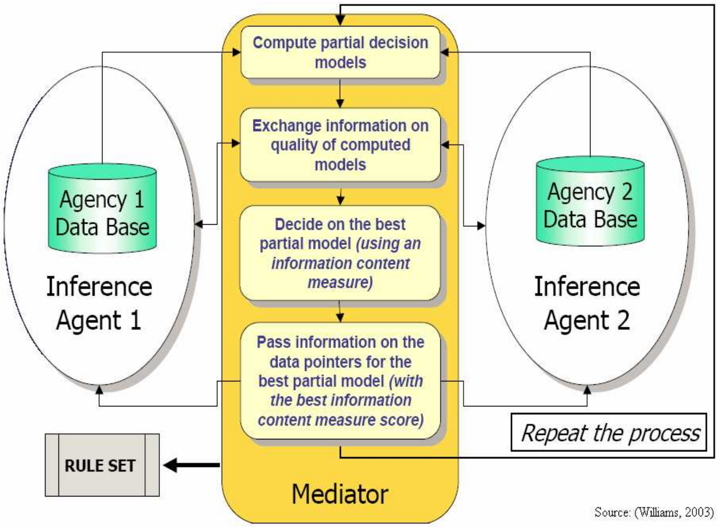
Link Analysis Algorithm

InferAgent is the company's main link analysis solution. By combining analysis and communication, this solution can identify hidden relationships such as changes in behavior or frequency in data distributed among databases through the use of "a combination of decision-tree algorithms with Bayesian networks" (Mena, 2004) (InferX). The company utilizes rule-based models to correlate data elements across distributed data repositories to infer information (Williams, 2003).

InferX's patented product suite uses what it terms independent *Knowledge Agents*™ *(KA)* "to transport algorithms, processes, and other necessary programs across networks to analyze data at its source" (InferX). The agents are installed via networks or the Web at each local site. These agents control the integrity of the local data sources and search local sites for patterns and behavior, passing on their correlated knowledge (which is done in conjunction with the *Prediction Agent*) without also passing the data.

After discovering unanticipated patterns from multiple databases, each agent reports results to the *Central Intelligent Server*™ *(CIS)* in real-time through the use of a Java-based Mediator which "synchronizes the information from each location and generates global models" (MDA, 2004). The CIS also generates scores and alerts users to take action. In this way the intelligence resides in the networks and not in the databases (InferX). A high-level diagram is presented to the lower left (InferX, 2004), and a conceptual diagram is presented below to the right (Williams, 2003).




Compute partial decision models
Exchange information on quality of computed models
Decide on the best partial model (using an information content measure)
Pass information on the data pointers for the best partial model (with the best information content measure score)
Mediator
Agency 1 Data Base
Inference Agent 1
Agency 2 Data Base
Inference Agent 2
RULE SET
Repeat the process

Source: (Williams, 2003)

In the description of the system presented in (InferX, b), users are first walked through a selection of the available databases. From these databases, appropriate tables and attributes are selected. This creates a data source that is saved as both a raw data file and a metadata file. Roles are then generated and a prediction model and a prediction data file are created. The results are presented to the user and can be emailed directly to the appropriate individuals.

Knowledge Engineering Cost

While some rules are required for the use of this link analysis solution, apparently not all rules require manual generation, as (InferX, b) claims that the rules and risk models are generated by the system itself and is not coded by people. The KEC is thus estimated to be medium.

Summary Table

Category: Commercial	
Hierarchy: NE	**Source Scope**: Inter
Company Name: InferX Corporation Company URL: http://www.inferx.com/	**Location**: McLean, VA, USA
Solution Name: InferAgent™ InferView™	
Domain Scope: general (distributed data)	**Application Type**: LA
Knowledge Engineering Cost: medium	Financial Cost: $250,000 - $1,000,000
Input Requirements/Preparation Required: distributed network of databases	
Link Analysis **Algorithm Name/Group**: combination of decision-tree algorithms with Bayesian networks **Labeling**: n/a **Labeling Supervision**: n/a **Model Generation**: hybrid **Model Generation Supervision**: semi-supervised **Process Description**: Independent *Knowledge Agents™ (KA)* are installed within each database to analyze the individual datasets. After discovering unanticipated patterns from multiple databases, each agent will report the results to the *Central Intelligent Server™ (CIS)* in real-time. The CIS will generate global models and pass the information on to users.	
Solution Output: Visual representation and textual results to generated rules	
Application to Law Enforcement: extensive	

Is performance evaluation available? no	Solution/demo available? no

Sources

InferX. Available: http://www.inferx.com/. Accessed January 12, 2006.

InferX (a). *InferX Fact Sheet*. Online. http://www.inferx.com/inferx_facts.pdf. Accessed January 12, 2006.

InferX (b). *A Next Generation Targeting System for Container Security Risk Assessment.* Flash multimedia presentation. Online. http://www.inferx.com/inferxcsra.zip. Accessed January 12, 2006.

InferX (2004). *Technical Specifications for the InferAgent™ Suite*. Online. http://www.inferx.com/technicalspec.pdf. Accessed October 7, 2005.

MDA (2004). Missile Defense Agency, Advanced Systems, Technology Applications Program. "Data Analysis: Datamat Systems Research, Inc./InferX". *2004 MDA Technology Applications Report.* 2004. Online. http://www.inferx.com/MDA_Techreview_2004.pdf. Accessed January 12, 2006.

Mena, Jesus (2004). "Homeland Security as Catalyst." *Intelligent Enterprise*. July 1, 2004. Online. http://www.intelligententerprise.com/showArticle.jhtml?articleID=22102265. Accessed June 2, 2005.

Williams, Al (2003). "InferX Corporation: An Innovative Approach to Turning Distributed Data into Decision-Relevant Knowledge." *Presentation at the NewTECH Showcase: Decision Support Tools for the Virginia Center for Innovative Technology*. August 19, 2003. Online. http://www.cit.org/pdf/events/08-19-03-inferx.pdf. Accessed January 12, 2006.

3.3.9 Inxight Software, Inc.

Company Introduction and Domain Scope

Inxight Software Inc. is based in Sunnyvale, CA and is focused on "information discovery from unstructured data sources" (Inxight). A spin-off from Xerox Palo Alto Research Center (PARC), the company was founded in 1997 and holds over 75 patents in information visualization, natural language processing, and information retrieval. The company works with 300 Global 2000 customers, including such companies as Air Products, Factiva, Hewlett Packard, LexisNexis, IBM, Oracle, Reuters, SAP, SAS, and Thomson. Inxight is also financed by In-Q-Tel and works with the U.S. Department of Defense and the Defense Intelligence Agency in their efforts.

The company provides both information extraction and link analysis solutions.

Output/Results

The extracted information is exported in XML format.

Application to Law Enforcement

Extensive. Inxight's technology is not only financed in part by In-Q-Tel, the Central Intelligence Agency's venture capital arm, but is also being used by many government agencies, such as the Department of Defense and the Defense Intelligence Agency. Inxight also works with many company's that provide their technology to others, such as ClearForest, Hummingbird, IBM, Oracle, SAS, and SAP.

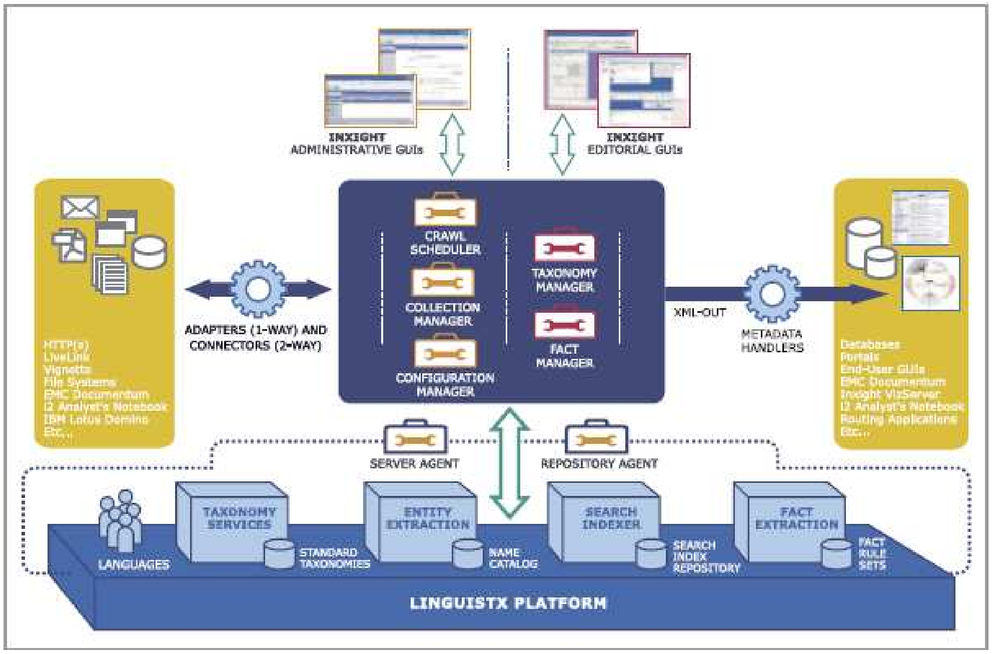
Evaluation

No performance results were found.

Financial

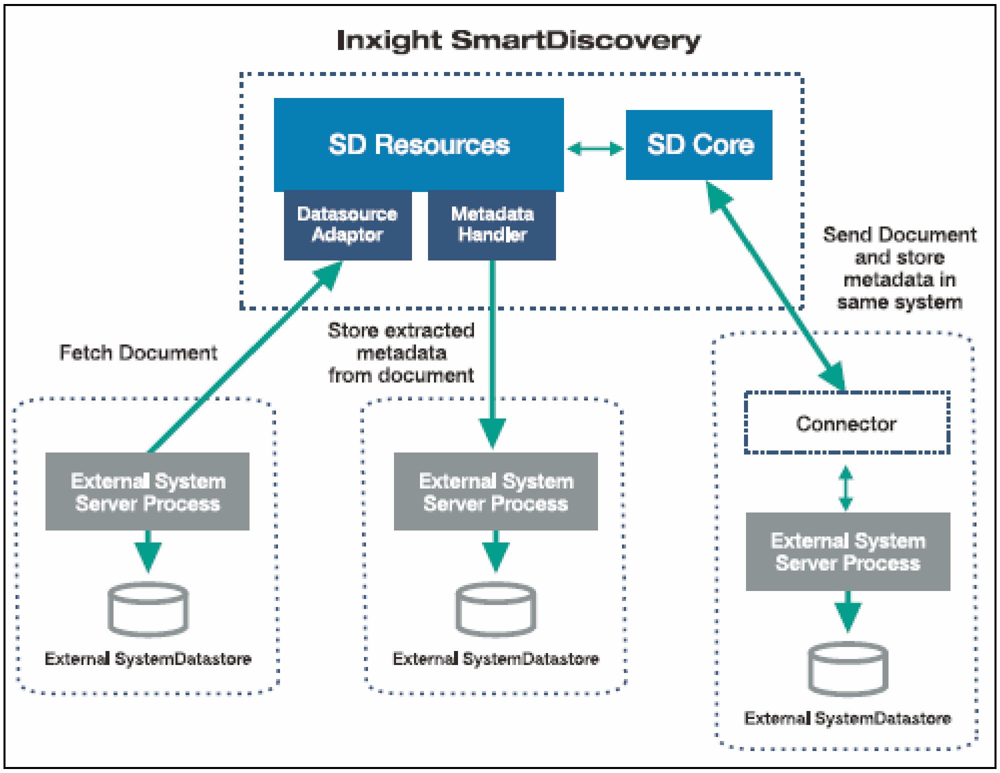
No details about the cost of the solution were found.

Software



Inxight offers a suite of tools for information extraction and link analysis. The company's flagship product, SmartDiscovery® incorporates several components that are also available individually.

Inxight has identified five key requirements that are involved in the knowledge transformation process. *Organizing* the data automatically classifies the data into topics/subjects as well as naming the entities. *Enriching* the content of the information involves "applying XML meta-tags to documents that embed characterizations of the document's topics, key entities, hyperlinks to related information, and summaries" (Inxight, a). The *Collection/Aggregation* requirement integrates content from multiple, disparate sources into a single useful source of information. *Normalization* (processing and refining the data) and *Data Personalization* (sending the information to the right person in the right format) are the final two requirements. Each of these requirements is met by one of the


Inxight SmartDiscovery
SD Resources
SD Core
Datasource Adaptor
Metadata Handler
Fetch Document
Store extracted metadata from document
Send Document and store metadata in same system
Connector
External System Server Process
External System Server Process
External System Server Process
External SystemDatastore
External SystemDatastore
External SystemDatastore

solution components they provide.

The highest level division of Inxight's solutions follows the company's five-step method by providing an Analysis Server and an Awareness Server. While the Awareness Server monitors the results of the analysis and communicates those results appropriately, the Analysis Server provides the information extraction and link analysis tasks and will, therefore, be the focus of the following description.

Information extraction capabilities are provided by Inxight ThingFinder, an automatic entity extraction component. Entities themselves are extracted by the LinguistX® Platform, working through several steps to extract named entities (see Algorithm). Currently, the company has developed 27 key entity types that can be extracted automatically without requiring any setup or manual creation of rules. These include the following named entity types: address, city, company, country, currency, date, day, holiday, internet address, measure, month, noun group, organization, percent, person (position, given name, family name, suffix, affiliation), phone number, place (regions, political areas, geographical areas), product, social security number, state, ticker symbol, time, time period, vehicle (make, model, color, VIN, license plate), and year. The company also offers ThingFinder Advanced/ThingFinder Professional as an add-on module to allow the user to define custom entity types using regular expression patterns (see Algorithm).

SmartDiscovery also incorporates taxonomy and categorization capabilities. These capabilities allow taxonomy structures and new categories to be developed based on both the context and content of the data through the use of terms, phrases, rules, sample documents, and filters – all while incorporating existing and/or publicly available taxonomies. With regards to document categorization, the various documents and sources can be classified by the XML meta-data that is generated and the documents can be grouped under several taxonomies.

Fact extraction, which learns events, activities, and relationships from text, is also provided by with the SmartDiscovery solution. Entity aliasing and co-referencing; normalization; fuzzy, partial, and order-free analysis; grammatical phrase, sentence, paragraph, and document recognition; sub-rule invocation; and grammatical analysis (pronominal resolution, subject-object, main verb identification, part-of-speech, concatenation, etc.) are all provided as part of this process.

The company also provides normalization and personalization through the use of their visualization component, VizServer™. VizServer provides three types of visualization: StarTree (which creates a network "tree" of relationships that can be moved through), TableLens (which looks for patterns, trends, and correlations in a table format), and TimeWall (which allows events to be chronologically viewed by category).

These solutions also support a large number of languages. Currently, over 30 languages are supported, including English, Chinese, Farsi, Arabic, German, Greek, Spanish, and Japanese.

The solution is available only through purchase. No demos or trial versions are available.

Inputs Required

Inxight solutions can accept data in a wide variety of forms. Over 220 file formats are supported, including Microsoft Office documents, pdf, XML, HTML, text and email.

Link Analysis Algorithm

The approach that Inxight takes is complex, as is evidenced by the many solution components that are available as part of their SmartDiscovery® system. The company has divided their capabilities into three general categories: entity extraction, relationship and event extraction, and visualization. *Entity extraction* creates metadata about the data within sources that can later be used to review, route, reference, and search. *Relationship and event extraction* allows users to create links between the extracted entities to identify and monitor trends and events associated with the entities (van Zuylen, 2004). *Visualization technologies* then permit the users to identify the specific information they are looking for (van Zuylen, 2004).

As mentioned in the Software section, the company's information extraction component, ThingFinder, is driven in large part by their LinguistX® Platform. By turning grammatical relationships into mathematical formulas (Shachtman, 2005), this platform can intelligently analyze text by providing *automatic language and character encoding identification* for over 30 languages. Once this step has been completed, a *document analysis* is performed to segment paragraphs and provide a high-level overview of the text. *Word segmentation (tokenization)*, *stemming*, and *de-compounding* are then used to granulize the text and reduce the text into base forms to be used in the learning processes. *Part-of-speech tagging* allows the forms to be given context before the *noun phrase extraction* utilizes the above steps to extract the information.

The company has provided 27 such extraction modules which automatically run through the entity extraction for the user. However, ThingFinder Advanced also allows the user to develop his or her own rules. In developing the rules, the user can "define custom entity types as patterns of contiguous tokens in regular expression syntax, enriched with morphological word stems and Part-of-Speech tags" (Inxight, b). Literal strings (i.e., a set sequence of characters, such as *a* or *Paris*), regular expression symbols (e.g., |, *, and ( )), part-of-speech tags (e.g., <bomb POS:Nn> refers to a *bomb* when used as a noun), and morphological stems (e.g., <STEM:attack> includes *attacks*, *attacking*, *attacked*, etc.).

At the end of this process, the entities have all be extracted and classified. ThingFinder also provides variant identification and grouping (to identify similar entities (e.g., Mr. Doe and John Doe)) and normalization (e.g., turning May 12 = 05/12) as well as handling misspellings to enhance the information extraction and link analysis tasks. As a final step, relevance ranking is also provided by the system to give the extracted entities a measurement to reflect their importance to the document as a whole. "A sentence's relevance…depends on the number of thematic words and proper names, its location in the document, and the length of the document" (van Zuylen, 2004).

Once the entities have been extracted, SmartDiscovery also performs *fact extraction*, which includes events, activities, and relationships from the data. "Using a visual, intuitive fact workbench environment, [the user] can define, test, and implement fact templates and rules that are unique" to the individual's needs (Inxight, 2005e). Therefore, while the solution is designed in a way to aid the user in the development of rules, manual rules are still generated to perform link analysis.

Link analysis tasks are also performed by the user. Search capabilities enable the user to input various criteria on which to analyze. Visual capabilities depicting the links are provided by VizServer and also aid the user in link analysis activities.

Knowledge Engineering Cost

In terms of link analysis, the solution does an excellent job of presenting the relationships to the user in a very efficient and concise manner. The solution automatically suggests taxonomies that can be used by the user, and categorizes documents based on entities found within the various sources. It also provides excellent visualization tools to allow the user to navigate through the various entities. During the fact extraction process, the user also is required to develop manual rule templates (see Algorithm). Additionally, the user is required to input search terms to be able to identify links. Given these reasons, we have classified the link analysis component of this solution as having a high knowledge engineering cost.

Summary Table

Category: Commercial	
Hierarchy: NE	**Source Scope**: Intra and Inter
Company Name: Inxight Software, Inc.	**Location**: Sunnyvale, CA, USA

Company URL: http://www.inxight.com/	
Solution Name: SmartDiscovery Analysis Server (LinguistX Platform, ThingFinder, ThingFinder Advanced, Fact Extraction, Taxonomy and Management Categorization), SmartDiscovery Awareness Server, VizServer	
Domain Scope: general	**Application Type**: IE and LA
Knowledge Engineering Cost: high	**Financial Cost**: unknown
Input Requirements/Preparation Required: over 220 file formats are supported, including Microsoft Office documents, pdf, XML, HTML, text and email	
Link Analysis **Algorithm Name/Group**: proprietary **Labeling**: n/a **Labeling Supervision**: n/a **Model Generation**: manual **Model Generation Supervision**: n/a **Process Description**: User generates rules that are used to determine *facts* that are learned from the text. Taxonomies and categorizations are also manually created within the system. However, human intervention is required to enter search terms before the relationships are presented graphically.	
Solution Output: results are output in XML	
Application to Law Enforcement: extensive	
Is performance evaluation available? no	**Solution/demo available**? no

Sources

Inxight. Available: http://www.inxight.com/. Accessed December 1, 2005.

Inxight (a). *Corporate Fact Sheet.* Available: http://www.inxight.com/pdfs/corp_fact_sheet.pdf. Accessed December 1, 2005.

Inxight (b). *ThingFinder Advanced with Custom Entity Extraction*. Online. http://www.inxight.com /pdfs/Inxight_ThingFinder_Advanced_ds.pdf. Accessed November 1, 2005.

Inxight (2004a). *Inxight SmartDiscovery: Entity Extraction.* Online. http://www.inxight.com/pdfs/EntityExtraction_FinalWeb.pdf. Accessed November 15, 2005.

Inxight (2004b). *Inxight SmartDiscovery: Taxonomy and Categorization.* Online. http://www.inxight.com/pdfs/Taxonomy_FinalWeb.pdf. Accessed November 15, 2005.

Inxight (2005a). *Inxight SmartDiscovery Analysis Adapters and Connectors.* Online. http://www.inxight.com/pdfs/SD_Adapters_Datasheet.pdf. Accessed December 22, 2005.

Inxight (2005b). *Inxight SmartDiscovery Analysis Server.* Online. http://www.inxight.com/pdfs/SmartDiscovery_AS.pdf. Accessed November 15, 2005.

Inxight (2005c). *Inxight SmartDiscovery Awareness Server.* Online. http://www.inxight.com/pdfs/SmartDiscovery_FinalWeb.pdf. Accessed December 22, 2005.

