Document Title:	Free Text Conversion and Semantic Analysis Survey
Author(s):	William M. Pottenger, Ph.D. ; Xiaoning Yang ; Stephen V. Zanias
Document No.:	219551
Date Received:	August 2007
Award Number:	2005-IJ-CX-K005

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.

Free Text Conversion and Semantic 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 information extraction. The purpose of this survey is to identify the leading information extraction capabilities 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 information extraction field and its terms, axes, capabilities, and algorithms, an in-depth analysis of 24 different solutions is provided. The remainder of the report discusses the next steps of the survey effort.

Key words: information extraction, named entity extraction, text analytics, 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.

Information extraction or "free text conversion and semantic analysis" can be considered one of the most vital components of the data mining routine as it forms the basis for knowledge derived from textual data sources. After the data has been cleaned and prepared, information extraction performs the data selection and data transformation steps to prepare the data for mining and evaluation. It is at this crucial step of the process that information extraction will occur, transforming ordinary textual data into actionable, categorized, and recognizable information.

This type of technology is precisely what is needed in law enforcement. With the vast quantities of textual data currently available, it is of critical importance to organize the existing data to *learn* the important *clues* and *leads* that can facilitate the pursuit of justice. According to the 451 Group, a technology research firm, only about 20 percent of all of the information in business or public service is stored within databases (Shachtman, 2005), meaning that for every single database record, there are more than three documents that contain under-utilized or unutilized information – information that may be critical to saving lives. According to a Gartner study, it is estimated that unstructured information is doubling in quantity every three months (Autonomy, 2005b).

A specific type of information extraction that can provide some of the greatest leads to law enforcement officers is a technology known as *named entity extraction (NEE).* This technology not only allows the important the clues and leads to be discovered and identified, but it also categorizes and groups the data to be more efficiently and effectively used. By placing each "nugget" of information in its proper context, similar clues and leads can be joined together to create a fuller and more complete picture of the situations in which officers find themselves engaged. Therefore, in order to provide the most meaningful survey possible, we have focused specifically on *named entity extraction (NEE)*, as it provides this extremely important set of data. While focusing specifically on NEE solutions limits the range of information extraction solutions, a more complete picture and more in depth survey can be conducted. It is important to note at this point that in order to avoid confusion and for simplification throughout this report, when we refer to *information extraction* in this survey, we are referring specifically to named entity extraction, where both the value and the type to which it belongs are extracted from the textual data. We will discuss this further in Section 2 below.

This paper presents our survey findings of the named entity extraction 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 (Pottenger and Zanias, 2005a), the government has taken great interest in the field of data mining in general, including information extraction. From May 2003 through April 2004, the GAO conducted a comprehensive survey of the analysis 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), reported that over 40% (52 of 128) of the federal departments were either using or planning to use data mining capabilities. The 199 efforts identified in the survey were categorized into various groupings, with criminal analysis and detection forming two of the top six categories.

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. We have found a great number of solutions developed at these levels, which only adds to the evidence of the growing need for information extraction tools. 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, information extraction is proving to be a vital technology to address many of the issues currently facing law enforcement. As previously mentioned there are vast quantities of textual data currently available, and information extraction makes it possible to organize this data to *learn* the important *clues* and *leads* that can facilitate the pursuit of justice.

There are a number of information extraction *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 information extraction task and to not limit ourselves to a particular type of product.) As much of this information exists in textual form, one way to confront this issue is to utilize information extraction technologies in order to transform this data into named entities that can easily be transformed into structured, searchable data. Link analysis solutions, on the other hand, allow this information (structured or otherwise) to be joined together to transform textual data into actionable information. Our survey work in this latter domain (link analysis) is presented in (Pottenger et al., 2006b).

There are many examples of solutions used in law enforcement that have produced impressive results. In addition to crime mapping tools ((Brown, 1998) and (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 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). The success experienced by the RPD was enabled through the use of free text conversion and semantic analysis – information extraction. As, according to Dr. McCue, "Data mining is 80 percent preparation and 20 percent analytics" (Leon, 2005), the department had a nearly impossible task before them to manually prepare all of their data; instead, using LexiQuest, text mining software available from SPSS, they were able to carry out their conversion and storage to data from hundreds of thousands of narrative reports and records (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.

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).

Many more examples of information extraction technologies and capabilities are surveyed in detail in Section 3.

1.2 Survey Method

As described in our proposal (Pottenger and Zanias, 2005a), there is a great need to understand the information extraction solutions that are currently available. These solutions have particular importance to the law enforcement community, as they transform 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 meaningful means of evaluating these types of tools. This includes not only the identification and development of meaningful metrics and compilation of representative datasets, but it also involves evaluating a seven step process for evaluation developed in Pottenger and Zanias (2005a) in order to determine whether the seven step process itself is an efficient means of carrying out this type of evaluation survey work.

It is our hope to bring coordination and organization to the intersection of law enforcement and data mining applications, specifically information extraction and link analysis. By identifying not only metrics and methodologies for evaluation but also cataloging numerous solutions and leading edge technologies, these solutions and technologies can be evaluated based on the metrics identified. 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 information extraction 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 the date of writing, 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 both academic and commercial. We have also consulted with law enforcement personnel to learn more about their needs and requirements as well. These categorizations have proved to be difficult, but it is our hope that utilizing these axes will aid in the development of more efficient and meaningful metrics. These axes, the survey results 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 importantly – 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 information extraction. In addition to defining terms that are used throughout the survey, we also present a set of information extraction categories that give insight into their capabilities. We also highlight the prominent algorithms in use in information extraction 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), followed by 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 set of 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 information extraction technology, and, therefore, these sections are primarily directed towards the law enforcement practitioner. The latter part of the solution summary (Information Extraction 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 Information Extraction Overview

2.1 Information Extraction 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 information extraction field. Often, the terms used in this field provide various nuances in meaning. In order to avoid confusion with the terms used in this survey, in this section we provide an definition for each of the terms used extensively throughout this report.

Obviously, named entities are vital to 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). In order to avoid confusion, a simple example is provided to illustrate the concept. If "Albert Einstein" was read from a given medium, it may seem obvious to the human reader that this refers to the famed scientist. Therefore, the pair <person, "Albert Einstein"> represents a named entity because the value ("Albert Einstein") is recognized as belonging to a particular category ("person").

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. As mentioned earlier, our research here specifically focuses on a portion of the Information Extraction (IE) space that is known as *named entity extraction* (NEE), which results in the creation of named entities as the output of the IE process.

This is best illustrated with another example. For instance, several different named entities can be extracted from the following sentence:

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

- <person> Albert Einstein </person>
- <date> March 14, 1879 </date>
- <country> Germany </country>
- <city> Ulm, Germany </city>

Obtaining such named entities from a textual data source is what we refer to as a *named entity (NE).* Because the process has not only identified a value, but has also assigned a type to it, a NE has been extracted.

Continuing with the above example, the following relationships can also be learned:

- <birth_date> Albert Einstein; March 14, 1879 </birth_date>
- <birth_location> Albert Einstein; Ulm, Germany </birth_location>

For the scope of this survey in information extraction, we are not concerned with relationships of this type since we have chosen to categorize this type of extraction as a form of *link analysis* (LA), and have prepared a separate survey of such technologies (Pottenger et al., 2006b). Although this distinction is not standard, we found it necessary not only due to the imprecise use of the term "information extraction" in the field, but also because the extraction of relationships has much in common with the link analysis solutions surveyed in Pottenger et al. (2006b).