Inxight (2005d). *Inxight SmartDiscovery: Fact Extraction.* Online. http://www.inxight.com/pdfs/FactExtraction_Web.pdf. Accessed November 15, 2005.

Inxight (2005e). *Inxight Software, Inc. Company Fact Sheet.* Online. http://www.inxight.com/pdfs/corp_fact_sheet.pdf. Accessed November 15, 2005.

Shachtman, Noah (2005). "With Terror in Mind, a Formulaic Way to Parse Sentences." *New York Times.* New York, NY. March 3, 2005. Online. http://www.nytimes.com/2005/03/03/technology/circuits/03next.html?ex=1135141200&en=b7e59924788a2cdb&ei=5070. Accessed August 11, 2005.

van Zuylen, Catherine (2004). *Insight: From Documents to Information: A New Model for Information Retrieval.* October, 2004. Online. http://www.inxight.com/pdfs/InxightInformationRetrieval.pdf. Accessed November 28, 2005.

3.3.10 Language Analysis Systems, Inc.

Company Introduction and Domain Scope

Language Analysis Systems, Inc. (LAS), a privately-held company, was founded in 1984 by Drs. Leonard Shaefer and Jack Hermansen and is currently based in Herndon, VA. Based on their work in linguistic and computational properties of personal names at Georgetown University in Washington, DC, the company's technology has been used extensively by U.S. Government agencies (including U.S. Intelligence and Border Patrol agencies) and technology firms and is installed in more than 200 countries around the world. According to a US government study on name search technology, "[t]he chief investigators at LAS are by far the most knowledgeable people working in the name search problem domain" (LAS).

The company's focus is on "supplying its multi-cultural name recognition products to international government and commercial clients with mission-critical name matching problems." Their technology is patent-pending, "the first patent-pending technology in the name-matching field since the Soundex algorithm was patented in 1922" (LAS). The company also has many partners that utilize its technology, such as Acxiom, ClearForest, IBM, ilogs, Infoglide, Microsoft, Oracle, MITRE, Lockheed Martin, Sun, and Visual Analytics.

Output/Results

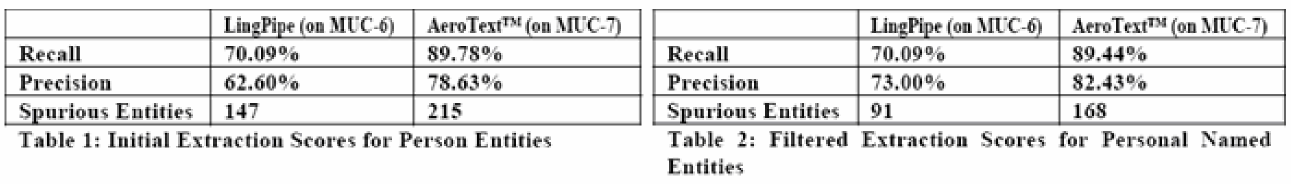
Linked names are identified by the system.

Application to Law Enforcement

Extensive. LAS's technology has been deployed extensively in government applications, including such agencies as the FBI. Additionally, LAS's technology was used by federal authorities to track the September 11th terrorists to their Florida connections.

Evaluation

Little information was found. However, in Williams and Patman (2005), an analysis of the incorporation of the *NameStats*™ solution demonstrated how the use of large name data stores with filtering logic could "significantly reduce the number of extracted spurious personal names without having any consequential impact on recall…[while also allowing] for the creation of broader rules to extract more entities without decreasing the precision." Using both Lockheed Martin's AeroText and Alias-I's LingPipe solutions to extract name information from the MUC-6 and MUC-7 corpora, the LAS technology results in the two tables below (Williams and Patman, 2005) demonstrate significant improvements in performance.



	LingPipe (on MUC-6)	AeroText™ (on MUC-7)
Recall	70.09%	89.78%
Precision	62.60%	78.63%
Spurious Entities	147	215

Table 1: Initial Extraction Scores for Person Entities

	LingPipe (on MUC-6)	AeroText™ (on MUC-7)
Recall	70.09%	89.44%
Precision	73.00%	82.43%
Spurious Entities	91	168

Table 2: Filtered Extraction Scores for Personal Named Entities

For more information and other examples, please refer to Williams and Patman (2005).

Financial

In 2004, pricing for the company's NameParser™ solution started at $10,000 (DMReview). Pricing for the other solutions was not available.

Software

LAS offers a wide range of solutions. NameClassifer™ identifies the cultural classification of a personal name. NameParser™ ensures consistency within the name data held within databases. By organizing the names into consistent formats, the information can be more accurately classified and more easily identified and searched. LAS NameHunter™ provides character-based search capabilities for names, while MetaMatch™ provides phonological (sound-based) search capabilities. NameStats™ utilizes NameParser™, but also identifies titles and name phrases within the name data and calculates frequency information. NameVariationGenerator™ "generates a set of possible alternative Romanized spellings of a name" (LAS, 2004). A cultural-specific linguistic rule-set converts the name into a sophisticated pattern to be used in comparison against the LAS Name Data Archive. NameGenderizer™ determines the gender of a given name and the NameReferenceLibrary (NRL)™ is a name-encyclopedia that contains "culture-specific information about names, their use, their meanings, and their patterns of spelling variation" (LAS, 2004). The Name Data Archive™ provided by the company includes nearly a billion names (Williams and Patman, 2005) to be used in comparison and analysis. LAS's most recent product, NameInspector™, "extends data profiling to name data" (LAS). This analysis tool detects parsing issues and erroneous names while also determining distribution information (gender and culture) and other valuable information (LAS, 2004) (LAS).

Online demos of LAS solutions are available with registration at the company's website (http://demos.las-inc.com/register_demo.asp).

Inputs Required

The solution requires that the name to be analyzed be provided to the system. In this manner, it requires named entities that are themselves names to perform. Names can be input in a variety of textual formats, such as Unicode or ASCII.

Link Analysis Algorithm

The Soundex algorithm has been the cornerstone of industry for nearly a century. "Soundex is the name-searching system still used by 90 percent of American businesses, almost every government department and major airline—even though it was originally developed for the 1890 census. It takes a person's last name, strips out the vowels and assigns codes to similar-sounding consonants to create a four character code—the first letter followed by three digits to represent the consonants" (Duffy, 2004). While simple, the algorithm will also produce the same code for similar individuals and is "blind to the cultural differences between names around the world. It treats three-syllable Asian names in the same manner as it treats eight-syllable Arabic or Hispanic names. The last name Zhang in China becomes Chang in Taiwan, Khiu in Thailand, Cheung in Singapore and Teoh in Malaysia. Soundex is incapable of recognizing that those names may indicate the same person and does not take into account cultural or language differences" (Duffy, 2004).

LAS's technology provides a more complete solution. First, the algorithm applies a *culture-specific matching criteria.* This involves identifying the name's culture of origin by applying a set of matching techniques to a given name. From the user's perspective, this is performed automatically. The second step is the *automatic application of linguistic rules for the culture/language context.* In order to parse the name and generate spelling variants, rule-based, algorithmic, statistical/probabilistic, or combinational approaches are used. Additionally, names can be phonetically or alphabetically compared. LAS has compiled many millions of names by which to use an automated statistical and linguistic approach (LAS, 2004).

The third step handles *noise tolerance* or typographical errors that can occur in data entry. *Recognition of equivalent but dissimilar name variants* allows the user to identify additional name links (such as Elizabeth and Betty or Paco and Francisco). *Ranked returns* are then presented in decreasing rank to the provided name; these rankings are based on sound, spelling, and cultural variation patterns. *Statistical and probabilistic search aids* are used to help the user identify the matching names and *syntactic flexibility* helps to accommodate for white-space variations and order-varied data (e.g., <last name>, <first name> versus <first name> <last name>). *Adjustment and tuning* can also be performed on the system to fit search results by adjusting the quality and quantity (i.e., precision and recall) of its results. Reference tools are also provided with the solution.

Knowledge Engineering Cost

Given the complexity of the system and the significant amount of name-list use and complexity of the patterns, the knowledge engineering cost of LAS's technology is high.

Summary Table

Category: Commercial	
Hierarchy: NE	**Source Scope**: Inter
Company Name: Language Analysis Systems **Company URL**: http://www.las-inc.com/index.shtml	**Location**: Herndon, VA, USA
Solution Name: NameInspector, NameParser, NameClassifier, NameHunter, MetaMatch, NameVariationGenerator, NameReferenceLibrary, NameGenderizer	
Domain Scope: name recognition	**Application Type**: LA
Knowledge Engineering Cost: high	**Financial Cost**: $10,000 quoted for NameParser (2004)
Input Requirements/Preparation Required: named *Name* entities were identified	
Information Extraction **Algorithm Name/Group**: proprietary **Labeling**: n/a **Labeling Supervision**: n/a **Model Generation**: manual **Model Generation Supervision**: n/a **Process Description**: Solution identifies the name's culture before applying rules to parse the data, generate additional spellings and patterns, and allowing search capabilities to perform link analysis among names.	
Solution Output: Name errors are corrected and data is normalized. Link analysis search tool produces name links and similarity rankings among possible names.	
Application to Law Enforcement: extensive	
Is performance evaluation available? no	**Solution/demo available**? yes

Sources

LAS. Language Analysis Systems, Inc. Available: http://www.las-inc.com/index.shtml.

LAS (2004). *Advanced Name Recognition Technology: An Overview.* Obtained via email correspondence on October 12, 2005.

Duffy, Diantry (2004). "What's in a Name?" *briefing, CSO Online.* January, 2004. Online. http://www.csoonline.com/read/010104/briefing_name.html. Accessed December 27, 2005.

DMR (2004). DMReview.com Web Editorial Staff. *LAS Announces New Name Parser.* April, 2004. Online. http://www.las-inc.com/media_coverage/2004/Apr04/04-23-04_DMReview.pdf. Accessed December 27, 2005.

Williams, Kemp and Patman, Frankie (2005). *Personal Entity Extraction Filtering Using Name Data Stores.* 2005. Online. https://analysis.mitre.org/proceedings/Final_Papers_Files/33_Camera_Ready_Paper.pdf. Accessed January 12, 2006.

3.3.11 Language Computer Corporation: Using predicate-Argument Structures for Information Extraction

Solution Introduction and Domain Scope

This solution was developed by researchers at the Language Computer Corporation, Richardson, Texas. It aims to extract information from textual data. This approach is scalable to different domains. For example, it could be used to extract stock market price changes or "death events" from newspapers. We categorize it as intra document LA, since it extracts events from Web pages, where each Web page is considered a source. A given event is extracted from a single Web page, instead of two or more Web pages.

Output/Results

The output is desired events. For example, a stock market change event could be "London gold fell $4.7 cents to $308.35." The output is also a filled template, as seen in the adjacent figure (Surdeanu et al., 2003).

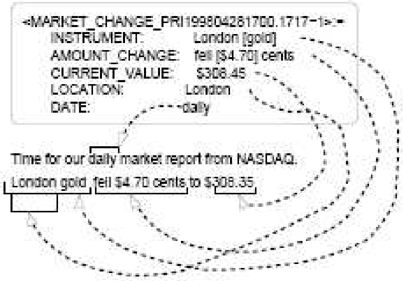

<MARKET_CHANGE_PRI199804281700.1717-1>:=
INSTRUMENT: London [gold]
AMOUNT_CHANGE: fell [$4.70] cents
CURRENT_VALUE: $308.45
LOCATION: London
DATE: daily
Time for our daily market report from NASDAQ.
London gold fell $4.70 cents to $308.35

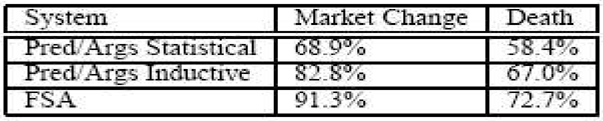
Figure 1: Templette filled with information about a market change event.

Application to Law Enforcement

Moderate. This solution could be used in law enforcement to extract the crime events but would need to be adapted.

Evaluation

Two domains are selected to evaluate the LA paradigm proposed in this solution: "market change" event and "death" event. An approach using a predicate is compared with an approach using finite state automata (FSA). In the adjacent table (Surdeanu et al., 2003), F-scores for the two domains are listed. Although the solution using FSA has the best performance, it requires an effort of 10 person days



System	Market Change	Death
Pred/Args Statistical	68.9%	58.4%
Pred/Args Inductive	82.8%	67.0%
FSA	91.3%	72.7%

Table 3: Templette F-measure (F_1) scores for the two domains investigated

per domain. On the other hand, the only human effort needed in the predicate solution is imposed by generating a mapping between arguments and template slots, which could be accomplished in less than two hours per domain. The assumption is that the training data has already been labeled. If this is not the case, extra KEC will be needed to label data. The table also shows that the predicate solution with inductive learning performs better than it does using a statistical method.

Inputs Required

The input is textual data, which could be online news or digital documents.

Link Analysis Algorithm

This solution describes a domain-independent LA paradigm which is based on predicate-argument structures. As labeling is already assumed to have occurred, no labeling is required in terms of this solution. The predicate-argument structures can be automatically identified by two different methods, one is the statistical method reported in Surdeanu et al. (2003); another is a new method based on inductive learning.

Some statistical methods have been used to predict argument roles and semantic roles (Surdeanu et al., 2003). The statistical technique labels predicate arguments on the output of a probabilistic parser. It consists of two tasks: (1) for each predicate, identifying the parse tree corresponding to the arguments of this predicate and (2) recognizing the role of each argument. For the example shown in the figure above (Surdeanu et al., 2003), the first step identifies two noun phrases. Then, in the second step, the two parameters are assigned "ARG1" and "ARG0" as roles, given the predicate "assailed."

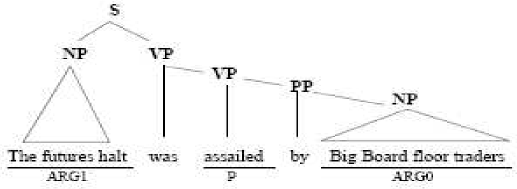


Figure 2: Sentence with annotated arguments

Statistical methods always don't perform well when data is sparse. As a result, another inductive learning algorithm (decision tree) is used to identify parameters or predicates and assign their roles. The C5 inductive decision tree learning algorithm (Surdeanu et al., 2003) is used to implement the classifier for argument constituents and the classifier for roles of arguments.

Following is the paradigm presented in Surdeanu et al. (2003):

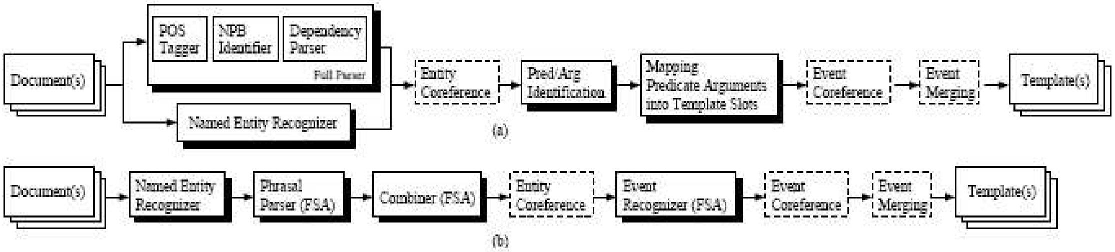


Figure 7: IE architectures: (a) Architecture based on predicate/argument relations; (b) FSA-based IE system

The figure demonstrates the use of predicate argument structures. First, documents are parsed using a full parser and named entities are recognized. Then, the parsed texts marked with named entity tags are passed to the "Entity Co-reference" module to revolve pronominal and nominal anaphors and normalize co-referring expressions. In "Pred/Arg Identification" module, one of the methods talked about above (statistical or inductive learning) is used to identify parameters and their roles. Then, for each domain, a mapping between predicate arguments and template slots is produced. One example of mapping can be seen in Figure 9 of Surdeanu et al. (2003), reproduced below.

The architectures in the above figure 7 (a) and (b) both share the same name entity recognition, co-reference and merging modules; the difference is the FSA-based approach uses a different phrasal parser and combiner. The comparison of performance for both of the architectures can be seen in the Evaluation section.

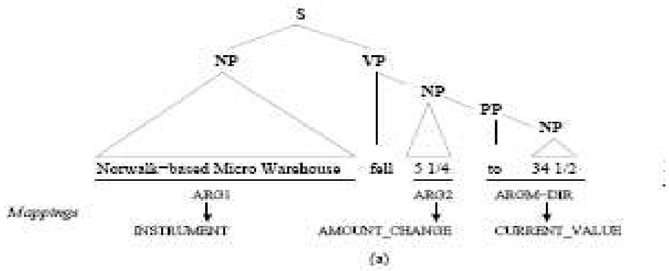


Figure 9: Predicate argument mapping example domain

Knowledge Engineering Cost

We conclude the KEC is medium, since the training data needs to be labeled manually, and the mapping between predicate arguments and template slots also needs to be done manually.

Summary Table

Category: Academic	**Hierarchy**: NE	**Source Scope**: Intra
Company Name: Language Computer Corporation **Company URL**: http://www.languagecomputer.com/	**Location**: Richardson, Texas, USA	
Solution Name: Using Predicate-Argument Structures for Information Extraction		
Domain Scope: general	**Application Type**: IE and LA	
Knowledge Engineering Cost: medium	**Financial Cost**: unknown	
Input Requirements/Preparation Required: textual data		
Link Analysis **Algorithm Name/Group**: using predicate-argument structure **Labeling**: manual **Labeling Supervision**: n/a **Model Generation**: automatic **Model Generation Supervision**: supervised **Process Description**: First documents are parsed using a full parser and NEs are recognized. Then the parsed texts marked with NE tags are passed to “Entity Co-reference” module to revolve pronominal and nominal anaphors and normalize co-referring expressions. In “Pred/Arg Identification” module, as statistical or inductive learning method is used to identify parameters and their roles. Then, for each domain, a mapping between predicate arguments and template slots is produced.		
Solution Output: events (filled predefined template)		
Application to Law Enforcement: moderate		
Is performance evaluation available? yes	**Solution/demo available**? no	

Sources

Surdeanu, Mihai; Harabagiu, Sanda; Williams, John and Aarseth, Paul (2003). “Using Predicate-Argument Structures for Information Extraction.” *In Proceedings of the 41st Annual Conference of the Association for Computational Linguistics (ACL-03).* Pages 8–15. Online. http://acl.ldc.upenn.edu/P/P03/P03-1002.pdf Accessed January 21, 2006.

3.3.12 Megaputer Intelligence Inc. / Megaputer Intelligence Ltd.

Company Introduction and Domain Scope

Beginning as a research and development group in Artificial Intelligence at Moscow State University in 1989, Megaputer Intelligence became a commercial entity first in 1993 in Moscow, Russia (Ltd) before incorporating in the United States (Inc) in 1997. According to the company's website, "The mission of Megaputer is to provide customers around the world with top quality software tools for transforming raw data into knowledge and facilitating better business decisions" (Megaputer). Although not a large company, Bloomington, Indiana-based Megaputer boasts quite an impressive client base working with over 300 customers globally, primarily in the customer support, analytics, safety, insurance, market research, and government industries. These include organizations and companies such as 3M, Best Buy, Taco Bell, the Center for Disease Control (CDC), Dow, Pfizer, Liberty Mutual, IBM, Raytheon, Boeing, EDS, Sprint, Ask Jeeves, Airbus, the National Institute of Standards and Technology (NIST), the US Navy, and several universities (e.g. the University of Pennsylvania, Rutgers). The company also has several partners, including Cambridge Technology Partners, Microsystems (Moscow) as well as major players IBM and Microsoft.

The company has software capabilities in both the information extraction and link analysis fields in their data mining packages.

Output/Results

The TextAnalyst process stores the knowledge base in a computer's RAM, where it is used to perform link analysis. Other than visual output through GUI tools, the stored data is not kept in a particular format, nor is the original source modified. However, textual reports are generated and can be saved.

Application to Law Enforcement

Moderate. While Megaputer offers a wide variety of options in the analysis of the data, it does not perform an in depth analysis of the data. However, the various algorithms and link analysis techniques applied by the solution represent good possibilities for law enforcement work.

Evaluation

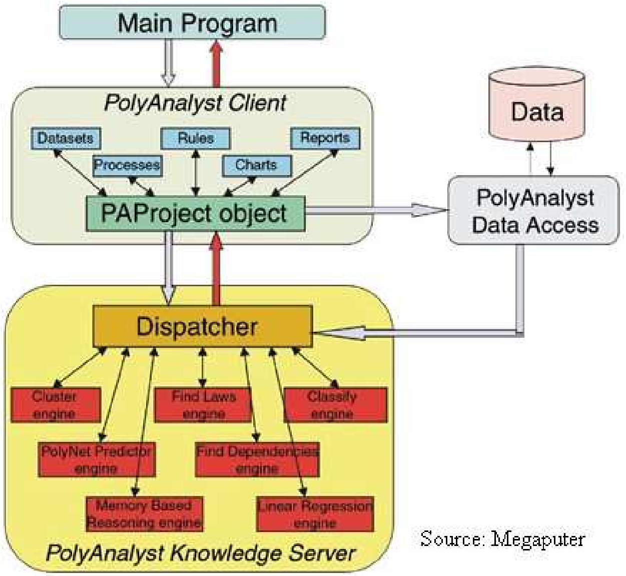
According to the company, TextAnalyst can process up to 20-40 MB of text and stored the entire knowledge base in RAM. For a given an amount of text, three to four times that amount of memory is required to store all of the relationships and links between terms and fragments discovered within the text.

Financial

In 2000, the price of the solution depended upon the algorithms chosen, ranging in price from $2,300 to $14,900, and the developer kit was an additional $16,000 (Apicella, 2000). The company claims to have the best "price/performance" ratio and is given support by Apicella's classification of the PolyAnalyst product as "competitively priced."

Software

Megaputer offers several different solutions that have applicability to a variety of clients. The company's base product, TextAnalyst "is a data mining tool for analyzing unstructured text. It is



Source: Megaputer

designed to derive key concepts from text articles by delivering semantic analysis and performing summarization" (Megaputer). However, it is important to note that the TextAnalyst solution was developed by Megaputer in cooperation with Microsystems, Ltd. (http://www.analyst.ru), and Megaputer serves as the worldwide distributor (outside of the Commonwealth of Independent States) of TextAnalyst. For an analysis of TextAnalyst, see the Algorithm section. TextAnalyst for Microsoft Internet Explorer provides information extraction capabilities within the internet browser and a COM component of the technology is also available. TextAnalyst SDK, available from Microsystems, allows users to customize their own information extraction programs.

Megaputer's main offering is the PolyAnalyst solution, "the world's most comprehensive and versatile suite of advanced data mining tools. PolyAnalyst incorporates the latest achievements in automated knowledge discovery to analyze both structured and unstructured data" (Megaputer). Version 4.6 is the latest offering, improving upon the program's efficiency, algorithms, and use (including drill down capabilities, etc.). The program's information extraction components (which the company refers to as Text Mining or Text Analysis) are provided primarily through the use of TextAnalyst algorithms. However, upon consolidating the data, PolyAnalyst employs a large number of data mining algorithms that can be used to analyze and mine the textual data. PolyAnalyst Knowledge Server is a DCOM-based solution that allows the technology to be used in an enterprise setting, while COM components allow the algorithms to be obtained individually.

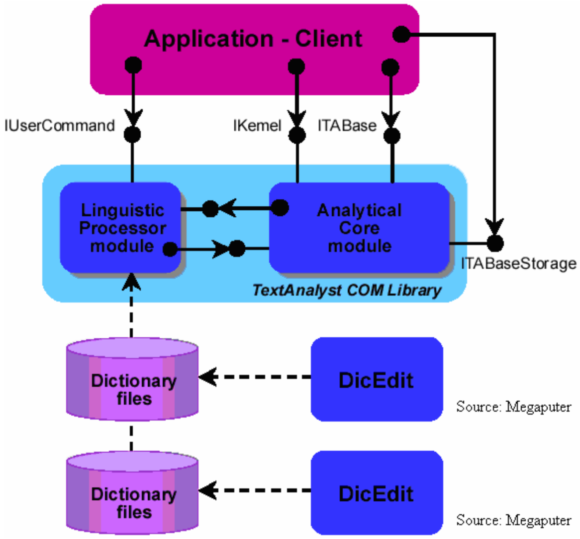
The company also offers a few other solutions. Client Shepherd provides a powerful link analysis visualization tool, presenting important customer information for business managers. WebAnalyst incorporates Megaputer's technology into websites to allow users to search and navigate the site (Megaputer). X-SellAnalyst aids users in e-commerce by analyzing user transactions and making recommendations in real-time to improve company growth (Megaputer, 2002).

Megaputer offers 30 day demos of nearly all of its offerings at http://www.megasysdev.com/webdown/prodlist with registration. Microsystems offers the TextAnalyst SDK at http://www.analyst.ru/index.php?lang=eng&dir=content/products/.

Inputs Required

PolyAnalyst can take in a wide variety of inputs ranging from structured data sources (as from ODBC or OLEDB links) to semi-structured (CLOB fields holding text documents) to unstructured textual files (such as HTML, .doc, .txt, flat files).

Link Analysis Algorithm



Source: Megaputer

Source: Megaputer

Megaputer/Microsystems's solution TextAnalyst™ (currently version 2.1) is an information extraction system. Utilizing both linguistic and Hopfield-like neural network technology, the user is able to search through textual samples, generate summaries (size is controlled by a semantic weight threshold), and further analyze the text. The component consists of two parts, a *Linguistics Processor* (the text preprocessing module) and the *Algorithmic Core* (the text analysis module). Through the use of a user-specified dictionary and linguistic rules, the user can control which word sequences and their attributes will be extracted from the text and included in the focus of a particular subject. The sequence is then passed to the Algorithmic Core, "where semantic analysis is performed with the help of neural network technology" (Megaputer). This creates a *semantic network* ("a set of the most important concepts from the text and the relations between these concepts weighted by their relative importance" (Megaputer)) and the terms in the dictionary are mapped to the terms located within the document. This creates a tree-like topic structure that represents the semantics of the investigated texts, with more important subjects located near the tree's root (Megaputer); clustering is also performed. Given the analyzed and organized data, the user is able to enter a natural language query. This query is "analyzed for semantically important words and all relevant sentences from the textbase documents are retrieved" (Megaputer).

Microsystems provide even more detail into the solution's offering. The solution "has been developed on the basis of neural network technology for complex, automatic semantic analysis of texts, semantic search, document subject classification and automatic creation of knowledge bases, hypertext links and abstracts" (Microsystems). TextAnalyst automatically identifies main topics (word-combinations and words) and their relationships. The solution also estimates their relative values and presents them hierarchically, indexing and classifying the sources. This allows for semantic information search, as well.

Megaputer follows a four-step process within the PolyAnalyst solution: *preprocessing*, *analysis*, *refining and comprehension*, and *reporting and scoring*. As already mentioned, TextAnalyst is used to generate a collection of the most important terms, count them, and tag the original sources with the discovered patterns of terms (a process termed *Semantic Text Analysis*) as well as incorporate "synonyms and particular instances of a term" (in *Focused Semantic Analysis*) to create the extracted information. Therefore, the values are extracted to hierarchical neural network and then statistically weighted prior to comparison. The source is then assigned to a taxonomy (*taxonomy categorization*) which was developed by the user or the system (automatically; can be adjusted later) (*taxonomy creation*). The system also handles eliminating duplicate records and allows batch (folder) processing.

After these steps have been completed, the user is able to apply many different algorithms by which the data can be analyzed. For instance, the PolyAnalyst Link Terms engine "reveals unexpected patterns and clusters of information hidden in data…[and] displays the results in a visual form facilitating further interactive manipulations of data" (Megaputer, 2003). TextOLAP "allows the user to define dimensions for the analysis of text notes and quickly roteate and dissect data across dimensions of interest in order to obtain aggregrated reports of interest" (Megaputer, 2003). The Link Analysis engine discovers and visualizes patterns of multi-order relationships, providing a link between data-driven automated discovery and the analyst's domain expertise (Megaputer, 2003). Additionally, the system allow the use of an additional dozen algorithms, listed below:

- Classify (Fuzzy logic classification)
- Cluster (Localization of Anomalies)
- Decision Forest
- Decision Tree (Information Gain)
- Discriminate (Unsupervised classification)
- Find Dependencies (Multidimensional distribution analysis)
- Find Laws (Symbolic Knowledge Acquisition Technology - SKAT)
- Market Basket Analysis (Transactional data processing)
- Memory Based Reasoning (Multiple group classification)
- PolyNet Predictor (Neural Net and GMDH hybrid)
- Stepwise Linear Regression
- Summary Statistics

For further information on these algorithms, please see http://www.megaputer.com/products/pa/algorithms/.

Knowledge Engineering Cost

Due to the large number of automated algorithms existent within the PolyAnalyst program, we have assigned a *medium* knowledge engineering cost rating. After the data has been processed, the algorithms are able to be utilized by the user through the GUI and automatically analyze the data.

Summary Table

Category: Commercial	
Hierarchy: not NE	**Source Scope**: Intra and Inter
Company Name: Megaputer Intelligence, Inc. **Company URL**: http://www.megaputer.com/	**Location**: Bloomington, Indiana, USA Moscow, Russia
Solution Name: TextAnalyst (TextAnalyst COM, TextAnalyst for MS Internet Explorer) PolyAnalyst (PolyAnalyst Knowledge Server, PolyAnalyst COM) Client Shepherd WebAnalyst X-SellAnalyst	
Domain Scope: general	**Application Type**: IE and LA
Knowledge Engineering Cost: medium	**Financial Cost**: $2,300 to $14,900; $16,000 (developer kit) (from 2000)
Input Requirements/Preparation Required: PolyAnalyst can analyze textual data of any structure.	
Link Analysis **Algorithm Name/Group**: (various) **Labeling**: hybrid **Labeling Supervision**: supervised **Model Generation**: (various) **Model Generation Supervision**: (various) **Process Description**: The solution uses user created and automatic-system generated taxonomies to group the data prior to applying one of several data mining algorithms to mine the data.	
Solution Output: The TextAnalyst process stores the knowledge base in a computer's RAM and the link analysis output is provided visually or stored in generated reports.	
Application to Law Enforcement: moderate	
Is performance evaluation available? yes	**Solution/demo available**? yes

Sources

Ananyan, S. and Kharlamov, A. *Automated Analysis of Natural Language Texts.* Online. http://www.megaputer.com/tech/wp/tm.php3.

Apicella, Mario (2000). "PolyAnalyst 4.1 Digs Through Data for Gold." *InfoWorld*. June 30, 2000. Online. http://www.infoworld.com/articles/es/xml/00/07/03/000703espoly.html. Accessed January 4, 2006.

Megaputer. *Megaputer Intelligence, Inc.* Available: http://www.megaputer.com/ Accessed January 4, 2006.

Megaputer (2002). *X-SellAnalyst™*. Online. http://www.megasysdev.com/down/wm/white_papers/x_sellanalyst.pdf. Accessed October 8, 2005.

Megaputer (2003). *PolyAnalyst for Text: Text Mining System.* Online. http://www.megasysdev.com/down/dm/pa/docs/PolyAnalyst_for_Text_brochure.pdf. Accessed October 8, 2005.

Microsystems. *Microsystems, Ltd.* Available: http://www.analyst.ru/ Accessed January 4, 2006.

3.3.13 NetOwl (SRA International)

Company Introduction and Domain Scope

NetOwl® is the text mining technology product line of Fairfax, Virginia-based SRA International, "a leading provider of information technology services and solutions - including strategic consulting; systems design, development and integration; and outsourcing and managed services - to clients in national security, civil government, and health care and public health" (SRA). The software began with research and development work for the U.S. government in the early 1990s and the first version of the solution was released in 1996.

NetOwl products are used extensively by the U.S. government as well as several major commercial entities, such as Edgar Online People, Thomson Gale, Gannet Co., Inc, iLumin, KnightRidder, and LexisNexis. Through SRA, the company also has many partners, including such companies as Microsoft, Oracle, Siebel Systems, and Tivoli. As another example, NetOwl technology is utilized in iLumin's Assentor® email surveillance and archiving product (NetOwl). NetOwl solutions have also has also garnered much recognition in conferences (see Evaluation section).

As the solution not only extracts entities, but also forms intrasource links between them, the solution is categorized as both an information extraction and link analysis technology.

Output/Results

In the papers presented at the MUC-7 conference, the results were input and output using Standard Generalized Markup Language (SGML)-marked up texts (Aone, et. al, 1998) (Krupka and Hausman, 1998). According to the most recent publication (NetOwl, 2005a), the system supports XML input and output, which includes the Web Ontology Language (OWL). "Many popular analytical tools such as OLAP, link analysis and visualization, GIS, and data mining tools can be applied to texts once they are structured by NetOwl Extractor" (NetOwl, 2005a). Additionally, translation of foreign language entities into English is also available.

Application to Law Enforcement

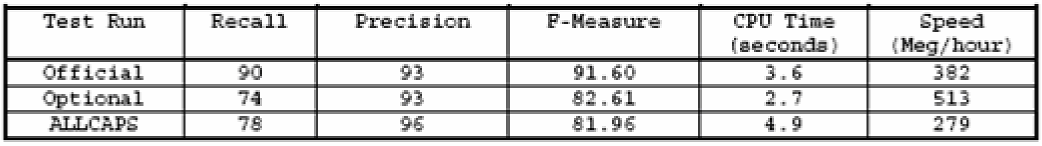
Extensive. NetOwl software originated in the 1990s for work specifically in the government domain. While the SRA subsidiary IsoQuest, Inc. was formed in 1996 to understand the market potential for their technology, government applications have remained a focus for NetOwl technology. NetOwl technology "has been deployed extensively through the U.S. Government" (NetOwl) and is also a recipient of federal funding. Beginning in February 2000, the company began to receive funding from In-Q-Tel "to apply its NetOwl® text mining technology to support specific user functions, including information retrieval for a daily briefing of world events…The In-Q-Tel funded enhancements applied the power of NetOwl to identify events and relationships and create structured data from unstructured text" (SRA, 2000a).

Needless to say, NetOwl also aids homeland security efforts. "It has become clear that the United States needs better means to handle the vast amounts of unstructured data that contain critical information necessary to defend our homeland. The Government receives unstructured data in many forms: hard-copy documents - even hand-written ones, faxes, e-mails, Web pages, etc. It comes in many different languages, some where the U.S. has very few human analysts skilled in them....Defending the homeland requires a seamless, technology-driven environment where analysts have at their fingertips data from a multitude of sources in a structured, usable format. NetOwl technology provides a means of achieving these goals" (NetOwl).

Evaluation

The company prides itself on the success of its product. According to the company's website, "NetOwl has demonstrated its accuracy through state-of-the-art performance over many years in Government-sponsored benchmarking for text mining technology. For example, NetOwl posted the highest score ever achieved for name extraction from unformatted text, a score which has never been equaled by another system. In recent benchmarking, it has also achieved the highest score for link extraction of any participant" (NetOwl).

NetOwl competed in the most recent Message Understanding Conference (MUC-7) held in the spring of 1998 (when NetOwl was a product of SRA subsidiary IsoQuest, Inc) using NetOwl Extractor 3.0. At this conference, the solution achieved the performance detailed in Krupka and Hausman (1998). The solution was run on a Pentium II 300 MHz processor and produced the following results for named entity extraction (Krupka and Hausman, 1998):



Test Run	Recall	Precision	F-Measure	CPU Time (seconds)	Speed (Meg/hour)
Official	90	93	91.60	3.6	382
Optional	74	93	82.61	2.7	513
ALLCAPS	78	96	81.96	4.9	279

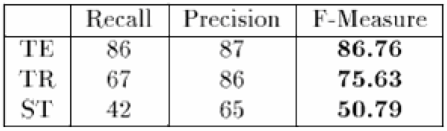
Table: NE Test Results Source: (Krupka and Hausman, 1998)

"The *Official* run utilized the full pattern rule base to perform the maximum analysis, achieving the best results at the slowest speed. The *Optional* run used about 20% of the rules to perform the minimum analysis, achieving a lower performance at the greatest speed" (Krupka and Hausman, 1998). The *ALLCAPS* run was configured to achieve a high precision due to the fact that case-sensitive rules could not be utilized; if manually re-tagging had been performed, the results would most likely have been improved (Krupka and Hausman, 1998). In summary, the solution "demonstrated that the drop in performance was mainly due to the document style combined with the change in domain of the formal test documents, and showed how to improve performance with simple additions to the lexicon….[NetOwl] demonstrated its high speed and low memory" (Krupka and Hausman, 1998). For more information, the reader is directed to (Krupka and Hausman, 1998).

The report also mentions that data runs were able to be performed on a Pentium 133 MHz laptop at 140 MB/hour and 190 MB/hour.

SRA also entered a separate solution in the MUC-7 conference, as documented in Aone, et al. (1998). As some of the technology used in their entry has now been incorporated into the NetOwl solution, a discussion of their results is included here. Termed the Information Extraction Engine (IE^2) System, the NetOwl Extractor 3.0 was used for entity named recognition using NameTag, PhraseTag, and EventTag elements (which are currently available as NameTag and Link and Event configurations within NetOwl Extractor Version 6 (NetOwl, 2005a)).

On the three tasks performed (Template Element (TE), Template Relation (TR), and Scenario Template (ST)), SRA achieved the results presented in the adjacent figure (Aone, et. al, 1998), the highest score in each of the three tasks entered (Aone, et. al, 1998). Additionally, time performance evaluations were conducted for each on each of the tasks using a SUN Ultra (167 MHz) with 128 MB of RAM to process 100 test texts: TE: 11 minutes, 17 seconds (an additional 5:38 was needed with coreference capabilities added); TR: 18:59; ST: 19:22.



	Recall	Precision	F-Measure
TE	86	87	**86.76**
TR	67	86	**75.63**
ST	42	65	**50.79**

Table: SRA's Scores for TE, TR and ST

Financial

No specific financial costs were available.

Software

NetOwl's solution is available in four different product offerings. The company's main product, NetOwl Extractor, incorporates the information extraction technology. Version 6 is the most recent version and uses "advanced computational linguistics and natural language processing technologies" to accurately find and classify key concepts in unstructured text (NetOwl). The solution extracts links and events connecting people, organizations, and items as well as identifying new patterns. A Java-based Visual Extractor enhances this process. (See Algorithm for more detail.)

NetOwl Summarizer uses the company's technology to generate abstracts and summaries of documents through a combination of linguistic, statistical, and learning techniques. The system is "trainable" and allows the user to select the length of the summary (NetOwl).

NetOwl InstaLink is the most recent offering provided by the company. This Java-based link analysis solution provides "advanced visualization, information extraction, and plan recognition technology to provide a visual means of linking critical information from disparate sources" (NetOwl). Link information is automatically updated with drag-and-drop capabilities to incorporate unstructured textual sources as well as a highly scalable data ingestion which accepts news feeds, document submissions, and structured data sources. The solution allows real-time maintenance and updatability of active situation displays (NetOwl).

NetOwl TextMiner is the company's main product offering, integrated a full text search engine, clustering capabilities, RDBMS, and various visualization tools in addition to NetOwl Extractor and NetOwl Summarizer. The solution automatically retrieves, analyzes, extracts, summarizes, and visualizes large amounts of unstructured data. It also combines search and retrieval, extraction, clustering, summarization, visualization, and translation capabilities. NetOwl solutions also offer multi-threading capabilities. Company-support is required for installation and maintenance as the company will determine needs, build and adjust the system, and provide consulting assistance.

A small demo of NetOwl's capabilities on a few sample documents (compared with AeroText and METIS) is provided on the web at http://im-dev-1.industrialmedium.com/xp/IC__working/AeroText/SMLA/040505_SMLA_IRAN.xml

Inputs Required

NetOwl solutions can take in a wide variety of unstructured and structured textual data. Over 200 different document types are supported, including UTF-8, XML, and OWL. Language support exists for English, Arabic, Chinese, Farsi (Persian), Korean, Thai, Russian, and all the Roman alphabet languages (Spanish, French, etc.).

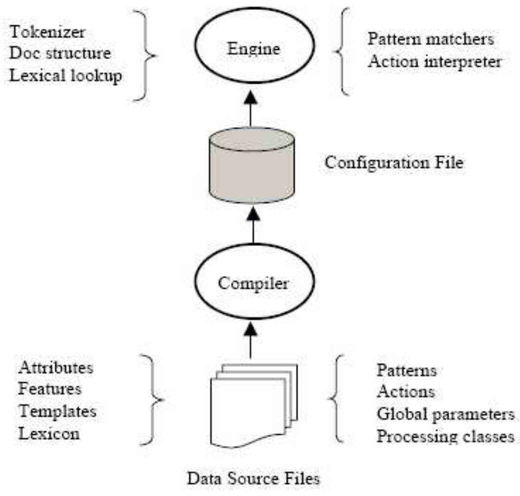


Figure 2: NetOwl Extractor Configuration Compiler

Information Extraction Algorithm

As mentioned in the Software section, the company's core technology is provided in its Extractor product offering. As it "extracts not only entities but also links and events that involve these entities" (NetOwl, 2005a), the NetOwl Extractor can be viewed as both an information extraction and a named entity link analysis solution. NetOwl extractor is available in two separate configurations: *NameTag* and *Link and Event*. The NameTag Configuration extracts seven types and over 60 subtypes of important entities. The seven category types (and a few examples of the subtypes) are: *Person* (*civilian*, *military*), *Organization* (*company*, *education*, *facility*, *religious*), *Place* (*astronomical*, *city*, *country*, *water*, *landform*), *Numeric* (*credit card*, *phone*, *SSN*, *VIN*), *Artifact* (*drug, vehicle*, *weapon*), *Time* (*age*, *date*, *duration*), *Address* (*email*, *IP*, *street*, *URL*), and *Concept* (*currency*). "The lexicon and pattern rule base define what the engine recognizes, a template (tag) specification and action definitions define what the engine extracts, and the processing classes define the distinct processing phases that the engine performs" (Krupka and Hausman, 1998). Name ambiguity is handled through the use of a rule completion phase which selects the most probable name interpretation; using each rule's numeric weight, the solution factors in the length of each interpretation and sums the values according to the type of tags (Krupka and Hausman, 1998). Strong evidence is indicated by a high rule weight, weak evidence by low rule weights, and negative rule weights indicate counter-evidence (Krupka and Hausman, 1998).