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 Information Extraction Axes

As with nearly every issue, there are multiple vantage points from which to classify, organize, and divide. The field of information extraction 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 link analysis axes (Pottenger et al., 2006b), information extraction presents its own unique set of axes.

One possible way to separate information extraction technologies is based on the *level of structure* of the data to be extracted. As information extraction is heavily dependent upon the way in which the text is organized, such a categorization would provide a great deal of information. Knowing how the information was organized would give insight into the type of information extraction algorithm as the different inputs could potentially cause the methods to be carried out by different means. For instance, the approach for extracting information from a database might differ from an algorithm that would extract named entities from narrative textual data. However, such an axis would also be more qualitative in nature since determining the type of data structure would involve non-discrete comparisons or categorizations. For instance, exactly how much more organized is a database table than a textual list or a narrative summary? Also, while it is obvious that some data is more structured than other data, quantifying this difference is no simple task.

Another possible way to divide information extraction 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 extracts every capitalized word and assigns it to type "person" would not be an example of a "sophisticated" technology. But how should complexity be measured? Similarly, practicality is an abstract concept and almost entirely dependent upon the context in which the solution might be used.

Another axis could be the *domain knowledge* required to use a particular IE solution. Many information extraction solutions are tailored for specific disciplines, such as chemical formulas or names. While identifying the applicable disciplines of a given technology, insight is gained into the *flexibility* of the algorithm as well as its *applicability*. If a solution can be used in several domains (perhaps both chemistry and law enforcement), it could represent a better technology. However, this does not take into account how well the algorithm performs in the different application domains, nor does this metric lend itself to quantitative measurement. In addition, with numerous disciplines and application domains from which information can be extracted, the range of values when using such axes would be difficult to enumerate.

A final axis is to divide on the *technique* or *algorithm* used in the solution itself, asking the question, "How is the named entity discovered and extracted?" Although using this axis could result in nearly as many algorithm categories as solutions (as each institution could have its own "proprietary" algorithm), dividing the solutions into *algorithm types* could prove to be useful. While such a categorization would likely be somewhat qualitative in nature, by analyzing the general approach taken by a particular solution, knowledge can be gained not only into the process employed in the solution, but also basic information about the domain knowledge, sophistication, and even the practicality of using a given solution. However, determining how many and exactly what the categories are is a complex task.

In this survey we have chosen to represent several of these axes through a combination of "attributes." The attributes not only help to analyze the solutions, but also give insight into the algorithms employed in the processes. The attributes are presented in the solution summaries in Section 3 and appear as fields in the summary table provided with each solution analyzed. In an effort to provide further categorization, we have qualitatively created nominal values associated with each of these attributes – we 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.

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

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.

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.

2.3 Common Information Extraction Algorithms

This section contains an overview of some common algorithms used in information extraction. These algorithms can be divided into two general categories: *knowledge based* and *machine learning* techniques. Machine learning techniques discussed include Hidden Markov Models (HMM), covering algorithms, and Support Vector Machines (SVM).

2.3.1 Knowledge-based

Knowledge based methods are different from methods which use machine learning algorithms as they rely on special domain knowledge. In order to obtain this knowledge, a great deal of human effort is required. There are several ways by which this can be accomplished in information extraction. One way would be to use a predefined dictionary, where tokens are matched to their type through dictionary lookups. Another way could have the domain expert manually craft patterns/rules to extract entities. For example, the pattern/rule to extract a calendar date could be mm/dd/yyyy. Such patterns can of course also be learned by learning algorithms, which require a lower KEC. Callan and Mitamura (2002) and Toral (2005) are two solutions which are completely knowledge-based.

There are other solutions which are a hybrid of knowledge-based and learning algorithms. An example of this type of solution would be one which uses learning algorithms to extract rules/patterns, but it also uses a predefined dictionary to help identify certain types of entities.

2.3.2 Hidden Markov Model (HMM)

A Hidden Markov Model (HMM) is a probabilistic finite state automaton comprised of a set of unobserved (hidden) states, a finite dictionary of discrete output symbols, and edges denoting transitions from one state to another. Each edge is associated with a transition probability value, and each state emits one symbol in the dictionary from a probability distribution for that state. As with all finite state automatons, HMMs have both a starting state and an end state. From the starting state, a HMM generates an output sequence. "Beginning from the start state, a HMM generates an output sequence $O = o_1, o_2 \ldots, o_k$ by making k transitions from one state to the next until the end state is reached. The i^{th} symbol o_i is generated by the i^{th} state based on that state's probability distribution of the dictionary symbols" (Borkar, 2001). There may be more than one path to generate a given sequence, each path with a different probability. The sum of these probabilities is the total probability of generating the output sequence. The HMM thus induces a probability distribution on sequences of symbols chosen from a discrete dictionary. The training data is employed to learn this distribution. During testing, the trained HMM can be used to determine the most likely sequence of hidden states to have emitted the observed sequence of symbols. The HMM is a very common technique used in information extraction; for example, Borkar (2001), Zhou and Su (2002), and Churches et al. (2002) use HMMs in their solutions.

2.3.3 Covering algorithms

A covering algorithm is yet another machine learning approach. The approach is to take each class in turn and discover rules that cover all instances in the class. At the same time, instances not in this class are excluded. It is called *covering* since at each stage a rule is identified which "covers" some of the instances. The output of a covering algorithm is a set of rules. While constructing rules, conditions are added to the rule if the condition improves the rule's accuracy. Wu and Pottenger (2003) and Wu and Pottenger (2005a) are examples of solutions using covering algorithms.

2.3.4 Support Vector Machines (SVM)

A support vector machine is a learning algorithm that can perform binary classification and regression estimation tasks. It maps the input vectors into a high-dimensional feature space using a nonlinear mapping and constructs an optimal hyperplane in the feature space (Vapnik, 1995). In the simplest case training an SVM involves discovery of a hyperplane that separates the positive training samples from the negative training samples by the largest possible margin. This hyperplane is then used to classify previously unseen samples, which are represented as vectors. The vectors that fall on one side of the hyperplane are classified as positive, while the others are classified as negative

(Mayfield et al., 2003). Mayfield et al. (2003) is a good example of a solution using a SVM in information extraction.

2.4 Overview Conclusion

As can be seen from the preceding sections, the information extraction field is extremely complex and an analysis of this topic is not a simple task. As noted, we have chosen to focus on named entity extraction, and the following section utilizes the terms, concepts, and axes expounded on in this section to analyze the current solutions available in the IE field.

3 Information Extraction 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 research institutions) and *commercial solutions* (those solutions currently offered as part of a business venture or available from the government). Within these two categories, the solutions are organized alphabetically to allow for simple searching and to remove any indication of partiality towards any of the solutions.

For the purposes of the survey, the focus was primarily on information extraction (IE) solutions rather than information retrieval (IR) solutions. This differentiation is important, since much of work has been done in both domains. According to (Diaz, 2004),

> IR retrieves relevant documents from collections while IE extracts relevant information from documents. In other words, by using IR techniques one gets relevant documents to analyze and by using IE techniques one gets facts out of the documents and analyzes those facts. Information Retrieval recovers from a collection a subset of documents which are (hopefully) relevant to a query, based on keyword searching. Information Extraction is different; its aim is to extract from the documents salient facts about pre-specified types of events, entities, or relationships (Diaz, 2004).