The Link and Event Configuration extracts over 100 types of links (such as affiliations and transactions). As it requires the named entities to carry out link analysis, this configuration also extracts all of the NameTag entities. The event extraction "does not just identify the presence of a certain event – it identifies the participants and their roles, and also attaches date and location information of the event" (NetOwl, 2005a). Link types include links based on *Place* (*place near*, *place parent location*), *Organization* (*founder*, *location*, *nationality*), *Person* (*address*, *affiliation*, *parent*, *phone*, *sibling*), *Artifact* (*maker*, *owner*), and *Address* (*component*). *Event* types include *Personnel Changes* (*hire*, *contract*), *Politics* (*appoint*, *elect*, *nominate*), *Law* (*acquit*, *arrest*, *jail*, *sue*), *Transactions* (*buy artifact*, *give money*, *travel*), *Conflicts* (*attack target*, *kill*, *surrender*), *Crime* (*extort money*, *steal*), *Finance* (*currency moves up/down*, *stock moves up/down*), *Business* (*acquire company*, *merge company*, *sell company*), *Vehicles* (*spacecraft launch*, *vehicle crash*), and *Family* (*die*, *marry*).

"NetOwl uses natural language processing, rather than keywords, to find information and has the ability to recognize a word as a person, place, or company" (SRA-IQT). Linguistic context analysis allows dynamic recognition and concept classification, while additionally providing alias resolution, normalization, and translation of entities from foreign languages to English. According to the company, their Extractor can also be viewed as "an automated meta-tagging tool, whereby organizations can tag and manage their enterprise content in an effective way" (NetOwl, 2005a). The extractions are dependent upon the use of the core Extractor engine and various *Configurations*. These configurations and ontologies are tailored to Subject Domains, such as Business, Finance, Homeland Security, Intelligence, Law Enforcement, Politics, and various languages. User-defined concepts are also able to be extracted through Creator Edition.

More detail on the inner-workings of the solution are provided in Krupka and Hausman (1998) and Aone, et al. (1998).

Knowledge Engineering Cost

Given the above descriptions, it is apparent that the rules are manually crafted and rely on the use of dictionaries and lexicons to extract the entities and learn the relationships. Because of this human intensive process, NetOwl has a high KEC.

Summary Table

Category: Commercial	
Hierarchy: NE	**Source Scope**: Intra and Inter
Company Name: NetOwl (SRA International, Inc.) **Company URL**: http://www.netowl.com/	**Location**: Fairfax, VA, USA
Solution Name: NetOwl Extractor; NetOwl Summarizer; NetOwl TextMiner; NetOwl InstaLink	
Domain Scope: general	**Application Type**: IE and LA
Knowledge Engineering Cost: high	**Financial Cost**: unknown
Input Requirements/Preparation Required: unstructured and structured textual data from over 200	

different document types and 10 languages	
Link Analysis **Algorithm Name/Group**: proprietary **Labeling**: n/a **Labeling Supervision**: n/a **Model Generation**: manual **Model Generation Supervision**: n/a **Process Description**: Manually-crafted rules are used to identify entities and the links between them based on the use of lexicons and pattern rule bases. Then, a template is used to carry out the extraction and link analysis processes. The InstaLink program also allows links to be formed through a GUI.	
Solution Output: XML-marked up texts and translations of foreign values into English	
Application to Law Enforcement: extensive	
Is performance evaluation available? yes	**Solution/demo available**? no

Sources

Aone, Chinatsu; Halverson, Lauren; Hampton, Tom; and Ramos-Santacruz, Mila (1998). *SRA: Description of the IE^2 System Used for MUC-7*. Online. http://www.itl.nist.gov/iaui/894.02/related_projects/muc/proceedings/muc_7_proceedings/sra_muc7.pdf. Accessed January 5, 2006.

Krupka, George R. and Hausman, Kevin (1998). *IsoQuest, Inc: Description of the NetOwl™ Extractor System as Used for MUC-7*. April, 1998. Online. http://www.itl.nist.gov/iaui/894.02/related_projects/muc/proceedings/muc_7_proceedings/isoquest.pdf. Accessed January 5, 2006.

NetOwl. Available: http://www.netowl.com/. Accessed January 5, 2006.

NetOwl (2005a). *NetOwl® Extractor Version 6.* Obtained via email correspondence. Received October 24, 2005.

SRA. *SRA International, Inc.* Available: http://www.sra.com/. Accessed January 5, 2006.

SRA (2000a). "In-Q-Tel Next Generation Intelligence Dissemination System.". *Services and Solutions: Success Stories.* Online. http://www.sra.com/services/index.asp?id=182. Accessed January 5, 2006.

3.3.14 SAS Institute, Inc.

Company Introduction and Domain Scope

SAS Institute was founded in 1976 out of North Carolina State University and is based in Cary, North Carolina. Claiming to be "the world's largest privately held software company" (SAS), SAS has nearly 400 offices worldwide for about 9,800 employees and recorded revenues of $1.53 billion in 2004. Kathleen Khirallah, a senior analyst at the Tower Group, was quoted in Dumiak and Sisk (2004) as stating that the SAS Institute is the "800-pound gorilla in financial services when it comes to analytics" because of its 30-year track record and the fact that its products are used by 90 percent of the Fortune 500. In fact, 96 of the top 100 companies on the FORTUNE Global 500® list are using SAS solutions (SAS, 2004b). SAS works in industries such as energy and utilities, financial services, government and education, healthcare, life sciences, manufacturing, retail, and telecommunications. Some of the company's major clients include Bank of America, Merrill Lynch, Burger King, Kohl's, The Limited, Lowe's Companies, Office Depot, Staples, Wal-Mart, Honda, Ford, Wells Fargo, and the U.S. Census Bureau. Partners include Accenture, IBM, Intel, Sun Microsystems, and Computer

Sciences Corporation. The company also sponsors data mining conferences and events, such as the M2005 conference (M2005, 2005).

The company has also been the recipient of numerous awards. It was recognized by IDC as the number one provider of data warehouse generation tools based on 2004 worldwide revenue and came in second in the data warehouse information access tools category (SAS). It was also highly ranked by *Retail Information Systems News* “for the overall performance, strategic value and ROI that [SAS] delivers to the retail industry through its retail intelligence software” (SAS, 2006). SAS solutions were also KMWorld Trend-Setting Products in 2004 and 2005 and Datamation Products of the Year in 2005 while the company was recognized among KMWorld’s ‘100 Companies that Matter’ in 2005 and Fortune’s 100 Best Companies to Work For (from 1998 – 2005).

SAS solutions provide a wide-range of applicability and the technology encompasses most of the data mining field. As the solutions not only extract data and links from text and link values to present predictive models and insight, SAS provides information extraction and both intra- and intersource link analysis solutions.

Output/Results

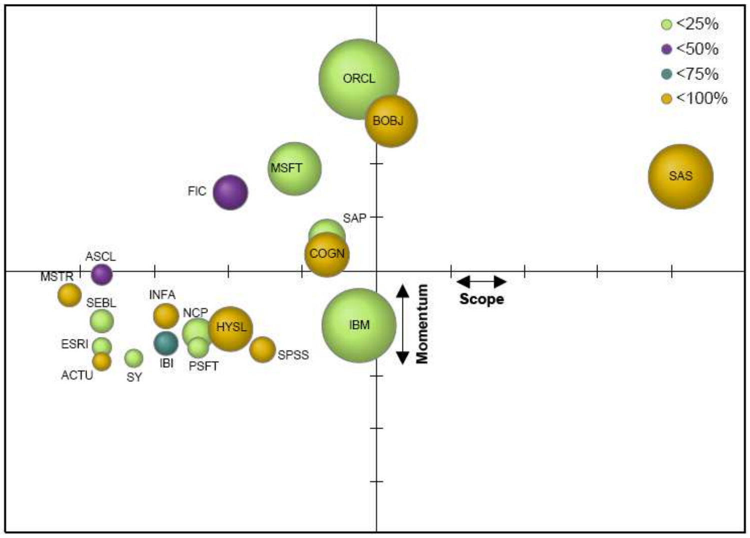
Enterprise Miner primarily produces graphical output. Due to the various different techniques available, the output will be heavily dependent upon the selected algorithm.

Application to Law Enforcement

Extensive. SAS is one of the largest companies with perhaps the most diverse and broad-ranging solutions available. It solutions are used by many leading companies and offer extremely powerful processing and analysis tools. As mentioned in the introduction, the company works across many industries and has numerous clients. For example, Nextel Communications Inc. currently uses SAS’s Enterprise Miner to make predictions based on text captured from call center dialogues and relate key phrases to customer churn (Mitchell, 2005).

Evaluation

IDC Business Analytics Competitive Market Map, 2003



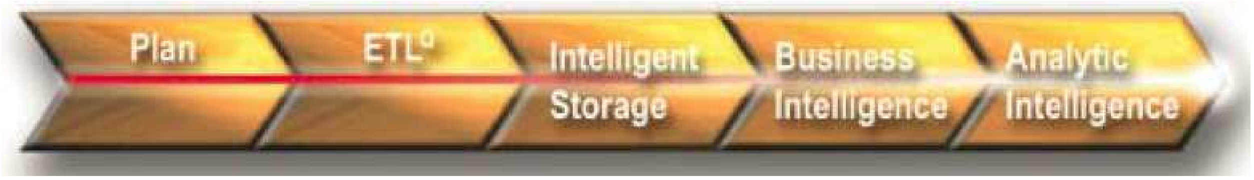
IDC conducted a survey of the business analytics (BA) software market in 2003 (Vesset and Morris, 2004) to evaluate the performance of a variety of companies, of which SAS was included. The companies were ranked on four axes: size (worldwide license and maintenance revenue of BA software), momentum (size-adjusted growth rate), scope (breadth and depth of product offerings as measured in nine categories), and reliance (extent of revenue generated by BA software). SAS was ranked very highly by the survey, coming in as the third largest BA vendor, the fourth highest momentum, and the broadest scope (by far, top three in five of the nine categories; the next closest only ranked top three in two). However, it also mentioned that the company’s reliance on BA revenue was very high (greater than 75%), which would put the company at risk from more diversified software companies. “Strong focus on BA software also puts SAS in the unique position of having a large size, broadest scope and yet being highly-specialized” (Vesset and Morris, 2004). The graph from the study is presented in the figure above.

Financial

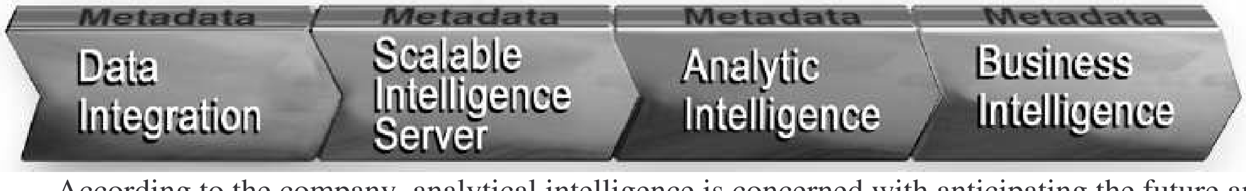
Little detail of the cost of SAS components and solutions is available. However, Charlesworth (2005) states that SAS's Marketing Optimization solution is "typically purchased by companies with in excess of 250,000 customers" and goes on to say that SAS claims "that the solution will ordinarily pay for itself in the first set of campaigns that it is deployed against" as "a typical customer can expect an uplift between 10% and 30%."

Software

SAS provides an immense selection of product offerings. According to the company, data mining is "the process of data selection, exploration and building models using vast data stores to uncover previously unknown patterns" (SAS). SAS uses its *Intelligence Value* Chain, a "framework for delivering high-value, enterprise-wide intelligence" (SAS, 2003a), to provide data mining capabilities to its customers; a diagram of this chain (SAS, 2003a) is presented in the figure above. The *Plan* phase uses roadmaps and industry-specific models, methodologies, and expertise to help develop customized solutions. Users can *e*xtract, *t*ransform, and *l*oad data from various, disparate and heterogeneous platforms and sources for integration into the system during the *ETL*^{Q} phase. According to Bloor Research (2004), this phase "provides data analysis and profiling, data cleansing, and ETL…capabilities based on a shared metadata repository" and through the use of natural language processing techniques. *Intelligence Storage* "efficiently tunes data storage specifically for enterprise intelligence creation and dissemination" (SAS, 2003a), while *Business Intelligence* allows workers to access and maintain the source data for use in various tasks. The final phase, *Analytic Intelligence* provides in-depth intelligence and supports decision making and information dissemination through the uses of predictive and descriptive modeling, forecasting, resource optimization, simulation, experimental design, and other capabilities. The integration of these five steps into a single, cohesive technology framework helps users optimize intelligence environments and align strategic organization objectives (SAS, 2003a).



This chain is implemented through the use of the SAS® Enterprise Intelligence Platform, which is shown in the figure below (SAS, 2005c). Through the use of *Data Integration, Scalable Intelligence Server, Analytic Intelligence*, and *Business Intelligence*, SAS is able to offer its customers a complete data consolidation and mining solution. *SAS Intelligence Platform* includes the SAS Enterprise ETL Server to clean and integrate data in a common data store as well as the SAS Enterprise Business Intelligence Server which allows many users to analyze data and generate reports (SAS).



According to the company, analytical intelligence is concerned with anticipating the future and "calculating the significance of the data to deliver informed inferences about the future and the best action plans to get there" (SAS, 2005d). Analytic intelligence has been further divided into several main capability groupings (with each category having several product offerings): *statistics* (SAS/STAT, SAS/INSIGHT, SAS/IML, SAS/LAB), *data and text mining*, *forecasting* and *econometrics* (SAS High Performance Forecasting, SAS/ETS, SAS/ETS Time Series Forecasting System), *quality improvement* (SAS/QC), and *operations research* (SAS/OR) (SAS). As *data and text mining* is most relevant to this survey, the solutions offered under this category will serve as the focus of this analysis.

SAS®9 is the company's flagship product offering and was released in March, 2004. According to the company's CEO Jim Goodnight, it represents "the most significant release in [the company's] history" as the platform integrates all of SAS's applications and communicates with other data sources and programs (SAS). The solution consists of several main components. The information extraction and link analysis components of the system are grouped into two categories. *SAS Text Miner* is the solution's information extraction component, discovering and extracting knowledge from text documents (SAS); the main applications of this solution include text collection, text processing, and knowledge extraction (SAS, 2002). *SAS Enterprise Miner* (currently version 5.2) provides data mining and link analysis solutions to analyze data through the use of a Java interface. Details of these two solutions are provided in the Algorithm section.

Demo versions of several SAS offerings are available at http://support.sas.com/.

Inputs Required

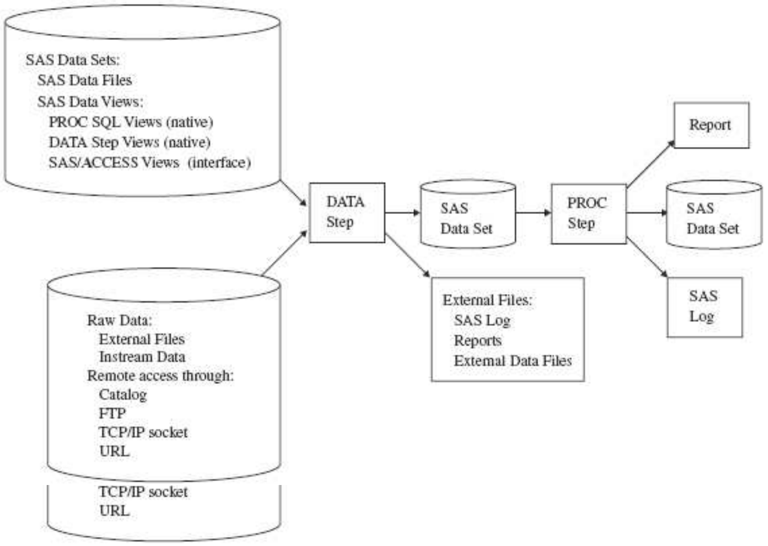
The Text Miner solution "combines a variety of information sources, including text and traditional databases" (SAS, 2005e) and can handle a wide variety of textual data formats, including PDF, extended ASCII, HTML, MS Word, and WordPerfect. Web crawling capabilities are also available. Customized routines and dictionaries are available in Dutch, English, French, German, Italian, Portuguese, and Spanish (SAS, 2005e).

Enterprise Miner can access more than 50 different file structures (SAS, 2005a).

Link Analysis Algorithm

As mentioned in the Software section, SAS's information extraction capabilities are housed within the SAS Text Miner solution and utilize a process known as *SAS processing*. SAS solutions use the *SAS language* to manage data and *SAS procedures* to handle data analysis and reporting. SAS processing has a DATA step to manipulate the data and a PROC step to analyze the data, produce output, or manage SAS files (SAS, 2005f). A high-level diagram of this process is presented in the adjacent figure (SAS, 2005f). Details of the SAS language are beyond the scope of this survey but can be found in (SAS, 2005f) and (SAS, 2005g).

Figure 2.1 SAS Processing



SAS considers text mining to be a three-step process: *accessing the unstructured text*, *parsing the text and turning it into actionable data*, and *analyzing the newly created data* (SAS, 2005e). Through the use of a graphical user interface, users can use automated procedures to extract and analyze the data. Terms and phrases are extracted from the text via rules from English, French, German, and Spanish texts. Stemming, spell correction (transposed letters, embedded spaces, etc.), stop lists, compound word splitting, and part of speech tagging are also performed. Users can specify entities and noun-groups such as abbreviations, country names, and organization names to be extracted from the text through the use of broad customizable data dictionaries (SAS, 2005e). Users can also establish synonym lists. Once the entities have been extracted, they are normalized and included in a matrix table (SAS, 2005e).

Text Miner can also transform parsed documents into numerical representation through the use of Singular Value Decomposition (SVD), rollup terms, or a combination of both. "SVD is a powerful technique for automatically relating similar terms and documents, eliminating an exhaustive need to manually generate specific ontologies or synonym lists…transform[ing] each document into an *n*-dimensional subspace" (SAS, 2005e). Rollup terms, then, also "reduces dimensionality by taking the *n* highest weighted terms and ignoring the rest" (SAS, 2005e).

In addition to information extraction, Text Miner also performs intrasource and intersource link analysis capabilities. User directed concept linking allows users to visualize complex hidden relationships among terms, phrases, and entities (through the use of the Interactive Results Browser) within a single source. Additionally, many text clustering algorithms can group sources based on themes automatically or through user-defined taxonomies (via the Taxonomy Browser), Expectation Maximization Clustering using spatial clustering techniques, hierarchical clustering using Ward's agglomerative method, K-means or SOM/Kohonen clustering, and structured-data profiles (such as age, etc.). Neural networks, memory-based reasoning, regression, and decision trees are also used. (SAS, 2005e).

According to (SAS, 2003b), the Text Miner solution also contains Inxight's LinguistX and ThingFinder solutions. The extent to which SAS utilizes Inxight's technology is not known.

SAS Enterprise Miner continues Text Miner's technology and allows users to apply a wide-variety of link analysis techniques and algorithms to the extracted information. Like Text Miner, Enterprise Miner uses a GUI to aid the user. The solution also supplies scoring code which is used to evaluate the entire model development process. Enterprise Miner can access more than 50 different file structures and is integrated with SAS ETL Studio through SAS Metadata Server. This studio provides for the definition of training tables and the retrieval and deployment of the scoring code. Data partitioning, outlier-filtering, model ensembles, and model comparisons can also be performed.

The techniques and algorithms included in Enterprise Miner include the following:

- Sampling (simple random, stratified, weighted, cluster, systematic, first *N*, rare event sampling)
- Transformations
    - Simple (log, square root, inverse, square, exponential, standardized)
    - Binning (bucketed, quartile, optimal binning for relationship to target)
    - Best power (maximize normality, maximize correlation with target, equalize spread with target levels)
    - User-defined (polynomial and nth degree interaction effects through the use of an editor).
- Data replacement (measures of centrality, distribution-based, tree imputation with surrogates, mid-medium spacing, robust M-estimators, default constant, user-defined)
- Descriptive statistics
    - Univariate statistics and plots (interval variables, class variables, distribution plots)
    - Bivariate statistics and plots (ordered Pearson and Spearman correlation plot, ordered chi-square plot, coefficient of variation plot)
    - Variable selection by logworth
- Clustering (user-defined, PMML score code, etc.)
- Self-organizing maps (Batch SOMs with Nadaraya-Watson or local-linear smoothing, Kohonen networks, etc.)
- Association rule/Market basket analysis
- Web path analysis
- Dimension reduction (variable selection, principle components, time series mining)

- Regression (linear and logistics, Stepwise, polynomials, cross validation, effect hierarchy rules, optimization with Conjugate Gradient, Double Dogleg, Newton-Raphson, etc.)
- Decision Trees (CHAID, C4.5, Prob. Chi-square test, Gini, Entropy, etc.)
- Neural Networks and Autoneural Neural
- Rule induction
- Two-stage modeling
- Memory-based reasoning (k-nearest neighbor, Patented Reduced Dimensionality Tree and Scan)

(SAS, 2005a).