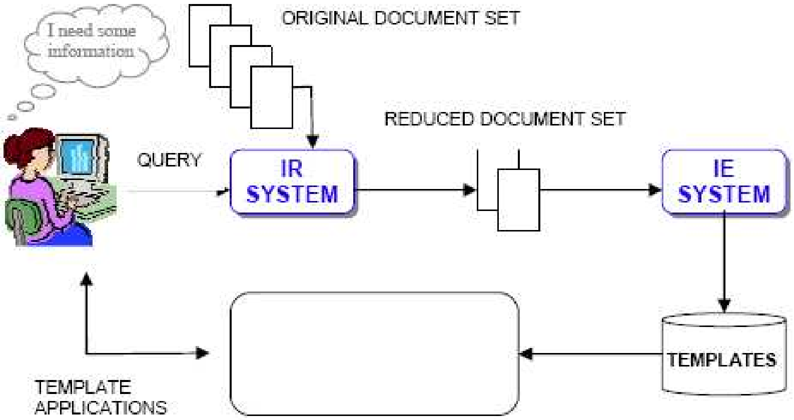


Figure 5.4 – Architecture of a coupled IR-IE system

The report goes on to state the differences in techniques between the two processes. IR primarily uses *Boolean searches* and *ranked-output* approaches while IE uses techniques such as *pattern matching*, *syntactic structure*, *name recognition*, *ontologies*, and *natural language processing*.

However, while these two approaches differ, IR system results can be used as input to an IE system, as seen in the diagram above (Diaz, 2004), or vice versa.

As this is the case in many solutions, it is often difficult to separate the two technologies. The approach of this survey has been to focus specifically on IE technologies, but, since the fields overlap, we have also covered solutions that involve both IR and IE. Other extensive work and studies have been done focusing specifically on the IR domain. SearchTools (2005) is an excellent example of such an IR survey. With this background in mind, we now proceed to the survey results. 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

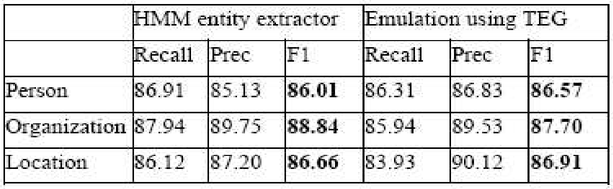
Moderate. This solution is not specially designed for law enforcement applications. The solution cannot be directly used as the named entities extracted in this solution are not comprehensive with respect to the law enforcement domain. However, as with many other IE solutions, it could be used in law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

Evaluation

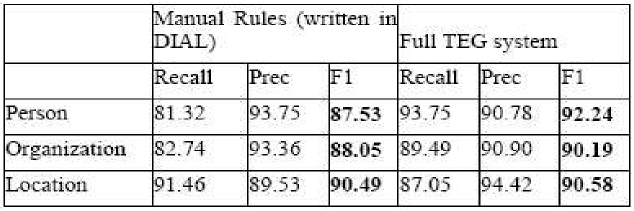
The performance of both named entity extraction and relationship extraction is evaluated in this solution. Named entity extraction is evaluated on MUC-7 data, and the relation extraction is evaluated on ACE-2 data.

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

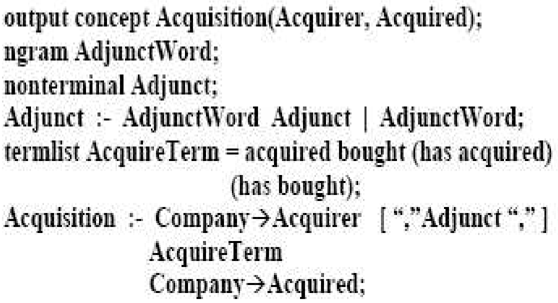
n/a

Inputs Required

Textual data

Information Extraction Algorithm

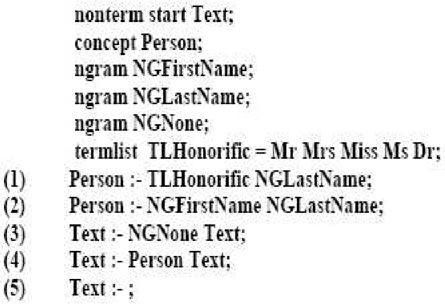
This solution is a hybrid statistical and knowledge-based IE and LA model, and it requires less manual crafting of rules and a smaller amount of training data than other approaches. The solution employs a SCFG (stochastic context-free grammar). Similar to a regular grammar, a string is accepted by a 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):



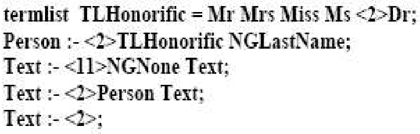
```
output concept Acquisition(Acquirer, Acquired);
ngram AdjunctWord;
nonterminal Adjunct;
Adjunct :- AdjunctWord Adjunct | AdjunctWord;
termlist AcquireTerm = acquired bought (has acquired)
                       (has bought);
Acquisition :- Company→Acquirer  [ “,”Adjunct “,” ]
               AcquireTerm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

Company Introduction and Domain Scope

This solution was developed by researchers at the Indian Institute of Technology, Bombay, India. It aims to automatically segment unformatted textual data into structured elements, such as in segmenting the unformatted address records into a specific formatted address record. This information extraction solution could be used in many domains.

Output/Results

Structured records are produced, which could be a formatted address record, a bibliography entry, etc. For example, the following is a formatted address record:

House Number	201
Street Name	West 4th St
City	Bethlehem
State	PA
Zip Code	18015

Application to Law Enforcement

Moderate. While this solution is not specifically designed for the law enforcement field, it seems that the technology presented could be used as its current focus is not domain-specific. For example, changing an unformatted address into a formatted address record is also a problem that needs to be solved for police criminal report data extraction. However, the solution may not currently have the capability to handle the many attributes which must be extracted in the law enforcement domain.

Evaluation

The paper (Borkar et al., 2001) goes into great detail describing the experiment results for this solution. DATAMOLD was measured on two real-life datasets (an address database and a bibliography database) and was compared to several different automatic approaches, namely Naïve-HMM, Independent-HMM, and Rule-learner. The effect of feature selection and the effect of training dataset size on accuracy was also evaluated. Performance related to running time etc. was not a concern of the evaluation, since their Nested HMM only has no more than a hundred states, and the tests on the largest dataset were completed within an hour.

For the address database, three different real-life address sources were used: *US addresses*, which were downloaded from an internet yellow-page directory and contained 740 addresses, *student addresses*, which contained 2,388 home addresses of students at the authors' university, and *company address*, which contained 769 customer addresses of a major national bank in a large Asian metropolis. During the experiments, the instances were first manually segmented into their corresponding elements. For each dataset, one-third of the data was used for training, and the remaining two-thirds for testing. The overall accuracy was 99% for the US dataset, 88.9% for the student dataset, and 83.7% for the company dataset.

Naïve-HMM, Independent-HMM, Rule-learner and DATAMOLD were tested on the three database datasets using accuracy to measure the performance. Results showed that DATAMOLD was significantly better than Independent-HMM and Naïve-HMM had 3%-10% lower accuracy than DATAMOLD. DATAMOLD was also considerably better than Rapier, which is the rule-learner that was used in the analysis. Detailed performance figures are available in (Borkar et al., 2001).

The bibliography data was gathered from two sources: a set of PDF files whose references were generated by bibtex and bibliographic references from Citeseer. The training set had 100 references and the test set contained 205 references. DATAMOLD AND Rapier were compared against this dataset. DATAMOLD generated an overall accuracy of 87.3%. Although this was lower than the accuracy produced by Rapier, DATAMOLD was able to tag all the tokens while Rapier left many tokens untagged.

Software

n/a

Inputs Required

The input is textual data, but for this solution, it is also restricted to certain types of textual data, such as addresses or bibliographies. A sample address text is "201 West 4th St. Bethlehem, PA 18015."

Information Extraction Algorithm

This solution uses a supervised learning algorithm to construct a probabilistic model, which is based on Hidden Markov Models (HMM). The training data is manually labeled, and some external domain-dependent information is included. The adjacent figure provides an overview of DATAMOLD (Borkar et. al, 2001).

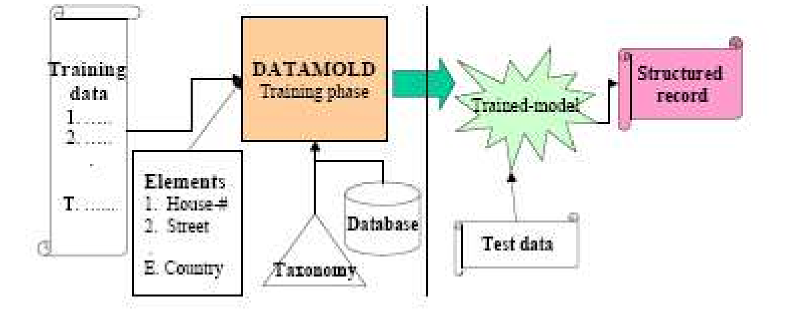


Figure 1: An overview of the working of DATAMOLD.

A basic HMM model is learned through a training and a testing step. During training, the structure of the HMM is decided, (e.g., the number of states and the edges between states) and a dictionary is trained. Then, the transition probabilities and emission probabilities are learned. During the testing portion, the element (attribute) for each symbol must be learned given an output symbol sequence, $S=S_1,S_2,\dots..S_k$. In order to do this, a path of length K which starts from S_1 and ends at S_k needs to be discovered. Generally, there may be more than one path for a given sequence; in this case, the path with the highest probability will be chosen. The Viterbi algorithm is used to discover the most possible path for a given sequence.

A naïve way to model the HMM is to have as many states as the number of elements (attributes) and ensure that all states are completely connected. This solution uses a nested structure of HMMs; each element has its own inner HMM to represent its internal structure. An outer HMM represents the sequence relationship between elements and, within the outer HMM, each inner HMM is treated as a single state.

The nested HMM is learned in two stages. In the first stage, the outer HMM is learned. The training data is now considered as a sequence of elements without considering the length of each element or the words within it. The outer HMM is then trained by these sequences. In the second stage, the structure of the inner HMM is learned. This time, the training data for each element is the sequence of all distinct tokens, which could consist of words, delimiters, and digits.

Knowledge Engineering Cost

The KEC for this solution is medium, since it requires manually labeled training data, but its module generation is automatic. If considering some domain-dependent information which would require a domain term dictionary to be manually developed, the total KEC will be high.

Summary Table

Category: Academic	
University Name: Indian Institute of Technology **University URL**: http://www.iitb.ac.in/	**Location**: Bombay, India
Solution Name: DATAMOLD	
Domain Scope: general	**Application Type**: IE
Knowledge Engineering Cost: medium	**Financial Cost**: n/a
Input Requirements/Preparation Required: textual data	
Named Entity Extraction **Algorithm Name/Group**: Hidden Markov Models (HMM) **Labeling**: manual **Labeling Supervision**: n/a **Model Generation**: automatic **Model Generation Supervision**: supervised **Process Description**: Given training data, a probabilistic model based on Hidden Markov Models (HMM) is built and later given an output symbol sequence. The most possible sequence of elements (attribute) for each symbol is output.	
Solution Output: structured records	
Application to Law Enforcement: moderate	
Is performance evaluation available? yes	**Solution/demo available**? no

Sources

Borkar, V; Deshmukh, K.; and Sarawagi, S (2001). "Automatic segmentation of text into structured records." *Proceedings of SIGMOD, 2001*. Online: http://citeseer.ist.psu.edu/cache/papers/cs/26886/ http:zSzzSzranger.uta.eduzSz~alpzSzixzSzreadingszSzp175-borkar-auto-classify-text-into-structured-records.pdf/borkar01automatic.pdf. Accessed January 18, 2006.

3.2.3 Johns Hopkins University: Named Entity Recognition using Hundreds of Thousands of Features

Solution Introduction and Domain Scope