The company also goes into detail the development of predictive models in (SAS, 2005b). Within this paper, SAS describes the five major stages of the model development life cycle: *Determination of the Business Objective, Data Management, Model Development, Model Deployment,* and *Model Management.* It is important to point out that the company also frequently points to its SEEMA (Sample, Explore, Modify, Model, Assess) methodology which "provides a natural workflow for predictive modeling tasks…[which] guides SAS' development process for its suite of analytical modeling solutions" (SAS, 2005b).

Knowledge Engineering Cost

In terms of link analysis, due to the large number of algorithms existent within the Enterprise Miner solution – each of which require various levels of KEC, we have assigned a *medium* knowledge engineering cost to SAS's solutions. After the data has been processed, the algorithms are able to be accessed by the user through the GUI and analyze the data.

Summary Table

Category: Commercial	
Hierarchy: NE	**Source Scope**: Intra and Inter
Company Name: SAS Institute, Inc. **Company URL**: http://www.sas.com/	**Location**: Cary, NC, USA
Solution Name: SAS® 9; *SAS Intelligence Platform* (SAS Enterprise ETL Server, SAS Enterprise Business Intelligence Server); SAS/STAT; SAS/INSIGHT; SAS/IML; SAS/LAB; SAS High Performance Forecasting; SAS/ETS; SAS/ETS Time Series Forecasting System; SAS/QC; SAS/OR; SAS Text Miner; SAS Enterprise Miner	
Domain Scope: general	**Application Type**: IE and LA
Knowledge Engineering Cost: medium	**Financial Cost**: unknown
Input Requirements/Preparation Required: Enterprise Miner: more than 50 different file structures	
Link Analysis **Algorithm Name/Group**: (various) **Labeling**: (various) **Labeling Supervision**: (various) **Model Generation**: (various) **Model Generation Supervision**: (various) **Process Description**: Enterprise Miner allows the use of numerous analysis techniques and algorithms.	
Solution Output: Enterprise Miner primarily produces graphical output. Due to the various different techniques available, the output will be heavily dependent upon the selected algorithm.	
Application to Law Enforcement: extensive	
Is performance evaluation available? yes	**Solution/demo available**? yes

Sources

Bloor Research (2004). *ETL^Q from SAS Institute*. Online. http://www.sas.com/news/analysts/bloor_etl_0404.pdf. Accessed January 13, 2006.

Charlesworth, Ian (2005). *Business Intelligence: Technology Audit – SAS Marketing Optimization*. Butler Technology Audit. June, 2005. Online. http://www.sas.com/reprints/butler_mo_0605.pdf. Accessed January 13, 2006.

Dumiak, Michael and Sisk, Dumiak (2004). "10 Technology Companies to Watch." *Bank Technology News*. August, 2004. Online. http://www.banktechnews.com/article.html?id=20040802NJ1TRC6O. Accessed January 13, 2006.

M2005 (2005). *M2005: Eighth Annual Data Mining Conference*. October 24-25, 2005. Available: http://www.sas.com/events/dmconf/. Accessed January 13, 2006.

Mitchell, Robert L (2005). "Anticipation Game." *ComputerWorld*. June 13, 2005. Online. http://www.computerworld.com/databasetopics/businessintelligence/story/0,10801,102375,00.html. Accessed August 5, 2005.

SAS. SAS Institute, Inc. Available: http://www.sas.com/. Accessed January 13, 2006.

SAS (2001). *Finding the Solution to Data Mining*. Online. http://www.sas.com/ctx/whitepapers/whitepapers_frame.jsp?code=279. Accessed January 13, 2006.

SAS (2002). *Data Mining in Drug Discovery: Uncovering Hidden Opportunities with SAS® Scientific Discovery Solutions and Enterprise Miner™*. Online. http://www.sas.com/ctx/whitepapers/whitepapers_frame.jsp?code=280. Accessed January 13, 2006.

SAS (2003a). *The SAS® Intelligence Value Chain (brochure)*. Online. http://www.sas.com/technologies/architecture/ivcbrochure0303.pdf. Accessed January 16, 2006.

SAS (2003b). *SAS® Text Miner (brochure)*. Online. http://www.sas.com/technologies/analytics/datamining/textminer/brochure.pdf. Accessed January 13, 2006.

SAS (2004a). *Beyond Business Intelligence*. Online. http://www.sas.com/ctx/whitepapers/whitepapers_frame.jsp?code=277. Accessed January 13, 2006.

SAS (2004b). *New SAS® 9 Software Revolutionizes the BI Industry*. March 30, 2004. Online. http://www.sas.com/news/preleases/033004/news9.html. Accessed January 13, 2006.

SAS (2005a). *Enterprise Miner 5.2 Fact Sheet*. Online. http://www.sas.com/technologies/analytics/datamining/miner/factsheet.pdf. Accessed January 13, 2006.

SAS (2005b). *Operationalizing Analytic Intelligence*. Online. http://www.sas.com/ctx/whitepapers/whitepapers_frame.jsp?code=277. Accessed January 13, 2006.

SAS (2005c). *The SAS® Enterprise Intelligence Platform: An Overview*. Online. http://www.sas.com/ctx/whitepapers/whitepapers_frame.jsp?code=235. Accessed January 16, 2006.

SAS (2005d). *The SAS® Enterprise Intelligence Platform: SAS® Analytic Intelligence.* Online. http://www.sas.com/ctx/whitepapers/whitepapers_frame.jsp?code=240. Accessed January 13, 2006.

SAS (2005e). *SAS® Text Miner Fact Sheet.* Online. http://www.sas.com/technologies/analytics/datamining/textminer/factsheet.pdf. Accessed January 13, 2006.

SAS (2005f). *SAS® 9.1.3 Language Reference: Concepts.* 2nd ed. 2005. Online. http://support.sas.com/documentation/onlinedoc/91pdf/sasdoc_913/base_lrconcept_8943.pdf. Accessed January 16, 2006.

SAS (2005g). *SAS® 9.1.3 Language Reference: Dictionary.* 3rd ed. 2005. Online. http://support.sas.com/documentation/onlinedoc/91pdf/sasdoc_913/base_lrdictionary_9200.pdf. Accessed January 16, 2006.

SAS (2006). *Retail Executives Rank SAS High on Overall Performance, Strategic Value, ROI.* January 9, 2006. Online. http://www.sas.com/news/preleases/010906/news1.html. Accessed January 13, 2006.

Stedman, Craig (2004). "SAS Releases Data Analysis Upgrade to Bid in Broaden Use." *ComputerWorld.* March 31, 2004. Online. http://www.computerworld.com/databasetopics/businessintelligence/story/0,10801,91791,00.html?nas=AM-91791. Accessed January 13, 2006.

Vesset, Dan and Morris, Henry D. (2004). *IDC Competitive Market Map – Evaluation of SAS Institute (Excerpt from IDC #30877).* August, 2004. Online. http://www.sas.com/news/analysts/idc_marketmap.pdf. Accessed January 13, 2006.

3.3.15 SPSS, Inc.

Company Introduction and Domain Scope

SPSS, Inc. represents one of the largest solution providers analyzed in this survey. Founded in 1968, the company now supports more than 250,000 customers that are served by over 1,200 employees in 60 countries. Headquartered in Chicago, Illinois, the company serves "virtually every industry, including telecommunications, banking, finance, insurance, healthcare, manufacturing, retail, consumer packaged goods, higher education, government, and market research" (SPSS). Customers include New York University, Lloyds TSB, Atlanta Police Department, Shenandoah Life Insurance, Puma, Canon, GE, Chase-Pitkin Home and Garden, The Gallop Organization, Southwestern Bell, British Telecom and Deloitte & Touche. SPSS partners include major players such as Accenture, HP, IBM, Microsoft, Oracle, PeopleSoft, Sun Microsystems, Sybase, and Teradata.

The company is also the recipient of numerous awards including a "Company to Watch in 2005" from Intelligent Enterprise Magazine and a Frost & Sullivan 2005 Product Innovation Award for its customer relationship management (CRM) analytics (SPSS). SPSS solutions also enjoy widespread use, as demonstrated in two recent polls conducted by KDnuggets, a leading knowledge discovery (KD) information web site. SPSS ranked highest in both the 2004 "text analysis/text mining software" poll and in the 2005 "data mining/analytical tools." In the first poll, the company's LexiQuest solution ranked over twice as high as the second place solution as it was used by 39% of the respondents (KDnuggets, 2005a). The second poll produced similar results; SPSS Clementine and SPSS solutions ranked as the top two solutions and was used by over a quarter of the respondents (KDnuggets, 2005b).

The solutions provided by SPSS allow information extraction and both intrasource and intersource link analysis.

Output/Results

Extracted information is stored within an existing data source, such as a database or data warehouse. Link analysis is primarily done on a modeling and visual basis. However, Clementine *streams* can be published and executed to export relationship data (SPSS, 2002a).

Application to Law Enforcement

Extensive. SPSS technology has been utilized by many law enforcement departments, including the Charlotte-Mecklenburg (North Carolina) Police Department, the Louisiana Commission on Law Enforcement, the Queensland Fire and Rescue Authority (Australia), the Virginia Department of Juvenile Justice, and the West Midlands (UK) Police Department.

SPSS's solutions were also used in Richmond, VA to cut down on crime. According to McCue, "One thing we realized is that the whole field of behavioral profiling of criminal investigative analysis is based on the concept that crime – even the most serious, violent crime – tends to be very homogeneous and predictable" (McKay, 2005). The Richmond, VA Police Department, under the direction of Dr. Colleen McCue, has been implementing many data mining techniques and applications. Working with SPSS and RTI International, the department has used the tools to predict random gunfire occurrences and helped to reduce New Year's Eve 2003 gunfire incidents by 47% over the previous year (Leon, 2005). The text/data mining capabilities also helped to save $15,000 in costs by having 50 fewer officers on duty, reduce citizen complaints by 47%, and increased the number of firearms removed from circulation by 245% (McKay, 2005).

McKay (2005) mentions other specific examples of public service applications such as city-wide information systems (as in Dallas, TX and Philadelphia, PA), Medicaid monitoring systems (New York), and school district information coordination (Broward County, FL).

Evaluation

Little detailed performance results were found. The company claims that Text Mining for Clementine "analyzes approximately one gigabyte of text per hour, with 90 percent or better accuracy" (SPSS) and maintains throughout their literature that their solutions obtain accuracies of 90% or better. SPSS (2002d) also details some benchmarking studies used to calculate the improvements the Server extensions provided to the data analysis; Linear scalability was verified during the tests as it took approximately 69 seconds to process one million records.

LexiQuest Mine is "capable of handling over 250,000 pages of text per hour" (SPSS, 2002c).

Financial

The company provides detailed financial costs for their solution components as well as training costs. GSA and academic pricing variations are available. Commercial prices for these components range from $199 to $7,452, averaging over $1,200 a component (pricing under the GSA schedule range from $164 to $1,235, with an average price of approximately $600). For instance, the SPSS Text Analysis for Surveys version 1.5 sells for $3,000.

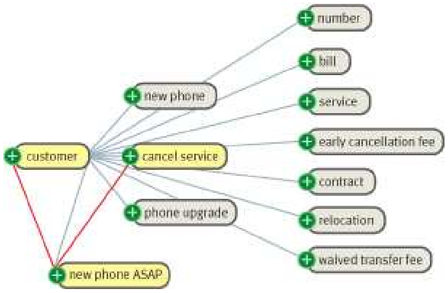
Pricing for Clementine was not available; however, installation of the solution can be performed through the use of a five-day, fixed-price *Clementine Data Mining Jumpstart* which involves the use of consultants to allow the solution to be quickly deployed. Additionally, Charlesworth (2005) reports that "[p]ricing for licenses and implementation depends on the implementation. Annual maintenance and support is 20% of the licensing costs."

Please visit http://www.spss.com/estore/softwaremenu/index.cfm for more pricing information.

Software

The SPSS solution "combines the natural language processing (NLP) linguistic technologies of our LexiQuest text mining products with the advanced data mining capabilities of our data mining workbench, Clementine" (SPSS). The company offers several variations of its solutions. *Text Mining for Clementine* is an open architecture that accesses the textual data and extracts the concepts using NLP technologies. Data mining techniques such as classification, clustering, and predictive modeling uses these concepts in model development. According to the company, the solution is "a text mining product that enables you to extract key concepts, sentiments, and relationships from textual or "unstructured" data and convert them to a structured format that can be used to create predictive models." English, French, German, Italian, Japanese, and Spanish can all be processed and, with the use of the Language Weaver option, Arabic and Chinese sources can also be handled. Specifically, this technology is used in the company's *PredictiveCallCenter*™, *PredictiveClaims*™, and *PredictiveMarketing*™ applications.

WebMining for Clementine includes analysis for web information sources. Based on the company's NetGenesis® technology, it provides open data collection, an *Importer* for processing Web data based on sophisticated rules, an *eDataMart* for storing and organizing data, a *Developer's Kit* for integrating data from other sources and activating e-metrics, and role-based reporting and delivery (SPSS).



LexiQuest Mine visualizes relationships that are contained within large text collections through the use of color-coded association maps, trend charts, and spreadsheet-style reports. A sample screen shot (SPSS) is provided in the adjacent figure. The English, French, and German languages are supported.

LexiQuest Categorize sorts and routes information by organizing large amounts of textual data, such as emails, call center notes, reports, and documents.

SPSS Text Analysis for Surveys analyzes text responses to open-ended survey questions.

Text Mining Builder allows the user "to modify the solution's built-in dictionaries to include terms such as acronyms and synonyms specific to [the] business, industry, or area of research" through the use of an "intuitive interface" (SPSS). The system comes with several pre-built libraries for CRM, genomics, survey, and Homeland Security applications. Spelling variations, words/phrases to ignore, new types (such as negative expressions), and non-linguistic entities (email addresses, currencies) can all be handled, as well. Dutch, English, French, German, Italian, and Spanish dictionaries are editable with this component.

Clementine® is the company's data mining workbench and enables the development of predictive data mining models and deployment of those models into an organization's operations (SPSS). The solution incorporates many link analysis technologies and algorithms, such as decision trees (SPSS, 2001b) (SPSS, 1999) and association rules (SPSS, 2001a). Recently released Clementine Server provides even greater speeds and analysis of larger datasets (SPSS, 2002d).

A complete list of SPSS's solutions is available at http://www.spss.com/products/alpha.cfm?letter=all&source=homepage&hpzone=products. Additionally, a series of online and downloadable demos of various SPSS solutions are available at http://www.spss.com/downloads/Papers.cfm?List=all&Name=all.

Inputs Required

Nearly any textual data format can be handled by the solutions, including HTML, XML, MS Office, PDF, and email. Numerous languages are also supported (see Software section).

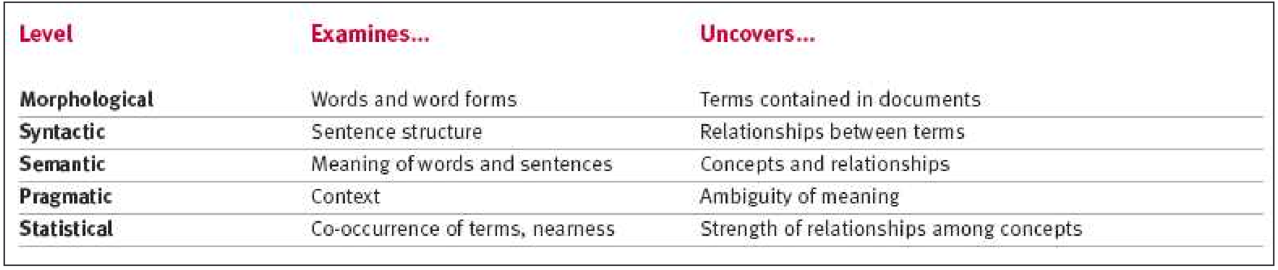
Link Analysis Algorithm

Apparently the majority of the company's information extraction technology is enabled through the use of LexiQuest linguistic extraction technology, which is used to "access and process virtually any type of unstructured data." The LexiQuest Mine solution uses NLP technologies to analyze text "not as a collection of words or letters but as a set of phrases and sentences whose grammatical structure provides a context for the meaning of the document" (SPSS). These processes are carried out through the use of five major components: *Database Manager*, *LexiQuest Mine*, *Database Server*, *LexiQuest Base of Text Mining*, and *Search Engine*. According to a company white paper SPSS (2002c),

> "LexiQuest Mine works by employing a combination of dictionary-based linguistics analysis and statistical proximity matching to identify key concepts, including multi-word concepts. Then, based on a linguistics analysis of the context and semantic nature of the words, it is able to identify their type (organization, product, etc.) as well as the degree of relationship between them and other concepts. These relationships are displayed in a dynamically produced graphical map…which can be used to develop a query based on the connections shown. This query is then run against the document base using Mine's internal search engine. The relevant documents are then returned with the key search concepts highlighted for easy identification within the broader text. Conversely, this query can be sent to a public search engine for further information collection efforts beyond the scope of the existing corpus."

According to Norris (2005), LexiQuest is based on the use of the CLEM expression language to manually generate rules by which to prepare and retrieve the data.

The company's white paper of predictive analysis (SPSS, 2003) defines several types of text mining. A *manual approach* requires people to read through the text. *Automated solutions based on statistics and neural networks* represent another approach, but results in a "fairly low" accuracy due to *noise* (irrelevant results) and *silence* (missed results). *Linguistics-based* solutions offer the best of both worlds; providing "the speed and cost effectiveness of statistics-based systems…" while offering "a far higher degree of accuracy" and "requiring far less human intervention" (SPSS, 2003). In this way, linguistics-based solutions can analyze text at all five different levels, as presented in the chart (SPSS, 2003) below:



Level	Examines…	Uncovers…
Morphological	Words and word forms	Terms contained in documents
Syntactic	Sentence structure	Relationships between terms
Semantic	Meaning of words and sentences	Concepts and relationships
Pragmatic	Context	Ambiguity of meaning
Statistical	Co-occurrence of terms, nearness	Strength of relationships among concepts

The white paper also talks about the six major steps in the extraction process:

1. *Document conversion and language identification* – Sources are first converted to a common format for use in further analysis and the portion of the document to be analyzed is specified. Additionally, the language must be identified. LexiQuest recognizes more than different 80 languages. Internal (static, compiled) and External (user-edited) *dictionaries* (lists of words, relationships, or other information that are used to specify or tune the extraction) are used. These can identify parts of speech as well as domain-specific entities through the use of *LexiQuest Packs.* External dictionaries exist as one of several different types: extraction, synonym, type, keyword, and global.

2. *The identification of candidate terms* – Candidate uni-terms (those not in the general dictionary), candidate multi-terms (containing one or more words), non-linguistic entities (such as phone numbers or dates), and upper-case letter strings (such job titles) are identified.
3. *The identification of equivalence classes among candidate terms and the integration of synonyms* – The terms are then compared and *equivalence classes* (a base form of a phrase, or a single form of two variants of the same phrase) are identified through the use rules. The rules are applied in the following order: user-specified, the most frequent form in the full body of text, and the shortest form in the full body of text (which usually corresponds to the base form).
4. *Type assignment* – Category types are assigned to the extracted components.
5. *Indexing, using a representative term for each equivalence class* – "The document collection is re-indexed by establishing a pointer between a text position and the representative term for each equivalence class" (SPSS, 2003).
6. *Pattern matching and events extraction* – Relationships among the named entities are identified through the use of algorithms provided in LexiQuest Mine and Text Mining for Clementine (SPSS, 2003).

Clementine provides SPSS's core link analysis technologies primarily through the use of modeling and visualization tools. "Clementine is a fully graphical end user tool based on a simple paradigm of selecting and connecting icons from a palette to form what SPSS calls a 'stream'" (Norris, 2005). As part of the analysis, the solution allows the user to select from a variety of algorithms. "By default, Clementine builds predictive models without the user having to specify technical details on how the mining techniques will be implemented" (Norris, 2005). Algorithms included in the solution include:

- Neural Networks: Kohonen Networks, Multi-Layer Perceptions and Radical Basis Function Networks
- Rule Inductions: Decision tree algorithms (C5.0 and C&RT), Quest, Chaid
- Regression Modeling: linear regression, logistic regression
- Clustering: L-Means, TwoStep Clustering
- Association Rule Discovery: Apriori Association Detection and Generalized Rule Induction, and
- Sequence Detection: (time associations).

The company prides itself on its predictive analysis process, which is used to direct, optimize, and automate decisions to improve processes through the use of advanced analytics and decision optimization (SPSS). *Advanced analytics* is used to coordinate and understand the relationships between past, present, and projected future actions through the use of statistical, mathematical, and other techniques such as data and text mining and visualization and reporting. Decision optimization uses engines for scoring, rule generation and application, recommendations, and optimizations to use the analysis and arrive at the best possible conclusion. The adjacent figure (SPSS) provides a visual example of this process.