This solution is developed in Mayfield et al. (2003) by researchers in the Applied Physics Laboratory at Johns Hopkins University, Laurel, Maryland. The technology aims to extract named entities from text and could be used in any domain. But it is especially suitable for language-independent extraction. The solution has been categorized as an Information Extraction solution, since it only extract entities from textual data and learns no relationships or links.

Output/Results

The output is extracted named entities. There are four types of named entities extracted in this solution: *person name*, *location name*, *organization name,* and names of entities that do not belong to the previous three types.

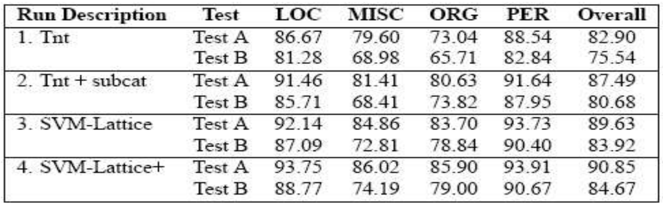
Application to Law Enforcement

Moderate. This solution could be used in law enforcement to extract named entities. Since it is especially designed as a language-independent solution, however, it is more suitable for use when different languages are involved. If a law enforcement officer wanted to extract named entities from non-English language reports, then the solution could have applicability.

Evaluation

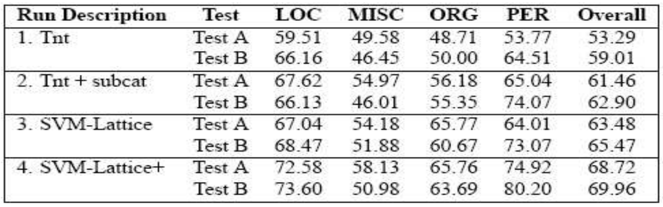
This solution is evaluated on CoNLL-2003 Shared Task training data (Sang and Meulder, 2003), which contains both English- and German-language training and test datasets. The performance of this solution is compared against two baselines using Thorsten Brants' TnT tagger (Brants, 2000).

The SVM-Lattice+ process works similarly to SVM-Lattice, except SVM-Lattice+ uses the output of SVM-Lattice and TnT+subcat as input features. It can been seen from tables 1 and 2 (Mayfield et al., 2003) that SVM-Lattice+'s performance was better than the baselines for both languages.



Run Description	Test	LOC	MISC	ORG	PER	Overall
1. Tnt	Test A	86.67	79.60	73.04	88.54	82.90
	Test B	81.28	68.98	65.71	82.84	75.54
2. Tnt + subcat	Test A	91.46	81.41	80.63	91.64	87.49
	Test B	85.71	68.41	73.82	87.95	80.68
3. SVM-Lattice	Test A	92.14	84.86	83.70	93.73	89.63
	Test B	87.09	72.81	78.84	90.40	83.92
4. SVM-Lattice+	Test A	93.75	86.02	85.90	93.91	90.85
	Test B	88.77	74.19	79.00	90.67	84.67

Table 1: English evaluation results. $F_{\beta=1}$ measures for subcategories, and overall.



Run Description	Test	LOC	MISC	ORG	PER	Overall
1. Tnt	Test A	59.51	49.58	48.71	53.77	53.29
	Test B	66.16	46.45	50.00	64.51	59.01
2. Tnt + subcat	Test A	67.62	54.97	56.18	65.04	61.46
	Test B	66.13	46.01	55.35	74.07	62.90
3. SVM-Lattice	Test A	67.04	54.18	65.77	64.01	63.48
	Test B	68.47	51.88	60.67	73.07	65.47
4. SVM-Lattice+	Test A	72.58	58.13	65.76	74.92	68.72
	Test B	73.60	50.98	63.69	80.20	69.96

Table 2: German evaluation results. $F_{\beta=1}$ measures for subcategories, and overall.

Inputs Required

Textual data.

Information Extraction Algorithm

The basic idea of this solution is to use a large number of features while performing named entity extraction. By employing a large numbers of features, it is not necessary to consider how well a feature is likely to work for a particular language before proposing it. Therefore, these features can be introduced with little concern for dependency among features and without significant knowledge of the target language. However, overfitting might be a problem for this solution, given the large number of features (parameters) used.

The algorithm used in this solution combines Support Vector Machine (SVM) and Lattice approaches. Each sentence is processed individually and a lattice is built with one column per word of the sentence (an additional column indicates a start state). In each column, there is one vertex for each possible tag, which is connected to every vertex in the next column that may legitimately follow it. Given such a lattice, the solution's aim is first to assign probabilities to each of the edges, and then discover the path with the highest likelihood. This path has the highest likelihood of being the correct tagging of this sentence.

A SVM is a binary classifier that uses a supervised learning algorithm to predict whether a given vector is a target class. In order to use an SVM with a lattice, a method is needed to generate probabilities. Platt's method is one such method: $P(y = 1 \mid f) = 1/(1 + exp(Ax + b))$ (Platt, 1999). Platt uses an iterative algorithm to determine sigmoid parameters A and B for given training set vectors and their margins. Platt's method works well when there are sufficient positive examples in the training dataset. However, the training dataset used in this solution is sparse. Two other methods are used to handle this problem: smoothing and manual estimation. By estimating A=-2 and b=0, the performance improved. While the authors state that a learning algorithm should ultimately lead to superior performance over estimation, their approach represents an improvement over Platt's.

The overall approach is described as following: first, features are taken from the training data to form sparse vectors, which is the input for their SVM package, SVMLight 5.00 (Mayfield et al., 2003). Second, the SVM is trained for each transition type identified in the training data and a classifier is generated. Third, test data is formed into vectors, and the generated classifier is used to calculate the margin. Fourth, the margin is mapped to a probability estimate using the static sigmoid described above. In the fifth and final step, a Viterbi-like algorithm is used to discover the most likely path through the lattice.

Knowledge Engineering Cost

We conclude the KEC as high, since the approach requires not only labeled data, but also a huge feature set huge (which also increases the labeling work).

Summary Table

Category: Academic	
Name: Johns Hopkins University **University URL**: http://www.jhu.edu/	**Location**: Laurel, Maryland, USA
Solution Name: Named Entity Recognition using Hundreds of Thousands of Features	
Domain Scope: general (especially language-interdependent domains)	**Application Type**: Named Entity IE
Knowledge Engineering Cost: high	**Financial Cost**: n/a
Input Requirements/Preparation Required: textual data	
Named Entity Extraction **Algorithm Name/Group**: a combination of lattice approach and SVM **Labeling**: manual **Labeling Supervision**: n/a **Model Generation**: automatic **Model Generation Supervision**: supervised **Process Description**: Training dataset is first represented in vectors before training an SVM for each transition type seen in the training data. After test data is formed into vectors, the classifier generated by the SVM is used to calculate the margin. Then, the margin is mapped to a probability estimate using the static sigmoid. Finally, a Viterbi-like algorithm is used to find the most likely path through the lattice.	
Solution Output: named entities	

Application to Law Enforcement: moderate	
Is performance evaluation available? yes	**Solution/demo available**? no

Sources

Brants, Thorsten (2000). "TnT-A Statistical Part-of-Speech Tagger." *Proceedings of 6th Applied NLP Conference, ANLP-2000. Seattle, Washington*. Online. http://citeseer.ist.psu.edu/cache/papers/cs/26650/http:zSzzSzacl.ldc.upenn.eduzSzAzSzA00zSzA00-1031.pdf/brants00tnt.pdf. Accessed January 26, 2006.

Mayfield, James; McNamee, Paul and Piatko, Christine (2003), "Named Entity Recognition using Hundreds of Thousands of Features." *Proceedings of CoNLL-2003, Edmonton, Canada*, pp. 184-187. Online. http://acl.ldc.upenn.edu/W/W03/W03-0429.pdf. Accessed January 9, 2006.

Platt, John C. (1999)." Probabilistic Outputs for Support Vector Machines and Comparisons to Regularized Likelihood Methods." *Advances in Large Margin Classifiers* Scholkopf, B.; Smola, A.; Bartlett, P. and Schuurmans, D., eds., pp. 61-74. MIT Press. Online. http://research.microsoft.com/~jplatt/SVMprob.ps.gz. Accessed January 26, 2006.

Sang, Erik F. Tjong Kim and De Meulder, Fien (2003). "Introduction to the CoNLL-2003 Shared Task: Language Independent Named Entity Recognition." *Proceedings of CoNLL-2003. Edmonton, Canada*. Online. http://acl.ldc.upenn.edu/W/W03/W03-0419.pdf. Accessed January 26, 2006.