It is also important to note that the company uses the Cross Industry Standard Process for Data Mining (CRISP-DM) methodology for data mining. Information on this methodology is available at http://www.crisp-dm.org/.

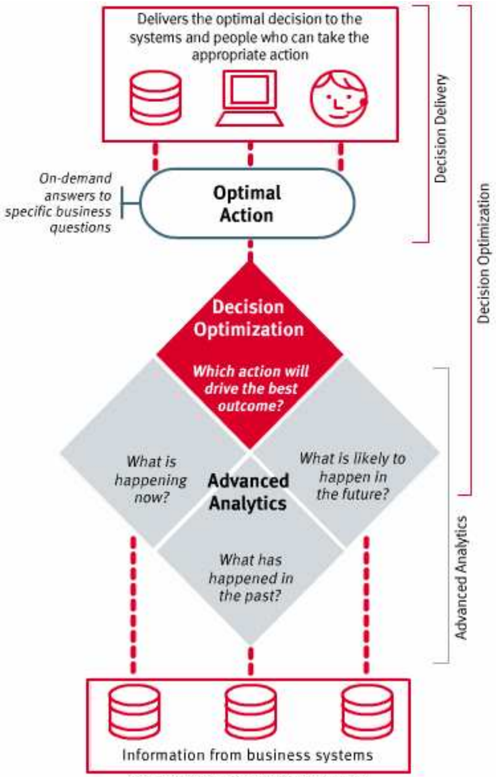

Delivers the optimal decision to the systems and people who can take the appropriate action
Decision Delivery
Decision Optimization
On-demand answers to specific business questions
Optimal Action
Decision Optimization
Which action will drive the best outcome?
What is happening now?
Advanced Analytics
What is likely to happen in the future?
What has happened in the past?
Advanced Analytics
Information from business systems

Figure 2. How predictive analytics works

Knowledge Engineering Cost

In terms of information extraction, SPSS solutions have a high KEC. This is due to the fact that the LexiQuest solution (the foundation of the IE components) is based on the use of the CLEM language for developing manually crafted rules. Additionally, the use of dictionaries and lexicons also substantiate the high KEC. Given the large number of options available to the users of SPSS's solutions, however, the KEC would have to be classified as medium to high, since numerous algorithms and techniques of various complexity and requiring different preparations are utilized.

Summary Table

Category: Commercial	
Hierarchy: NE	**Source Scope**: Intra and Inter
Company Name: SPSS, Inc. **Company URL**: http://www.spss.com/	**Location**: Chicago, IL, USA
Solution Name: Text Mining for Clementine; PredictiveCallCenter™, PredictiveClaims™, and PredictiveMarketing™; WebMining for Clementine (NetGenesis®); LexiQuest Mine; LexiQuest Categorize; SPSS Text Analysis for Surveys; Text Mining Builder; Clementine®	
Domain Scope: general	**Application Type**: IE and LA
Knowledge Engineering Cost: medium	**Financial Cost**: various ($199 - $7,452 for components)
Input Requirements/Preparation Required: textual data	
Link Analysis **Algorithm Name/Group**: various **Labeling**: various **Labeling Supervision**: various **Model Generation**: various **Model Generation Supervision**: various **Process Description**: Link analysis is able to be performed using a variety of algorithms (see Algorithm)	
Solution Output: models and visual representations	
Application to Law Enforcement: extensive	
Is performance evaluation available? no	**Solution/demo available**? yes

Sources

Azoff, Michael. *SPSS Enterprise Platform for Predictive Analysis.* Butler Technology Audit. Online. ftp://hqftp1.spss.com/pub/web/wp/SPSS%20-%20Enterprise%20Platform%20for%20Predictive%20Analytics%20(TA000904BIN).pdf. Accessed January 10, 2006.

Charlesworth, Ian (2005). *Business Intelligence: Technology Audit – SPSS PredictiveClaims Version 1.0.* Butler Technology Audit. July, 2005. Online. ftp://hqftp1.spss.com/pub/web/wp/Butler%20Group%20Audit%20On%20PredictiveClaims.pdf. Accessed January 10, 2006.

KDnuggets (2005b). *Data Mining/Analytic Tools You Used in 2005.* KDnuggets Poll, May, 2005. Online. http://www.kdnuggets.com/polls/2005/data_mining_tools.htm. Accessed January 10, 2006.

KDnuggets (2005a). *Text Analysis/Text Mining Software You Used in 2004.* KDnuggets Poll, January, 2005. Online. http://www.kdnuggets.com/polls/2005/text_mining_tools.htm. Accessed January 4, 2006.

Leon, Mark (2005). "Data Mining Reaps Law Enforcement Rewards." *Database Pipeline*. May 3, 2005. Online. http://www.databasepipeline.com/shared/article/printableArticleSrc.jhtml?articleId=162100971. Accessed June 2, 2005.

McCue, Colleen (2003). "Data Mining and Crime Analysis in the Richmond Police Department." *SPSS Executive Report*. Online. http://www.spss.com/registration/premium/consol056.cfm?WP_ID=132. Accessed July 5, 2005.

McKay, Jim (2005). "Magnifying Data." *Government Technology*. May, 2005 (April 27, 2005). Online. http://www.govtech.net/magazine/story.php?id=93797&issue=5:2005. Accessed June 28, 2005.

Norris, Dave (2005). *Clementine Data Mining Workbench from SPSS.* Bloor Research report. Online. ftp://hqftp1.spss.com/pub/web/wp/Clementine%209%20BloorReport%20LR.pdf. Accessed January 10, 2006.

SPSS. Available http://www.spss.com/. Accessed January 10, 2006.

SPSS (1999). *AnswerTree Algorithm Summary.* Online. ftp://hqftp1.spss.com/pub/web/wp/ATALGWP-0599.pdf. Accessed January 10, 2006.

SPSS (2001a). *The SPSS Association Rules Component.* Online. ftp://hqftp1.spss.com/pub/web/wp/ARCWP-0101.pdf. Accessed January 10, 2006.

SPSS (2001b). *The SPSS C&RT Component.* Online. ftp://hqftp1.spss.com/pub/web/wp/CRTWP-0101.pdf. Accessed January 10, 2006.

SPSS (2002a). *Clementine® Solution Publisher.* SPSS Technical Report. Online. ftp://hqftp1.spss.com/pub/web/wp/CLMP6WP-0301.pdf. Accessed January 10, 2006.

SPSS (2002b). *LexiQuest Categorize.* Online. ftp://hqftp1.spss.com/pub/web/wp/LQCategorizeWP.pdf. Accessed January 10, 2006.

SPSS (2002c). *LexiQuest Mine.* Online. ftp://hqftp1.spss.com/pub/web/wp/LQMineWP.pdf. Accessed January 10, 2006.

SPSS (2002d). *Performance on Large Datasets: Clementine® Server.* Online. ftp://hqftp1.spss.com/pub/web/wp/CLEMPERWP-0802.pdf. Accessed January 10, 2006.

SPSS (2003). *Meeting the Challenge of Text: Making Text Ready for Predictive Analysis.* SPSS White Paper. Online. ftp://hqftp1.spss.com/pub/web/wp/LQWP_NQ.pdf. Accessed July 5, 2005.

4 Conclusion

Building on the work presented in this survey, we will continue our survey utilizing the seven-step process we have laid out for this work in Pottenger and Zanias (2005b). As mentioned in Section 1.2, this status report presents our work up to the present date in our efforts to bring coordination to the intersection of law enforcement and data mining applications. We have completed our preliminary survey results and identified several *axes* or categorizations by which these solutions can be identified.

These categorizations were then used to analyze and organize the solutions identified within the information extraction field, as well as to facilitate our next steps in continuing our research work.

Our work will continue to cover both commercial and academic solutions. We are pleased to have accomplished a preliminary categorization of solutions currently available, as well as those under development. In the coming months, we will continue our efforts to identify metrics and methodologies for evaluating various solutions as well as to develop a repository of ground truth datasets for use in evaluating law enforcement data mining solutions. After accomplishing these goals, our attention will then turn to focusing on the evaluation of representative solutions to continue the evaluation of our seven-step methodology.

As we have already begun through our survey work of the existing solutions and technologies, we are also beginning to understand where the current "cutting edge" of technology exists in the field. In order to incorporate all of our work at the conclusion of this report, this will be one of the focal points of our final report.

5 Future Directions

As mentioned in our proposal paper Pottenger and Zanias (2005b), our final result is to produce a comprehensive report summarizing the solutions categorized, metrics/methodologies identified, datasets developed and future directions identified. In doing this, we hope to accomplish our goal to advance law enforcement data mining research and development and provide law enforcement officials with valuable information and criteria for evaluating current data mining capabilities.

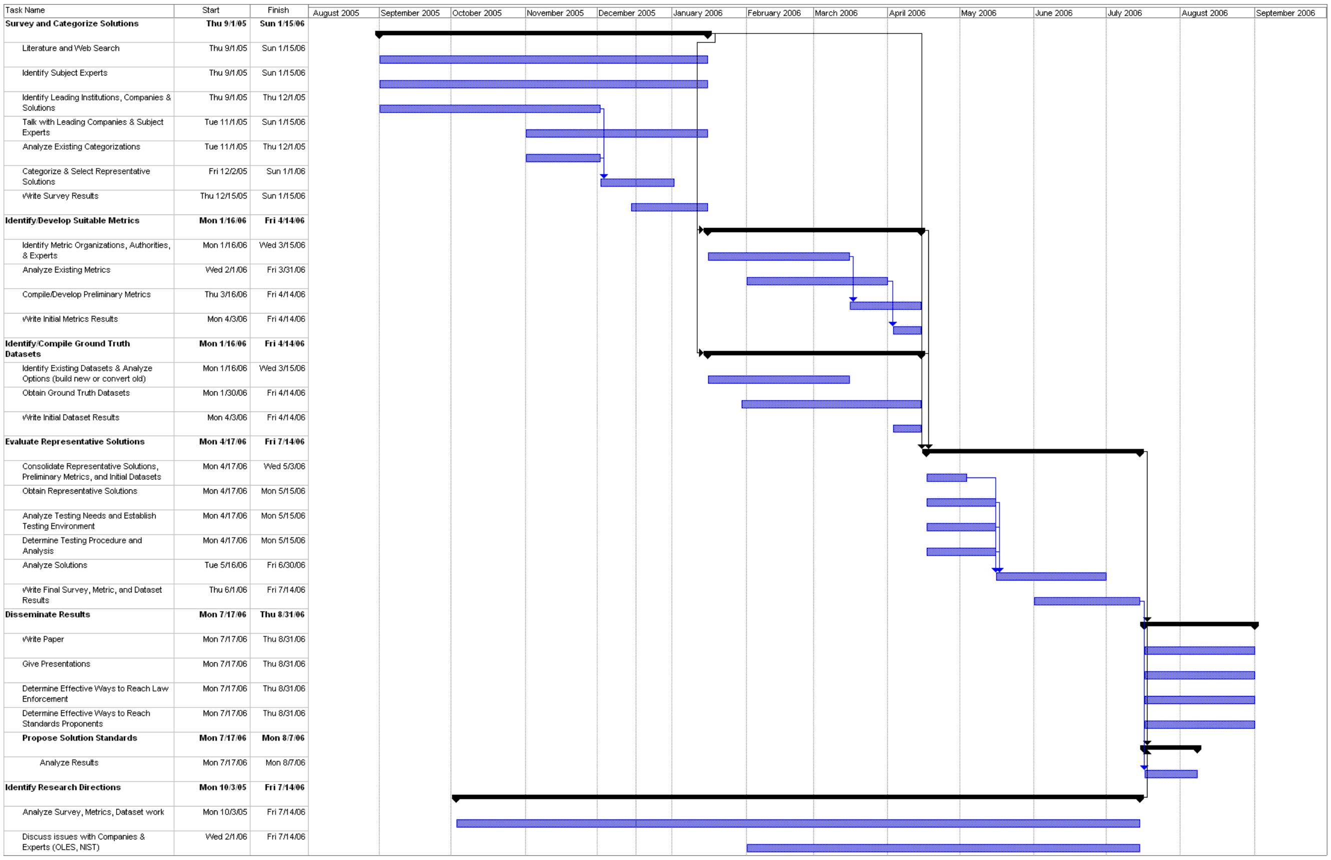
As mentioned in Section 1.2, our survey method calls for the accomplishment of seven steps. To date, we have successfully completed the first, and perhaps most extensive, portion of our survey effort: the survey of the information extraction field and the organization of the solutions into categories. The results of this work have been presented in this report.

Per our original timetable, we have also begun to work on the metric/standards identification and the dataset compilation stages of the project. However, the work required to complete the solution survey has required substantially more time than we had originally anticipated. One of these factors was the time involved in identifying and obtaining information on the various solutions – together with the time to understand and analyze them – was more than we had originally expected. Performing this task was one of the most crucial aspects of the project, as it will provide the information and background for the rest of the process. Therefore, in order to have a better grasp of the field and its technology to be able to perform a more complete analysis, we chose to allow additional time to focus on the survey work.

Another factor was also the difficulty in classifying the solutions. As evidenced throughout this report, the difficultly in categorizing and classifying information extraction technologies is significant. Not only was it necessary to group the solutions into categories, but in order to assess the suitability of various categories we needed to gain more insight into the field. Consequently, there was a great deal of analysis and reanalysis throughout the process. Regardless, we are pleased with our progress, and are looking forward to continuing the survey.

Given this additional time needed to complete the survey work, we have had to revise the project timeline put forth in the proposal document. The revised timeline is presented below, which also represents the modified project timeline to extend from September 1, 2005 until August 31, 2006. In adjusting to the revised start to our timeline, our solution survey work will now constitute one less month on the timeline (although the work was still performed prior to the start date). This will allow us additional time to focus specifically on the metric identification and dataset compilation phases of the project.

As originally specified, our evaluation period will follow this step and will be concluded by our assessment and general evaluation and recommendation phases.



Task Name	Start	Finish
Survey and Categorize Solutions	**Thu 9/1/05**	**Sun 1/15/06**
Literature and Web Search	Thu 9/1/05	Sun 1/15/06
Identify Subject Experts	Thu 9/1/05	Sun 1/15/06
Identify Leading Institutions, Companies & Solutions	Thu 9/1/05	Thu 12/1/05
Talk with Leading Companies & Subject Experts	Tue 11/1/05	Sun 1/15/06
Analyze Existing Categorizations	Tue 11/1/05	Thu 12/1/05
Categorize & Select Representative Solutions	Fri 12/2/05	Sun 1/1/06
Write Survey Results	Thu 12/15/05	Sun 1/15/06
Identify/Develop Suitable Metrics	**Mon 1/16/06**	**Fri 4/14/06**
Identify Metric Organizations, Authorities, & Experts	Mon 1/16/06	Wed 3/15/06
Analyze Existing Metrics	Wed 2/1/06	Fri 3/31/06
Compile/Develop Preliminary Metrics	Thu 3/16/06	Fri 4/14/06
Write Initial Metrics Results	Mon 4/3/06	Fri 4/14/06
Identify/Compile Ground Truth Datasets	**Mon 1/16/06**	**Fri 4/14/06**
Identify Existing Datasets & Analyze Options (build new or convert old)	Mon 1/16/06	Wed 3/15/06
Obtain Ground Truth Datasets	Mon 1/30/06	Fri 4/14/06
Write Initial Dataset Results	Mon 4/3/06	Fri 4/14/06
Evaluate Representative Solutions	**Mon 4/17/06**	**Fri 7/14/06**
Consolidate Representative Solutions, Preliminary Metrics, and Initial Datasets	Mon 4/17/06	Wed 5/3/06
Obtain Representative Solutions	Mon 4/17/06	Mon 5/15/06
Analyze Testing Needs and Establish Testing Environment	Mon 4/17/06	Mon 5/15/06
Determine Testing Procedure and Analysis	Mon 4/17/06	Mon 5/15/06
Analyze Solutions	Tue 5/16/06	Fri 6/30/06
Write Final Survey, Metric, and Dataset Results	Thu 6/1/06	Fri 7/14/06
Disseminate Results	**Mon 7/17/06**	**Thu 8/31/06**
Write Paper	Mon 7/17/06	Thu 8/31/06
Give Presentations	Mon 7/17/06	Thu 8/31/06
Determine Effective Ways to Reach Law Enforcement	Mon 7/17/06	Thu 8/31/06
Determine Effective Ways to Reach Standards Proponents	Mon 7/17/06	Thu 8/31/06
Propose Solution Standards	**Mon 7/17/06**	**Mon 8/7/06**
Analyze Results	Mon 7/17/06	Mon 8/7/06
Identify Research Directions	**Mon 10/3/05**	**Fri 7/14/06**
Analyze Survey, Metrics, Dataset work	Mon 10/3/05	Fri 7/14/06
Discuss issues with Companies & Experts (OLES, NIST)	Wed 2/1/06	Fri 7/14/06

Figure: Project Timeline

5.1 *Next Steps in Survey Process*

Below, we explain in more detail the steps that remain in the completion of our project.

Identify/Develop Suitable Metrics

We have already made great strides towards evaluating and ranking solutions. Currently, the metrics provided in the law enforcement community have primarily been subjective and based on personal opinion. Ranking solutions on a subjective basis, while useful, can lead to problems as one person's standards can be completely different from another's. In order to produce more objective, quantitative rankings, the identification of metrics is required.

Our work presented in Section 2.2 mentioned briefly our work in the establishment of "axes" on which to view these solutions. Recognizing these important criteria is vital to metric development, and our analysis of these issues will continue. Additionally, the feedback and insight that we have been able to obtain from the officers and industry experts has been crucial to identifying these axes and our communication with these individuals will continue over the next several months. Furthermore, we plan to continue to utilize our expertise in the research and computer science fields by focusing on the technical metrics that can be used to evaluate data mining solutions. By keeping in mind the practical requirements of officers, we will be focusing our attention to further develop metrics in the evaluation of Knowledge Engineering Cost (KEC) and other technical research and computer science metrics as well. We have also continued to keep in mind the execution time performance metrics such as throughput, latency, etc. in our study.

Identify/Compile Ground Truth Datasets

Similarly, the need for an authoritative, accurate, encompassing, and anonymized set of law enforcement data is crucial to the success of this project and the advancement of law enforcement data mining solutions. By evaluating the solutions and metrics on a suitable set of data, we can be more confident of the solutions' capabilities to handle the needs of law enforcement applications. It is important to note that, not only will this dataset be used to evaluate the various data mining solutions identified in the survey, but it will also be made available to other researchers and developers in the law enforcement area as a standard data source on which to evaluate their own and other solutions. We still have been unable to discover any such datasets that are designed for the specific purpose of general law enforcement solution testing and evaluation, but are hoping to be able to identify data partners in the coming months as we delve further into this aspect of the project.

Evaluate Representative Solutions, Propose Solution Standards, Identify Research Directions, Dissemination of Survey Results

Our plans for the remaining steps of the process remain the same as proposed in our proposal paper. A minor change exists in the selection of solutions. Due to the difficulty in identifying categories, the solutions which will be evaluated will be chosen based on several factors, rather than on a single categorical metric. As the evaluation stages require the use of metrics and the compiled dataset, our exact process will become more clearly focused as we conclude these two aspects of the project. Throughout our project, we have especially kept in mind the need to disseminate the information to practitioners as well as researchers and are currently looking into developing additional methods to enhance their utilization of the results of this survey.

6 Acknowledgements

This project was supported by grant number 2005-93046-PA-IJ awarded by the National Institute of Justice, Office of Justice Programs, US Department of Justice. Points of view in this document are those of the authors and do not necessarily represent the official position or policies of the US Department of Justice.

We are also grateful for the help of other co-workers, family members and friends. Co-authors Stephen V. Zanias and William M. Pottenger also gratefully acknowledge the continuing help of our Lord and Savior, Yeshua the Messiah (Jesus the Christ) in our lives and work. Amen.

7 References

Adamic, L. and Adar, E. (2003). "Friends and Neighbors on the Web." *Soc. Networks*. 25(3).

AeroText. Available: http://www.aerotext.com/. Accessed August 5, 2005.

AeroText (2003). *AeroText Products: Extracting Intelligence from Text*. May, 2003. Online. http://www.lockheedmartin.com/data/assets/3497.pdf. Accessed January 9, 2006.

Agrawal, R.; Imielinske, T. and Swami, A. (1993). "Mining Association Rules between Sets of Items in Large Databases." *In Proc. of the 1993 ACM SIGMOD Int'l. Conference on Management of Data.* Pages 207-216, Washington, D.C., June 1993. Online. http://citeseer.ist.psu.edu/cache/papers/cs/4475/http:zSzzSzwww.cs.uni-bonn.dezSzIIIzSzlehrezSzvorlesungenzSzDataMiningzSzWS97zSz.zSzWS96zSzliteraturzSzagrawal93:mining.pdf/agrawal93mining.pdf Accessed January 21, 2006.