3.2.4 Lehigh University: A Semi-supervised Algorithm for Pattern Discovery in Information Extraction from Textual Data

Solution Introduction and Domain Scope

This solution (Wu and Pottenger, 2003) was developed by researchers at Lehigh University in Bethlehem, Pennsylvania. It aims to discover patterns from textual data which can then be used to extraction information from previously unseen data. This solution is an information extraction solution, since it extracts named entities from textual data, but learns no relationships or links.

Output/Results

The output is regular expression rules, which represent the patterns of named entities. These rules can be used later to extract information from previously unseen data.

Application to Law Enforcement

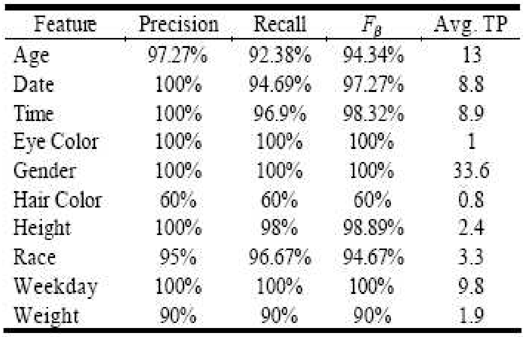
Extensive. The approach is suitable for use in the law enforcement field. After training and discovery of the regular repression rule, it can be used to extract named entities from textual data held by law enforcement agencies. The solution is currently being used by the Bethlehem (PA) Police Department to extract information from narrative police reports.

Evaluation

Domain expert labeled *segments* were used for training the solution. Then, two different methods were used to evaluate the training results. The first method tested *segment evaluation*, which determined whether segment labels were correctly predicted. The second method evaluated the performance of the model with respect to an exact match of the feature of interest. The metric F_β (with β=1 to balance precision and recall) was used to evaluate the test performance with both methods, based on 10-fold cross-validation.

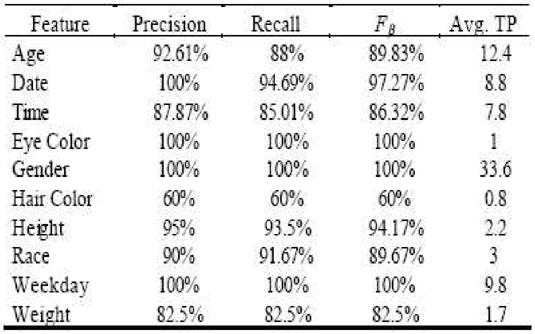
The training set consists of 100 incident reports obtained from Fairfax County, Virginia. These reports were automatically segmented into 1404 segments. The first column of the tables below depicts the 10 features currently supported by the solution. *Eye Color* and *Hair Color* were not well represented in their dataset due to their infrequent appearance in the Fairfax County data.

Table 1. 10-fold cross-validation test performance based on segment evaluation



Feature	Precision	Recall	F_β	Avg. TP
Age	97.27%	92.38%	94.34%	13
Date	100%	94.69%	97.27%	8.8
Time	100%	96.9%	98.32%	8.9
Eye Color	100%	100%	100%	1
Gender	100%	100%	100%	33.6
Hair Color	60%	60%	60%	0.8
Height	100%	98%	98.89%	2.4
Race	95%	96.67%	94.67%	3.3
Weekday	100%	100%	100%	9.8
Weight	90%	90%	90%	1.9

Table 2. 10-fold cross-validation test performance based on exact match



Feature	Precision	Recall	F_β	Avg. TP
Age	92.61%	88%	89.83%	12.4
Date	100%	94.69%	97.27%	8.8
Time	87.87%	85.01%	86.32%	7.8
Eye Color	100%	100%	100%	1
Gender	100%	100%	100%	33.6
Hair Color	60%	60%	60%	0.8
Height	95%	93.5%	94.17%	2.2
Race	90%	91.67%	89.67%	3
Weekday	100%	100%	100%	9.8
Weight	82.5%	82.5%	82.5%	1.7

The table on the left above presents the results that occurred when segment evaluation was used; as can be seen, the performance of this approach is high. For *Eye Color*, *Gender* and *Weekday*, the test performance is perfect (100%), which is partially due to the modification of the lexicon used in part of speech tagging to label these features during pre-processing. For *Age, Date, Time, Height, Race,* and *Weight*, the results are also excellent (F_β =90%). For *Hair Color*, however, the performance is not as good. As noted, this is due to the lack of *Hair Color* segments in test sets.

The table on the right shows the performance of exact match. An exact match occurs when a sub-string extracted by an RRE is exactly the same as the string that would be identified by a domain expert. In Table 2, it turns out that the RREs discovered automatically for these five features are exactly the same patterns developed manually by human experts who studied this same dataset. Other features, although not perfect, also have reasonable good performance (over 80%).

Software

The classification code is available at http://hddi.cse.lehigh.edu. The code for training (PERL) is available on request.

Inputs Required

Textual data from plain text formatted documents.

Labeling Algorithm

Semi-supervised algorithm.

Information Extraction Algorithm

The authors' approach employs a semi-supervised algorithm to discover patterns in textual data. The patterns discovered are represented as reduced regular expressions. This algorithm is termed 'semi-supervised' since it requires less knowledge engineering cost than other approaches. Given a training dataset, this algorithm is able to automatically find patterns and generate corresponding reduced regular expressions, which can be used later to extract information from previously unseen data. Labeling involves only segment labeling, not individual word labeling. For example, if a segment contains certain a word whose feature is type *A*, then the type of this segment is labeled as *A*. The overall process involves three main steps:

1. First Data needs to be Pre-processed

The IE system first divides textual data into segments. Then, the user needs only to label segments instead of words, saving time and energy which reduces the knowledge engineering cost. Nonetheless, a domain expert must first identify the features to be extracted. For example date, address, person name, and vehicle name all could be features. If in a particular application, only vehicle names are needed, there is only one feature: vehicle name. Each segment which contains vehicle names is assigned a "vehicle name" label.

2. Learning Reduced Regular Expressions

The goal of the algorithm is to discover sequences of words and/or part of speech tags that, for a given feature, have high frequency in the true set of segments and low frequency in the false set. A greedy covering algorithm is used to generate the reduced regular expressions. After one rule is generated, all the segments covered by it are removed from the true set. The remaining segments become a new set, and a new rule is discovered to cover the segments in it. This process is repeated until the number of segments left in the true set is less than or equal to a user-defined threshold.

The regular expression rule discovery process is presented in the figure below (Wu and Pottenger, 2003). First, the most common element of an RRE (root) is discovered, then the algorithm extends the 'length' of RRE by the "AND" learning process. In the "OR" learning process, the 'width' of the RRE is extended. Next, optional elements are discovered during the "Optional" learning process. The algorithm then proceeds with the "NOT" learning process, and finally discovers the start and the end of the RRE. This approach has been demonstrated to work effectively to extract a variety of entities including date, address, person name, vehicle name, and several other named entities (such as definitions).

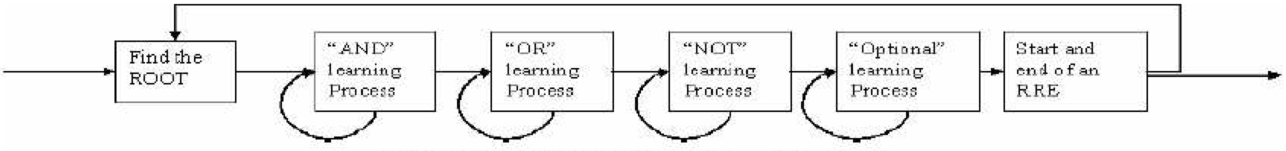


Figure 1: RRE Discovery Process

3. Post Processing

One RRE is generated in each iteration of the second step, which creates a sub-pattern of the current feature. Once all RREs for a given feature have been discovered (i.e., all segments labeled for the feature are covered), the system uses the "OR" operator to combine the RREs into a single rule.

Knowledge Engineering Cost

We conclude KEC as *medium*, since this solution uses a semi-supervised algorithm which only requires labeling a segment instead of labeling the exact location of a word. This reduces the KEC compared to supervised learning algorithms.

Summary Table

Category: Academic	
Company Name: Lehigh University **Company URL**: http://hddi.cse.lehigh.edu/	**Location**: Bethlehem, PA, USA
Solution Name: A Semi-supervised Algorithm for Pattern Discovery in Information Extraction from Textual Data	
Domain Scope: general	**Application Type**: NE IE
Knowledge Engineering Cost: medium	**Financial Cost**: n/a