Ananyan, S. and Kharlamov, A. *Automated Analysis of Natural Language Texts.* Online. http://www.megaputer.com/tech/wp/tm.php3. Accessed January 26, 2006.

Aone, C. and Ramos-Santacruz, M. (2000) “REES: A large-scale Relation and Event Extraction System.” *In Proceedings of the 6th Applied Natural Language Processing Conference*. Online. http://www.cs.mu.oz.au/acl/A/A00/A00-1011.pdf Accessed January 21, 2006.

Aone, Chinatsu; Halverson, Lauren; Hampton, Tom; and Ramos-Santacruz, Mila (1998). *SRA: Description of the IE^2 System Used for MUC-7.* Online. http://www.itl.nist.gov/iaui/894.02/related_projects/muc/proceedings/muc_7_proceedings/sra_muc7.pdf. Accessed January 5, 2006.

Apicella, Mario (2000). “PolyAnalyst 4.1 Digs Through Data for Gold.” *InfoWorld*. June 30, 2000. Online. http://www.infoworld.com/articles/es/xml/00/07/03/000703espoly.html. Accessed January 4, 2006.

Attensity. Available: http://www.attensity.com/ Accessed January 16, 2006.

Attensity (2005a). *Attensity Text Analytics Suite: Overview*. Online. http://www.attensity.com/www/pdf/AttenWorkstation_4_13_05.pdf. Accessed January 26, 2006.

Attensity (2005b). *Natural Language Processing and Text Extraction*, October 2005. Obtained via email correspondence. Received October 21, 2005.

Autonomy. Available: http://www.autonomy.com/. Accessed January 16, 2006.

Autonomy (2003a). *Autonomy Technology White Paper.* 2003. Online.http://www.autonomy.com/downloads/Marketing/Autonomy%20White%20Papers/Autonomy%20Technology%20WP%2020040105.pdf. Accessed January 16, 2006.

Autonomy (2003b). *Performance & Scalability White Paper.* August, 2003. Online. http://www.autonomy.com/downloads/Marketing/Autonomy%20White%20Papers/Performance%20and%20Scalability%20WP%2020050811.pdf. Accessed January 16, 2006.

Autonomy (2003c). *XML White Paper.* Online. http://www.autonomy.com/downloads/Marketing/Autonomy%20White%20Papers/Autonomy%20XML%20WP%2020031003.pdf. Accessed October 10, 2005.

Autonomy (2005a). *Autonomy IDOL Server™ 5 Technical Brief.* Online. http://www.autonomy.com/downloads/Technical%20Briefs/Servers/TB%20IDOL%20server%205%200305.pdf. Accessed October 10, 2005.

Autonomy (2005b) *Document Management Technical Brief.* Online. http://www.autonomy.com/downloads/Technical%20Briefs/Servers/TB%20Document%20Management%20Server%200205.pdf. Accessed October 10, 2005.

Ayres, Jay; Flannick, Jason; Gehrke, Johannes and Yiu, Tomi (2002). “Sequential Pattern Mining Using a Bitmap Representation.” *Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining.* Online. http://www.cs.cornell.edu/johannes/papers/2002/kdd2002-spam.pdf. Accessed January 11, 2006.

Azoff, Michael. *SPSS Enterprise Platform for Predictive Analysis.* Butler Technology Audit. Online. ftp://hqftp1.spss.com/pub/web/wp/SPSS%20-%20Enterprise%20Platform%20for%20Predictive%20Analytics%20(TA000904BIN).pdf. Accessed January 10, 2006.

Ben-Dov, M.; Wu, W. and Feldman, R. (2004). “Improving Knowledge Discovery by Combining Text-Mining and Link Analysis Techniques.” *SIAM Int. Conf. on Data Mining.* Online. http://www.uclic.ucl.ac.uk/paul/research/Moty1.pdf. Accessed January 25, 2006.

Bloor Research (2004). *ETL^{Q} from SAS Institute.* Online. http://www.sas.com/news/analysts/bloor_etl_0404.pdf. Accessed January 13, 2006.

Bock, Geoffrey E. (2002). “Meta Tagging and Text Analysis from ClearForest: Identifying and Organizing Unstructured Content for Dynamic Delivery through Digital Networks.” *Patricia Seybold Group*. Online. http://www.instinct-soft.com/WhatsNew/Research.asp Accessed August 8, 2005.

Brown, Donald E (1998). “The Regional Crime Analysis Program (RECAP): A Framework for Mining Data to Catch Criminals.” *IEEE.* January, 1998. Online. http://vijis.sys.virginia.edu/publication/RECAP.pdf Accessed June 13, 2005.

Charlesworth, Ian (2005). *Business Intelligence: Technology Audit – SAS Marketing Optimization*. Butler Technology Audit. June, 2005. Online. http://www.sas.com/reprints/butler_mo_0605.pdf. Accessed January 13, 2006.

Charlesworth, Ian (2005). *Business Intelligence: Technology Audit – SPSS PredictiveClaims Version 1.0*. Butler Technology Audit. July, 2005. Online. ftp://hqftp1.spss.com/pub/web/wp/Butler%20Group%20Audit%20On%20PredictiveClaims.pdf. Accessed January 10, 2006.

ClearForest. Available: http://www.clearforest.com/ Accessed December 17, 2005.

ClearForest (a). *White Paper - Tagging Textual Data: Why? What? How?* Available: http://www.clearforest.com/WhatsNew/Research.asp Accessed August 8, 2005.

ClearForest (2003). “Endeca and ClearForest Announce Strategic Partnership For Advanced Searching of Unstructured Data” March 31, 2003. Online. http://www.clearforest.com/whatsnew/PRs.asp?year=2003&id=34. Accessed December 2, 2005.

CNNMoney (2006). “Google Gets More Personal.” *CNNMoney.com.* January 12, 2006. Online. http://money.cnn.com/2006/01/12/technology/google_enterprise.reut/index.htm. Accessed January 22, 2006.

Das, Amitabha; Ng, Wee-Keong and Woon, Yew-Kwong (2001). “Rapid Association Rule Mining.” *Proceedings of the Tenth International Conference on Information and Knowledge Management*. October 2001. Online. http://portal.acm.org/citation.cfm?id=502665&coll=Portal&dl=ACM&CFID=61549669&CFTOKEN=14947090 Accessed January 21, 2006.

Delphes. *Delphes Technologies International*. Available: http://www.delphes.com/. Accessed January 23, 2006.

Delphes (2003). *White Paper: Integrated Information System*. Online. http://www.delphes.com/pdf/en/white_paper.pdf. Accessed January 23, 2006.

Delphes (2004a). *Extranet and Internet Solutions.* Online. http://www.delphes.com/pdf/en/internet.pdf. Accessed January 23, 2006.

Delphes (2004b). *Intranet Portal Solutions.* Online. http://www.delphes.com/pdf/en/intranet.pdf. Accessed January 23, 2006.

Delphes (2005). *Data Sheet – Intelligence Knowledge Service.* Online. http://www.delphes.com/pdf/en/datasheet.pdf. Accessed January 23, 2006.

Denwattana, Nuansri and Getta, Janusz R (2001). "A Parameterised Algorithm for Mining Association Rules." *Proceedings of the 12th Australasian Conference on Database Technologies ADC '01*. January, 2001. Online. http://portal.acm.org/citation.cfm?id=545543&coll=Portal&dl=ACM&CFID=61549669&CFTOKEN=14947090 Accessed January 21, 2006.

Di Sciullo, Anna Maria and Fong, Sandiway (2001). "Efficient Parsing for Word Structure". *In the Proceedings of the Sixth Natural Language Processing Pacific Rim Symposium.* November 27-30, 2001. Online. http://www.afnlp.org/nlprs2001/pdf/0034-03.pdf. Accessed January 23, 2006.

DMR (2004). DMReview.com Web Editorial Staff. *LAS Announces New Name Parser.* April, 2004. Online. http://www.las-inc.com/media_coverage/2004/Apr04/04-23-04_DMReview.pdf. Accessed December 27, 2005.

Duffy, Diantry (2004). "What's in a Name?" *briefing, CSO Online.* January, 2004. Online. http://www.csoonline.com/read/010104/briefing_name.html. Accessed December 27, 2005.

Dumiak, Michael and Sisk, Dumiak (2004). "10 Technology Companies to Watch." *Bank Technology News.* August, 2004. Online. http://www.banktechnews.com/article.html?id=20040802NJ1TRC6O. Accessed January 13, 2006.

Eidetica. Available: http://www.eidetica.com/. Accessed January 24, 2006.

Eidetica (a). *Content Matters (Brochure).* Online. http://www.eidetica.com/content/downloads/Eidetica-brochure.pdf. Accessed January 24, 2006.

EMC^2 (2006). *EMC^2 Partners: Delphes Technology International*. Online. http://www.emc.com/partnersalliances/partner_pages/delphes.jsp. Accessed January 23, 2006.

Endeca. Available: http://endeca.com/index.html. Accessed January 4, 2005.

Endeca (2005a). *Endeca InFront® for Online Retail*. Online. http://endeca.com/resources/pdf/Endeca_InFront_Overview.pdf. Accessed January 4, 2005.

Endeca (2005b). *The Endeca Navigation Engine*. Online. http://endeca.com/resources/pdf/Endeca_Technical_Overview.pdf. Accessed October 8, 2005.

Endeca (2005c). *Endeca Product Data Navigator.* Online. http://endeca.com/resources/pdf/ProductDataNavigator_Overview.pdf. Accessed January 4, 2005

Endeca (2005d). *The Endeca ProFind® Platform for Search and Guided Navigation® Solutions.* Online. http://endeca.com/resources/pdf/Endeca_ProFind_Overview.pdf. Accessed October 8, 2005.

Endeca (2005e). *New Search and Discovery for the Federal Government.* Online. http://endeca.com/resources/pdf/Endeca_ProFind_Overview_Govt.pdf. Accessed January 4, 2005.

Endeca (2005f). *Product Data Information Access and Retrieval: The Missing Component of Manufacturers' PLM Strategy: Endeca Business White Paper for Manufacturers.* Online. http://endeca.com/resources/pdf/Endeca_Manufacturing_BusinessWP.pdf. Accessed January 4, 2006.

Entrieva (2003). "Retrieving Information." *KMWorld.* Vol. 12, Issue 8. September, 2003. Online. http://www.kmworld.com/Articles/ReadArticle.aspx?ArticleID=8558. Accessed January 9, 2006.

Feldman, Ronen; Aumann, Yonatan; Libetzon, Yair; Ankori, Kfir; Schler, Jonathan and Rosenfeld, Benjamin. (2001). "A Domain Independent Environment for Creating Information Extraction Modules." *CIKM 2001*. Pages: 586-588. Online. http://www.cs.biu.ac.il/~aumann/papers/IEInvironment.pdf. Accessed November 1, 2005.

Feldman, Ronen; Aumann, Yonatan; Finkelstein-Landau, Michal; Hurvitz, Eyal; Regev, Yizhar; Yaroshevich, Ariel (2002). "A Comparative Study of Information Extraction Strategies." *In Proceedings of the Third international Conference on Computational Linguistics and intelligent Text Processing* (February 17 - 23, 2002). A. F. Gelbukh, Ed. Lecture Notes In Computer Science, vol. 2276. Springer-Verlag, London, 349-359. Online. http://www.springerlink.com/media/d48072xvvj3urngu8g8h/contributions/v/y/f/0/vyf09pl32j4nhkxh.pdf. Accessed December 17, 2005.

Feldman, Susan (2005). "Product Flash: Endeca's Latitude: Easy Access to Business Intelligence." *IDC #32716.* January, 2005. Online. http://endeca.com/resources/pdf/idc_bi.pdf. Accessed January 4, 2006.

FINDER. *Florida Integrated Network for Data Exchange and Retrieval*. Available: http://druid.engr.ucf.edu/datasharing/index.html Accessed November 11, 2005.

Franklin, Daniel (2002). "Data Miners: New Software Connects Key Bits of Data that Once Eluded Teams of Researchers." *Time: Online Edition*. December 23, 2002. Online. http://ai.bpa.arizona.edu/go/intranet/papers/GlobalBusiness.pdf. Accessed June 2, 2005.

Ganiz, Murat C., Pottenger, William M. and Janneck, Christoper D. (2005) "Recent Advances in Literature Based Discovery." *Technical Report*. Online. http://www.cse.lehigh.edu/~billp/pubs/JASISTLBD.pdf Accessed January 10, 2006.

GAO (2004). "Data Mining: Federal Efforts Cover a Wide Range of Uses." *Governmental Accountability Office.* May, 2004. Technical Report Number GAO-04-548. Online. http://www.gao.gov/new.items/d04548.pdf Accessed June 13, 2005.

Gordon, M., Lindsay, R.K., and Fan, W. (2001). "Literature Based Discovery on the World Wide Web." *ACM Transactions on Internet Technology*. 2(4), 261-275. Online.

http://delivery.acm.org/10.1145/610000/604597/p261-gordon.pdf?key1=604597&key2=1479128311&coll=GUIDE&dl=GUIDE&CFID=63393987&CFTOKEN=85662425 Accessed January 21, 2006.

Gorr, Wilpen (2004). "Crime Forecasting: Special Interest Group." *Wharton School, University of Pennsylvania.* December 13, 2004. Online. http://www-marketing.wharton.upenn.edu/forecast/Crime/index.html Accessed July 6, 2005.

Graham-Rowe, Duncan (2004). "Cyber Detective Links Up Crimes." *NewScientist.com.* December 5, 2004. Online. http://www.newscientist.com/article.ns?id=dn6734 Accessed June 2, 2005.

Habegger, B. and Quafafou, M. (2002). "Multi-Pattern Wrappers for Relation Extraction from the Web." *ECAI-02*. Lyon, France. pp. 395-399, July 2002. Online. http://www.grappa.univ-lille3.fr/~habegger/finals/ecai-2002.pdf Accessed January 21, 2006.

Hasegawa, T.; Sekine, S. and Grishman, R.. (2004). "Discovering Relations among Named Entities from Large Corpora." *In Proceedings of the Annual Meeting of Association of Computational Linguistics (ACL).* Online. http://www.cs.nyu.edu/~sekine/papers/acl04-hasegawa.pdf Accessed January 21, 2006.

Haser, Tom and Childs, Lois (2002). "Drug Discovery through Information Extraction Technology." Presentation at *NIH BCIG*. April 18, 2002. Online. http://www.altum.com/bcig/events/seminars/2002_04.pdf and http://www.altum.com/bcig/events/seminars/2002_04.htm. Accessed January 9, 2006.

Hauck, Roslin V.; Atabakhsh, Homa; Ongvasith, Pichai; Gupta, Harch and Chen, Hsinchun (2002). "Using CopLink to Analyze Criminal-Justice Data." *Computer* . Volume 35, Issue 3. March 2002. Pages: 30-37. Online. http://ai.bpa.arizona.edu/go/intranet/papers/CopLinkAnalyzeCriminalData.pdf Accessed January 25, 2006.

Hill, Ryan (2005). *Lockheed Martin Signs NetMap Analytics as Authorized Distributor of AeroText™ Information Extraction Software.* August 3, 2005. Online. http://www.netmapanalytics.com/press/AeroText.pdf. Accessed January 9, 2006.

Hingston, P. (2001). "Using Finite State Automata for Sequence Mining." *In Proceedings of the 25th Australasian Computer Science Conference*. Melbourne, Australia. Pages 105-110. Online. http://crpit.com/confpapers/CRPITV4Hingston1.pdf Accessed January 21, 2006.

Hira, Nadira A. (2005). "25 Breakout Companies 2005." *Fortune*. May 16, 2005. Online. http://www.fortune.com/fortune/breakout/snapshot/0,23871,21,00.html. Accessed August 11, 2005.

InferX. Available: http://www.inferx.com/. Accessed January 12, 2006.

InferX (2004). *Technical Specifications for the InferAgent™ Suite*. Online. http://www.inferx.com/technicalspec.pdf. Accessed October 7, 2005.

InferX (a). *InferX Fact Sheet*. Online. http://www.inferx.com/inferx_facts.pdf. Accessed January 12, 2006.

InferX (b). *A Next Generation Targeting System for Container Security Risk Assessment.* Flash multimedia presentation. Online. http://www.inferx.com/inferxcsra.zip. Accessed January 12, 2006.

Inxight. Available: http://www.inxight.com/. Accessed December 1, 2005.

Inxight (a). *Corporate Fact Sheet.* Available: http://www.inxight.com/pdfs/corp_fact_sheet.pdf. Accessed December 1, 2005.

Inxight (b). *ThingFinder Advanced with Custom Entity Extraction*. Online. http://www.inxight.com/pdfs/Inxight_ThingFinder_Advanced_ds.pdf. Accessed November 1, 2005.

Inxight (2004a). *Inxight SmartDiscovery: Entity Extraction.* Online. http://www.inxight.com/pdfs/EntityExtraction_FinalWeb.pdf. Accessed November 15, 2005.

Inxight (2004b). *Inxight SmartDiscovery: Taxonomy and Categorization.* Online. http://www.inxight.com/pdfs/Taxonomy_FinalWeb.pdf. Accessed November 15, 2005.

Inxight (2005a). *Inxight SmartDiscovery Analysis Adapters and Connectors.* Online. http://www.inxight.com/pdfs/SD_Adapters_Datasheet.pdf. Accessed December 22, 2005.

Inxight (2005b). *Inxight SmartDiscovery Analysis Server.* Online. http://www.inxight.com/pdfs/SmartDiscovery_AS.pdf. Accessed November 15, 2005.

Inxight (2005c). *Inxight SmartDiscovery Awareness Server.* Online. http://www.inxight.com/pdfs/SmartDiscovery_FinalWeb.pdf. Accessed December 22, 2005.

Inxight (2005d). *Inxight SmartDiscovery: Fact Extraction.* Online. http://www.inxight.com/pdfs/FactExtraction_Web.pdf. Accessed November 15, 2005.

Inxight (2005e). *Inxight Software, Inc. Company Fact Sheet.* Online. http://www.inxight.com/pdfs/corp_fact_sheet.pdf. Accessed November 15, 2005.

Jeh, G. and Widom, J. (2002). “SimRank: A Measure of Structural-Context Similarity.” *In KDD 2002.*

Jin, E; Girvan, M. and Newman, M. (2001). “The Structure of Growing Social Networks.” *Phys. Rev. E*. 64(046132).

Kanellos, Michael (2005). “Tech's Part in Preventing Attacks.” *CNet News.com.* July 9, 2005. Online. http://news.com.com/Techs+part+in+preventing+attacks/2100-7348_3-5778470.html Accessed July 11, 2005.

Katz, L. (1953). “A New Status Index Derived from Sociometric Analysis.” *Psychometrika*. 18(1), March 1953.

KCC. *Knowledge Computing Corporation*. Available: http://www.knowledgecc.com/ Accessed June 6, 2005.

KDnuggets (2005a). *Text Analysis/Text Mining Software You Used in 2004.* KDnuggets Poll, January, 2005. Online. http://www.kdnuggets.com/polls/2005/text_mining_tools.htm. Accessed January 4, 2006.

KDnuggets (2005b). *Data Mining/Analytic Tools You Used in 2005.* KDnuggets Poll, May, 2005. Online. http://www.kdnuggets.com/polls/2005/data_mining_tools.htm. Accessed January 10, 2006.

KMWorld. *KMWorld Buyers Guide: Lockheed Martin Corporation.* Online. http://www.kmworld.com/buyersGuide/ReadCompany.aspx?CategoryID=77&CompanyID=17. Accessed January 9, 2006.

Kogut, Paul and Holmes, William. *AeroDAML: Applying Information Extraction to Generate DAML Annotations from Web Pages.* Online. http://semannot2001.aifb.uni-karlsruhe.de/positionpapers/AeroDAML3.pdf. Accessed January 9, 2006.

Kontostathis, A. and Pottenger, W. M. (2006). "A Framework for Understanding LSI Performance." *Information Processing & Management*. Volume 42, Issue 1, Pages 56-73. January 2006. Online. http://www.cse.lehigh.edu/~billp/pubs/IPM.pdf Accessed January 24, 2006.

Krupka, George R. and Hausman, Kevin (1998). *IsoQuest, Inc: Description of the NetOwl™ Extractor System as Used for MUC-7.* April, 1998. Online. http://www.itl.nist.gov/iaui/894.02/related_projects/muc/proceedings/muc_7_proceedings/isoquest.pdf. Accessed January 5, 2006.

Kum, H. C.; Pei, J.; Wang, W. and Duncan, D. (2003). "ApproxMAP: Approximate Mining of Consensus Sequential Patterns." *In International Conference on Data Mining, IEEE*. Online. http://www.siam.org/meetings/sdm03/proceedings/sdm03_36.pdf Accessed January 21, 2006.

LAS. Language Analysis Systems, Inc. Available: http://www.las-inc.com/index.shtml.

LAS (2004). *Advanced Name Recognition Technology: An Overview.* Obtained via email correspondence.