Input Requirements/Preparation Required: textual data	
Named Entity Extraction	

Algorithm Name/Group: Covering Algorithm **Labeling**: manual **Labeling Supervision**: n/a **Model Generation**: automatic **Model Generation Supervision**: semi-supervised **Process Description**: Textual data is divided into segments that are manually labeled. Then, a covering algorithm is used to generate regular expressions for extracting named entities.	
Solution Output: regular expressions	
Application to Law Enforcement: extensive	
Is performance evaluation available? yes	**Solution/demo available**? yes

Sources

Wu, Tianhao and Pottenger, William M. (2003). "A Semi-supervised Algorithm for Pattern Discovery in Information Extraction from Textual Data." *Proceedings of the Seventh Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD-03). Seoul, Korea, April/May, 2003*. Online. http://www.cse.lehigh.edu/~billp/pubs/PAKDD03.pdf. Accessed January 11, 2006.

3.2.5 Lehigh University: A Semi-Supervised Active Learning Algorithm for Information Extraction from Textual Data

Company Introduction and Domain Scope

This solution was developed by researchers at Lehigh University in Bethlehem, Pennsylvania. It aims to discover patterns in extracted information from textual data. It is especially developed for extraction of law enforcement narrative text sources, but it could also be used in other domains to perform information extraction. Therefore, this solution has been categorized as an information extraction solution.

Output/Results

The solution discovers reduced regular expressions (RREs), which can be used to extract information from textual data.

Application to Law Enforcement

Extensive. This solution was specifically designed for law enforcement use. It aims to extract demographic and modus operandi features from police reports, e.g. suspect's height, weight, race, etc. The technology is useful in extracting entities from narrative police reports that are subsequently stored as structured data for use in information retrieval or link analysis.

Evaluation

The solution has been evaluated in two ways. The first was the performance of the information extraction, which was evaluated using the metrics of precision, recall and $F_{ß}$. The second approach evaluated the labeling effort (as a percentage) saved by using the active learning algorithm.

For the information extraction performance, the algorithm with active learning was compared to the algorithm without active learning, which had already been shown to have good performance for information extraction. The results of the evaluation were mixed. For some attributes, e.g. time, the algorithm with active learning produced even better results than the algorithm without active learning. For other attributes, e.g., eye color, gender, week day, the two algorithms had the same performance. On still other attributes (e.g., hair color, height, race, and weight), the algorithm with active learning performed worse than the one without active learning. However, for these latter four features, compared with the ≥90% reduction in labeling effort (i.e., knowledge engineering cost), the decrease of information extraction performance is acceptable. This illustrates that fact that there is a tradeoff between information extraction performance and reduction of labeling effort. More details can be found in Wu and Pottenger (2005a).

The evaluation using active learning showed that the labeling effort used to develop the training set was greatly reduced. Test results showed that use of active learning resulted in a significant reduction in labeling effort for nine out of the 10 attributes tested.

Software

TMI BPD_IE 1.1 and 1.0 are implementations of the RRE Discovery classification algorithms and are available at http://hddi.cse.lehigh.edu. The training version is available on request.

Inputs Required

Textual data.

Information Extraction Algorithm

This solution combines a semi-supervised algorithm together with an active learning algorithm. The semi-supervised algorithm was described above in Section 3.2.4. The active learning algorithm further reduces knowledge engineering cost. The following illustrates the operation of the active learning algorithms: first, a user inputs commonplace attribute descriptions to the active learning algorithm. For example, "six feet tall" could be a seed for the attribute "height." Based on the given seeds, the semi-supervised algorithm discovers reduced regular expressions to represent the context surrounding the seeds. Then, all segments which have similar contexts as the seeds are considered to be the candidate items in the training data true set for the given attribute. The training set developer then interactively selects those segments that actually contain the desired attribute. This process is called *active learning*, since a human user first provides the algorithm some seeds, or examples, that the technique uses as a basis for discovering more candidates; the user then evaluates these candidate segments and selects those that are relevant.

Upon completion of the active learning, the true set and false set for training have been generated. The semi-supervised learning algorithm described above in Section 3.2.4 is used then applied to the training data to discover reduced regular expression rules. The entire learning process is shown in the adjacent figure (Wu and Pottenger, 2005a). We classify this process as a hybrid learning method, since it not only uses active learning, but also utilizes the RRE-based semi-supervised learning algorithm to both extract a context around the seeds and to discover a pattern for the target feature.

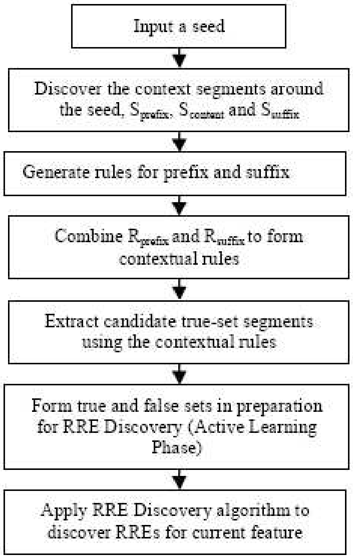

Input a seed
Discover the context segments around the seed, S_{prefix}, $S_{content}$ and S_{suffix}
Generate rules for prefix and suffix
Combine R_{prefix} and R_{suffix} to form contextual rules
Extract candidate true-set segments using the contextual rules
Form true and false sets in preparation for RRE Discovery (Active Learning Phase)
Apply RRE Discovery algorithm to discover RREs for current feature

FIG. 8. Active learning flow chart

Knowledge Engineering Cost

The KEC for this solution is medium. Although it is medium, it is lower than the KEC for the solution outlined in Section 3.2.4 because this solution uses an active learning algorithm to reduce the KEC for labeling the training data. Otherwise, the two solutions are similar, such as the use of segment labeling (which reduces KEC) instead of labeling the exact location of features. Given the training data, the reduced regular expression rules are automatically generated, so the KEC is medium.

Summary Table

Category: Academic	
University Name: Lehigh University **Company URL**: http://hddi.cse.lehigh.edu/	**Location**: Bethlehem, PA

Solution Name: A Semi-Supervised Active Learning Algorithm for Information Extraction from Textual Data	
Domain Scope: general	**Application Type**: IE
Knowledge Engineering Cost: medium	**Financial Cost**: n/a
Input Requirements/Preparation Required: Textual data	
Named Entity Extraction **Algorithm Name/Group**: a semi-supervised algorithm combined with an active learning algorithm **Labeling**: hybrid **Labeling Supervision**: supervised **Model Generation**: automatic **Model Generation Supervision**: semi-supervised **Process Description**: Given training data, a semi-supervised learning algorithm is used to discover the RREs for each attribute. An active learning algorithm is used to further reduce the KEC for labeling the data. First, seeds are input to the active learning algorithm. Then, the context around the seeds is used to identify candidates for the true set for a given attribute. Finally, the candidates are manually pruned and the rule is automatically generated from the remaining segments.	
Solution Output: Reduced Regular Expression (RRE) rules	
Application to Law Enforcement: extensive	