Leon, Mark (2005). "Data Mining Reaps Law Enforcement Rewards." *Database Pipeline.* May 3, 2005.Online. http://www.databasepipeline.com/shared/article/printableArticleSrc.jhtml?articleId=162100971 Accessed June 2, 2005.

Li, S.; Wu, T and Pottenger, W. M. (2005) "Distributed Higher Order Association Rule Mining Using Information Extracted from Textual Data." *SIGKDD Explorations*. Volume 7, Issue 1, June 2005. Online. http://www.cse.lehigh.edu/~billp/pubs/SIGKDDExplorations.pdf Accessed January 21, 2006.

M2005 (2005). *M2005: Eighth Annual Data Mining Conference.* October 24-25, 2005. Available: http://www.sas.com/events/dmconf/. Accessed January 13, 2006.

McCue, Colleen (2003). "Data Mining and Crime Analysis in the Richmond Police Department." *SPSS Executive Report.* Online. http://www.spss.com/registration/premium/consol056.cfm?WP_ID=132. Accessed July 5, 2005.

McKay, Jim (2005). "Magnifying Data." *Government Technology*. May, 2005 (April 27, 2005). Online. http://www.govtech.net/magazine/story.php?id=93797&issue=5:2005. Accessed June 28, 2005.

MDA (2004). Missile Defense Agency, Advanced Systems, Technology Applications Program. "Data Analysis: Datamat Systems Research, Inc./InferX". *2004 MDA Technology Applications Report.* 2004. Online. http://www.inferx.com/MDA_Techreview_2004.pdf. Accessed January 12, 2006.

Megaputer. *Megaputer Intelligence, Inc.* Available: http://www.megaputer.com/ Accessed January 4, 2006.

Megaputer (2002). *X-SellAnalyst*™. Online. http://www.megasysdev.com/down/wm/white_papers/x_sellanalyst.pdf. Accessed October 8, 2005.

Megaputer (2003). *PolyAnalyst for Text: Text Mining System.* Online. http://www.megasysdev.com/down/dm/pa/docs/PolyAnalyst_for_Text_brochure.pdf. Accessed October 8, 2005.

Mena, Jesus (2004). "Homeland Security as Catalyst." *Intelligent Enterprise*. July 1, 2004. Online. http://www.intelligententerprise.com/showArticle.jhtml?articleID=22102265. Accessed June 2, 2005.

Microsystems. *Microsystems, Ltd.* Available: http://www.analyst.ru/ Accessed January 4, 2006.

Mitchell, Robert L (2005). "Anticipation Game." *ComputerWorld*. June 13, 2005. Online. http://www.computerworld.com/databasetopics/businessintelligence/story/0,10801,102375,00.html. Accessed August 5, 2005.

Mnookin, Seth (2003). "Crime: A Google for Cops." *Newsweek*. March 3, 2003. pg. 9.

Mooney, R.; Melville, P.; Tang, L.; Shavlik, J.; Dutra, I.; Page, D. and Costa, V. Santos (2002). **"**Relational Data Mining with Inductive Logic Programming for Link Discovery." *Proceedings of the National Science Foundation Workshop on Next Generation Data Mining*. Baltimore, Maryland. Online: http://citeseer.ist.psu.edu/cache/papers/cs/32831/ftp:zSzzSzftp.cs.wisc.eduzSzmachine-learningzSzshavlik-groupzSzmooney.nsf02.pdf/mooney02relational.pdf Accessed January 10, 2006.

Mordoff, Keith (2004). *Lockheed Martin's NEW AeroText™ Version 4.0 Helps Users Tackle Data Overload, Pinpoint Critical Information.* April 14, 2005. Online. http://www.lockheedmartin.com/data/assets/10586.pdf. Accessed August 9, 2005.

Nahm, Un Yong and Mooney, Raymond J. (2001). "Mining Soft-Matching Rules from Textual Data." *Proceedings of the Seventeenth International Joint Conference on Artificial Intelligence(IJCAI-01).* Seattle,WA. Pages: 979-984. August, 2001. Online. http://www.cs.utexas.edu/users/ml/papers/discotex-ijcai-01.pdf. Accessed January 12, 2006.

NetOwl. Available: http://www.netowl.com/. Accessed January 5, 2006.

NetOwl (2005a). *NetOwl® Extractor Version 6.* Obtained via email correspondence. Received October 24, 2005.

Newman, M. (2001). "Clustering and Preferential Attachment in Growing Networks." *Phys. Rev. E.* 64(025102).

Nieland, Henk (1999). "Eidetica – A New CWI Spin-off Company." *Research and Development, ERCIM News, No. 37.* April, 1999. Online. http://www.ercim.org/publication/Ercim_News/enw37/nieland.html. Accessed January 24, 2006.

NIST (2001). "Definitions of terms used in Information Extraction." *NIST, Information Technology Laboratory, Information Access Division, The Retrieval Group*. January 12, 2001. Online. http://www.itl.nist.gov/iaui/894.02/related_projects/muc/index.html. Accessed December 19, 2005.

NLECTC (1999). "CopLink: Database Detective." *TECHbeat.* Summer, 1999. Online. http://ai.eller.arizona.edu/COPLINK/publications/detective/detective.htm Accessed June 2, 2005.

Noble, David (a). *Fusion of Open Source Information.* Online. http://www.ebrinc.com/files/Noble_Fusion.pdf. Accessed January 9, 2006.

Noble, David (b). *Structuring Open Source Information to Support Intelligence Analysis.* Online. http://www.ebrinc.com/files/Noble_Structuring.pdf. Accessed January 9, 2006.

Norris, Dave (2005). *Clementine Data Mining Workbench from SPSS.* Bloor Research report. Online. ftp://hqftp1.spss.com/pub/web/wp/Clementine%209%20BloorReport%20LR.pdf. Accessed January 10, 2006.

Pei, Jian; Han, Jiawei; Mortazavi-Asl, Behzad; Pinto, Helen; Chen, Qiming; Dayal, Umeshwar and Hsu, Mei-Chun (2001). "PrefixSpan: Mining Sequential Patterns Efficiently by Prefix-Projected Pattern Growth." *In Proc. 2001 Int. Conf. Data Engineering (ICDE'01).* Heidelberg, Germany. Pages 215-224. Online. http://www-sal.cs.uiuc.edu/~hanj/pdf/span01.pdf Accessed January 21, 2006.

Pottenger, W.M. and Zanias, S.V. (2005a) *Free Text Conversion and Semantic Analysis Survey*. August, 2005. NIJ Proposal Number 2005-93045-PA-IJ.

Pottenger, W.M. and Zanias, S.V. (2005b) *Link Analysis Survey*. August, 2005. NIJ Proposal Number 2005-93046-PA-IJ.

Pottenger, William M.; Yang, Xiaoning and Zanias, Stephen V. (2006). *Free Text Conversion and Semantic Analysis Survey Status Update*. January, 2006. NIJ Proposal Number 2005-93045-PA-IJ.

Pratt, W. and Yetisgen-Yildiz, M. (2003). "LitLinker: Capturing Connections across the Biomedical Literature." *Proceedings of the International Conference on Knowledge Capture (K-Cap'03)*. Florida. October, 2003. Online. http://www.ischool.washington.edu/wpratt/Publications/KCap-p032-pratt.pdf Accessed January 21, 2006.

Regev, Y., Finkelstein-Landau, M., and Feldman R. (2002). "Rule-based Extraction of Experimental Evidence in the 15 Biomedical Domain – the KDD Cup 2002 (Task 1)." *SIGKDD Exploration. Newsl.* 4, 2 Dec, 2002, pages: 90-92. Online. http://delivery.acm.org/10.1145/780000/772874/p90-regev.pdf?key1=772874&key2=8532584311&coll=GUIDE&dl=GUIDE&CFID=63236164&CFTOKEN=96493586. Accessed December 17, 2005.

Relue, Richard; Wu, Xindong and Huang Hao (2001). "Efficient Runtime Generation of Association Rules." *Proceedings of the Tenth International Conference on Information and Knowledge Management*. October, 2001. Online. http://delivery.acm.org/10.1145/510000/502664/p466-relue.pdf?key1=502664&key2=5618128311&coll=Portal&dl=ACM&CFID=61549669&CFTOKEN=14947090 Accessed January 21, 2006.

Roberts, Gregory (2003). *AeroText™ Products: Executive Summary Information.* Online. http://www.lockheedmartin.com/data/assets/3504.pdf. Accessed January 9, 2006.

Rosenfeld, Benjamin; Feldman, Ronen; Fresko, Moshe; Schler, Jonathan; and Aumann, Yonatan (2004). "TEG – A Hybrid Approach to Information Extraction." *CIKM'04 Conference (Washington, DC, USA*) November 8-13, 2004, Online. http://delivery.acm.org/10.1145/1040000/1031280/p589-rosenfeld.pdf?key1=1031280&key2=8291408311&coll=GUIDE&dl=GUIDE&CFID=66467799&CFTOKEN=25735454. Accessed January 19, 2006.

SAS. SAS Institute, Inc. Available: http://www.sas.com/. Accessed January 13, 2006.

SAS (2001). *Finding the Solution to Data Mining.* Online. http://www.sas.com/ctx/whitepapers/whitepapers_frame.jsp?code=279. Accessed January 13, 2006.

SAS (2002). *Data Mining in Drug Discovery: Uncovering Hidden Opportunities with SAS® Scientific Discovery Solutions and Enterprise Miner™.* Online. http://www.sas.com/ctx/whitepapers/whitepapers_frame.jsp?code=280. Accessed January 13, 2006.

SAS (2003a). *The SAS® Intelligence Value Chain (brochure).* Online. http://www.sas.com/technologies/architecture/ivcbrochure0303.pdf. Accessed January 16, 2006.

SAS (2003b). *SAS® Text Miner (brochure).* Online. http://www.sas.com/technologies/analytics/datamining/textminer/brochure.pdf. Accessed January 13, 2006.

SAS (2004a). *Beyond Business Intelligence.* Online. http://www.sas.com/ctx/whitepapers/whitepapers_frame.jsp?code=277. Accessed January 13, 2006.

SAS (2004b). *New SAS® 9 Software Revolutionizes the BI Industry.* March 30, 2004. Online. http://www.sas.com/news/preleases/033004/news9.html. Accessed January 13, 2006.

SAS (2005a). *Enterprise Miner 5.2 Fact Sheet.* Online. http://www.sas.com/technologies/analytics/datamining/miner/factsheet.pdf. Accessed January 13, 2006.

SAS (2005b). *Operationalizing Analytic Intelligence.* Online. http://www.sas.com/ctx/whitepapers/whitepapers_frame.jsp?code=277. Accessed January 13, 2006.

SAS (2005c). *The SAS® Enterprise Intelligence Platform: An Overview.* Online. http://www.sas.com/ctx/whitepapers/whitepapers_frame.jsp?code=235. Accessed January 16, 2006.

SAS (2005d). *The SAS® Enterprise Intelligence Platform: SAS® Analytic Intelligence.* Online. http://www.sas.com/ctx/whitepapers/whitepapers_frame.jsp?code=240. Accessed January 13, 2006.

SAS (2005e). *SAS® Text Miner Fact Sheet.* Online. http://www.sas.com/technologies/analytics/datamining/textminer/factsheet.pdf. Accessed January 13, 2006.

SAS (2005f). *SAS® 9.1.3 Language Reference: Concepts.* 2nd ed. 2005. Online. http://support.sas.com/documentation/onlinedoc/91pdf/sasdoc_913/base_lrconcept_8943.pdf. Accessed January 16, 2006.

SAS (2005g). *SAS® 9.1.3 Language Reference: Dictionary.* 3rd ed. 2005. Online. http://support.sas.com/documentation/onlinedoc/91pdf/sasdoc_913/base_lrdictionary_9200.pdf. Accessed January 16, 2006.

SAS (2006). *Retail Executives Rank SAS High on Overall Performance, Strategic Value, ROI.* January 9, 2006. Online. http://www.sas.com/news/preleases/010906/news1.html. Accessed January 13, 2006.

Seifert, Jeffrey W (2004). "Data Mining: An Overview." *Congressional Research Service Order Code RL31798.* December 16, 2004. Online. http://www.fas.org/irp/crs/RL31798.pdf Accessed July 7, 2005.

Shachtman, Noah (2005). "With Terror in Mind, a Formulaic Way to Parse Sentences." *New York Times.* New York, NY. March 3, 2005. Online. http://www.nytimes.com/2005/03/03/technology/circuits/03next.html?ex=1135141200&en=b7e59924788a2cdb&ei=5070. Accessed August 11, 2005.

Siegal, L.G. and Molof, M. J. (1979). *A Handbook For Planning and Performing Criminal Justice Evaluation*. McLean, VA: MITRE Corporation.

Solomon, Jay (2005). "Investing in Intelligence: Spy Agencies Seek Innovation Through Venture-Capital Firm." *The Wall Street Journal* (Eastern edition). pg A.4. September 12, 2005. Online. http://endeca.com/about_endeca/news/n_091205_wsj.html Accessed January 4, 2005.

SPSS. Available http://www.spss.com/. Accessed January 10, 2006.

SPSS (1999). *AnswerTree Algorithm Summary.* Online. ftp://hqftp1.spss.com/pub/web/wp/ATALGWP-0599.pdf. Accessed January 10, 2006.

SPSS (2001a). *The SPSS Association Rules Component.* Online. ftp://hqftp1.spss.com/pub/web/wp/ARCWP-0101.pdf. Accessed January 10, 2006.

SPSS (2001b). *The SPSS C&RT Component.* Online. ftp://hqftp1.spss.com/pub/web/wp/CRTWP-0101.pdf. Accessed January 10, 2006.

SPSS (2002a). *Clementine® Solution Publisher.* SPSS Technical Report. Online. ftp://hqftp1.spss.com/pub/web/wp/CLMP6WP-0301.pdf. Accessed January 10, 2006.

SPSS (2002b). *LexiQuest Categorize.* Online. ftp://hqftp1.spss.com/pub/web/wp/LQCategorizeWP.pdf. Accessed January 10, 2006.

SPSS (2002c). *LexiQuest Mine.* Online. ftp://hqftp1.spss.com/pub/web/wp/LQMineWP.pdf. Accessed January 10, 2006.

SPSS (2002d). *Performance on Large Datasets: Clementine® Server.* Online. ftp://hqftp1.spss.com/pub/web/wp/CLEMPERWP-0802.pdf. Accessed January 10, 2006.

SPSS (2003). *Meeting the Challenge of Text: Making Text Ready for Predictive Analysis.* SPSS White Paper. Online. ftp://hqftp1.spss.com/pub/web/wp/LQWP_NQ.pdf. Accessed July 5, 2005.

SRA. *SRA International, Inc*. Available: http://www.sra.com/. Accessed January 5, 2006.

SRA (2000a). "In-Q-Tel Next Generation Intelligence Dissemination System.". *Services and Solutions: Success Stories.* Online. http://www.sra.com/services/index.asp?id=182. Accessed January 5, 2006.

Srinivasan, P. (2004). "Text Mining: Generating Hypotheses from MEDLINE." *Journal of the American Society for Information Science and Technology*. 55(5), 396-413. Online. http://mingo.info-science.uiowa.edu/padmini/jasist03.pdf Accessed January 21, 2006.

Stedman, Craig (2004). "SAS Releases Data Analysis Upgrade to Bid in Broaden Use." *ComputerWorld.* March 31, 2004. Online. http://www.computerworld.com/databasetopics/businessintelligence/story/0,10801,91791,00.html?nas=AM-91791. Accessed January 13, 2006.

Sundaresan, Neel and Yi, Jeonghee Yi (2000)**. "**Mining the Web for Relations." *Proceedings of the 9th International World Wide Web Conference on Computer Networks: the International Journal of Computer and Telecommunications Networking***.*** Amsterdam, The Netherlands, pages: 699-711 Online. http://www9.org/w9cdrom/363/363.html Accessed January 21, 2006.

Surdeanu, Mihai; Harabagiu, Sanda; Williams, John and Aarseth, Paul (2003). "Using Predicate-Argument Structures for Information Extraction." *In Proceedings of the 41st Annual Conference of the Association for Computational Linguistics (ACL-03).* Pages 8–15. Online. http://acl.ldc.upenn.edu/P/P03/P03-1002.pdf Accessed January 21, 2006.

Swanson, D.R. (1988). Migraine and magnesium: eleven neglected connections. Perspectives in Biology and Medicine, 31(4), 526-557.

Taylor, Sarah M. (2004). "Information Extraction Tools: Deciphering Human Language." *IT Professional.* Vol. 06, no. 6, pages: 28-34. November/December, 2004. Online. http://ieeexplore.ieee.org/iel5/6294/30282/01390870.pdf?tp=&arnumber=1390870&isnumber=30282. Accessed January 9, 2006.

Tumasonis, Romanas and Dzemyda, Gintautas (2004). "A Probabilistic Algorithm for Mining Frequent Sequences." *Eighth East-European Conference on Advances in Databases and Information Systems*. Budapest, Hungary. Online. http://www.sztaki.hu/conferences/ADBIS/8-Tumasonis.pdf Accessed January 21, 2006.

Van der Eijk, C.; Van Mulligen, E.; Kors, J.A.; Mons, B. and Van den Berg, J. (2004). "Constructing an Associative Concept Space for Literature-Based Discovery." *Journal of the American Society for Information Science and Technology*. 55(5), 436-444.

van Zuylen, Catherine (2004). *Inxight: From Documents to Information: A New Model for Information Retrieval.* October, 2004. Online. http://www.inxight.com/pdfs/InxightInformationRetrieval.pdf. Accessed November 28, 2005.

Vesset, Dan and Morris, Henry D. (2004). *IDC Competitive Market Map – Evaluation of SAS Institute (Excerpt from IDC #30877).* August, 2004. Online. http://www.sas.com/news/analysts/idc_marketmap.pdf. Accessed January 13, 2006.

Weeber, M.; Vos, R.; Klein, H.and de Jong-van den Berg, L.T.W. (2001). "Using Concepts in Literature-Based Discovery: Simulating Swanson's Raynaud– Fish Oil and Migraine–Magnesium Discoveries." *Journal of the American Society for Information Science and Technology*. 52(7), 548-557. Online. http://www.inf.ed.ac.uk/teaching/courses/tts/papers/weeber.pdf Accessed January 21, 2006.

Williams, Al (2003). "InferX Corporation: An Innovative Approach to Turning Distributed Data into Decision-Relevant Knowledge." *Presentation at the NewTECH Showcase: Decision Support Tools for the Virginia Center for Innovative Technology*. August 19, 2003. Online. http://www.cit.org/pdf/events/08-19-03-inferx.pdf. Accessed January 12, 2006.

Williams, Kemp and Patman, Frankie (2005). *Personal Entity Extraction Filtering Using Name Data Stores.* 2005. Online. https://analysis.mitre.org/proceedings/Final_Papers_Files/33_Camera_Ready_Paper.pdf. Accessed January 12, 2006.

Witten, Ian H. and Frank, Eibe (2000). *Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations.* New York, NY. Morgan Kaufmann Publishers.

Wu, T. and Pottenger, W. M. (2005a). "A Semi-Supervised Active Learning Algorithm for Information Extraction from Textual Data." *Journal of the American Society for Information Science and Technology*. JASIST, Volume 56, Number 3, Pages: 258-271. Online. http://www.cse.lehigh.edu/~billp/pubs/JASISTArticle.pdf. Accessed September 1, 2005.

Wu, Tianhao and Pottenger, William M. (2005b). "A Very Brief Comparison of AeroText with Lehigh University's Approach to Information Extraction." Private communication from authors received on August 15, 2005.

Xu, J. and Chen, H. (2004). "Fighting Organized Crimes: Using Shortest-Path Algorithms to Identify Associations in Criminal Networks". *Decision Support Systems.* Volume 38, pages 473-487. Online. http://ai.bpa.arizona.edu/go/intranet/papers/Printed.pdf. Accessed January 4, 2006.

Xu, Xiaowei; Mete, Mutlu and Yuruk, Nurcan. "Mining Concept Associations for Knowledge Discovery in Large Textual Databases." *Proceedings of the 2005 ACM Symposium on Applied computing*. March, 2005. Online. http://portal.acm.org/citation.cfm?id=1066802&coll=Portal&dl=ACM&CFID=61549669&CFTOKEN=14947090 Accessed January 21, 2006.

Zelenko, Dmitry and Aone, Chinatsu (2002). "Kernel Methods for Relation Extraction." *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Philadelphia, July 2002, pp. 71-78. Association for Computational Linguistics. Online. http://citeseer.ist.psu.edu/

cache/papers/cs/26987/http:zSzzSzwww.ai.mit.eduzSzpeoplezSzjimmylinzSzpaperszSzZelenko02.pdf/zelenko02kernel.pdf Accessed January 21, 2006.

Zhua, J.; Goncalves, A.; Uren, V.; Motta, E. and Pachecob, R.. (2005). "CORDER: COmmunity Relation Discovery by Named Entity Recognition." *K-CAP'05.* October 2–5, 2005. Banff, Alberta, Canada. Online. http://kmi.open.ac.uk/people/jianhan/s32-zhu.pdf Accessed January 21, 2006.