Is performance evaluation available? yes	**Solution/demo available**? yes

Sources

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.

3.2.6 National University of Singapore: Named Entity Recognition with a Maximum Entropy Approach

Solution Introduction and Domain Scope

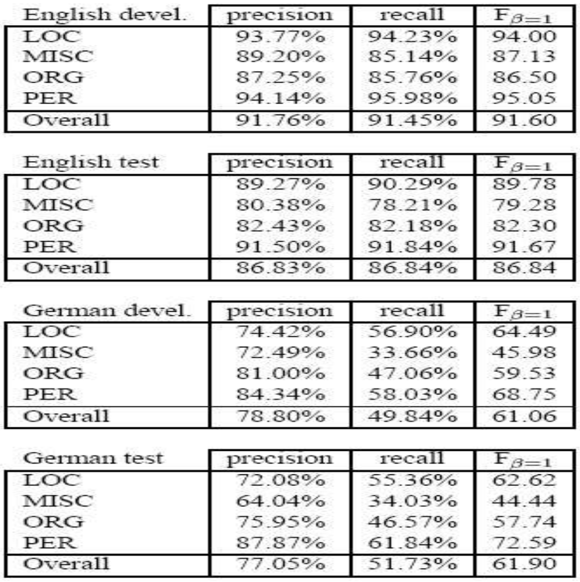
This solution described in (Chieu and Ng, 2003) was developed by researchers in DSO National Laboratories and the National University of Singapore, Singapore. It aims to extract named entities from text. It could be used in any domain which requires the extraction of named entities from text. Since it only extract entities from textual data and formulates no relationships or links, it has been classified as an information extraction solution.

Output/Results

The output is extracted named entities. There are four types of named entities extracted in this solution: person name, location name, organization name, and miscellaneous.

Application to Law Enforcement

Moderate. This solution could be used in Law Enforcement to extract named entities.



English devel.	precision	recall	$F_{\beta=1}$
LOC	93.77%	94.23%	94.00
MISC	89.20%	85.14%	87.13
ORG	87.25%	85.76%	86.50
PER	94.14%	95.98%	95.05
Overall	91.76%	91.45%	91.60

English test	precision	recall	$F_{\beta=1}$
LOC	89.27%	90.29%	89.78
MISC	80.38%	78.21%	79.28
ORG	82.43%	82.18%	82.30
PER	91.50%	91.84%	91.67
Overall	86.83%	86.84%	86.84

German devel.	precision	recall	$F_{\beta=1}$
LOC	74.42%	56.90%	64.49
MISC	72.49%	33.66%	45.98
ORG	81.00%	47.06%	59.53
PER	84.34%	58.03%	68.75
Overall	78.80%	49.84%	61.06

German test	precision	recall	$F_{\beta=1}$
LOC	72.08%	55.36%	62.62
MISC	64.04%	34.03%	44.44
ORG	75.95%	46.57%	57.74
PER	87.87%	61.84%	72.59
Overall	77.05%	51.73%	61.90

Table 2: Results for development and test set for the two languages by ME1

The number of unstructured digital documents is increasing day by day. Manually extracting useful information from these digital documents is expensive and inefficient. An information extraction tool such as this could be utilized by law enforcement agencies to do this work.

Evaluation

This solution is evaluated on CoNLL-2003 Shared Task training data (Sang and Meulder, 2003), which contains training and test data for the English and German languages. The best result is in table above (Chieu and Ng, 2003). Their system works well for the Location name and Person name classes, but does not perform well for the Organization name and Miscellaneous name classes. The reason for poor Miscellaneous name extraction is that the range for miscellaneous is too general, e.g., it includes both movie names and theater names.

Inputs Required

Input is textual data.

Information Extraction Algorithm

This solution uses a maximum entropy approach to the named entity extraction (NEE) task, in which it not only uses the local features (those occurring within a single sentence), but also global features (occurrences of each word within the same document). The maximum entropy classifier is used to classify each word as one of the following types: the beginning of an NE, a word inside an NE, the last word of an NE, or a unique word in an NE.

Given words in a sentence *s* in a document *D*, the probability of the classes $c_1,.....c_n$ assigned to them is defined as: $P(c_1,.....,c_n \mid s,D) = \prod_{i=1}^{n} P(c_i \mid s,D) * P(c_i \mid c_{i-1})$, where $P(c_i \mid s, D)$ is determined by the maximum entropy classifier, and $P(c_i \mid c_j)$ is one if the sequence is admissible (otherwise it is 0).

Feature representation is discovered using one of two methods discussed in this solution: ME1 (which only uses the knowledge in the training data) and ME2 (which uses some external knowledge such as some additional features derived from name lists). The training data is first preprocessed to compile a number of lists that are used by both ME1 and ME2. These lists are derived automatically from the training data. For example, the Frequent Word List (FWL) consists of words that occur in more than five different documents. Then, the basic features which are used by both ME1 and ME2 are divided into two types: local and global. Local features of a token *w* are derived from the sentences containing *w*. A “First word, case, and Zone” is one example of a feature, while “Case and Zone of w_{+1} and w_{-1}” provides a second example. Global features are derived by looking up other occurrences of *w* within the same document, e.g., unigram, bigrams, class suffixes, etc. In addition to the basic features used by both ME1 and ME2, ME2 uses additional features derived from name lists that have been compiled from the Internet and labeled by the researchers. The name list is pairs of words and their class, e.g., “Tom Kenny: PERSON.”

Knowledge Engineering Cost

This solution’s KEC has been classified as medium, since data needs to be manually labeled and the learning algorithm is automatic. If using ME2 for feature representation, the KEC will be slightly higher than ME1, since it also needs to manually add the additional entries to the automatically compiled NCS (Name Class Suffixes) list.

Summary Table

Category: Academic	
University Name: DSO National Laboratories and National University of Singapore	**Location**: Singapore

University URL: http://www.nus.edu.sg/	
Solution Name: Named Entity Recognition with a Maximum Entropy Approach	
Domain Scope: general	**Application Type**: IE (NE)
Knowledge Engineering Cost: medium	**Financial Cost**: n/a
Input Requirements/Preparation Required: textual data	
Named Entity Extraction **Algorithm Name/Group**: Maximum Entropy **Labeling**: manual **Labeling Supervision**: n/a **Model Generation**: automatic **Model Generation Supervision**: supervised **Process Description**: A maximum entropy approach is used for this NEE task. Given words in a sentence, the probability of the classes c_1,…..c_n is defined as $P(c_1,.....,c_n \mid s,D)=\prod_{i=1}^{n} P(c_i \mid s,D) * P(c_i \mid c_{i-1})$, in which $P(c_i \mid s, D)$ is determined by the maximum entropy classifier, and $P(c_i \mid c_j)$ is one if the sequence is admissible, otherwise 0. Local features, global features and external knowledge (for ME2) are used.	
Solution Output: named entities	
Application to Law Enforcement: moderate	
Is performance evaluation available? yes	**Solution/demo available**? no

Sources

Chieu, Hai Leong and Ng, Hwee Tou (2003). "Named Entity Recognition with a Maximum Entropy Approach." *Proceedings of the Seventh Conference on Natural Language Learning (CoNLL-2003), Edmonton, Alberta, Canada* pp. 160-163. Online. http://www.comp.nus.edu.sg/~nght/pubs/conll03.pdf. Accessed January 26, 2006.

Sang, Erik F. Tjong Kim and De Meulder, Fien (2003). "Introduction to the CoNLL-2003 Shared Task: Language Independent Named Entity Recognition." *Proceedings of CoNLL-2003. Edmonton, Canada*. Online. http://acl.ldc.upenn.edu/W/W03/W03-0419.pdf. Accessed January 26, 2006.

3.2.7 Open University: ESpotter

Solution Introduction and Domain Scope

ESpotter is a solution designed by researchers in the Open University, UK. ESpotter is a browser plug-in that can accurately and efficiently perform Named Entity Recognition. It could be used in many domains, but it is only currently able to handle Web pages. We categorize it as IE, since it only extracts named entities from Web pages and does not identify relationships or links.

Output/Results

Named entities recognized, which are highlighted on the Web pages. There are ten types of entities extracted, such as person name and organization name.

Application to Law Enforcement

Moderate. Such a tool could be helpful in law enforcement applications, since it could help police officers quickly identify the valuable information from the overwhelming amount of information on the Web.

Evaluation

ESpotter was tested on ten selected Websites and five web pages per Website. For ten types of entities (People, Organization, Location, Date, etc.), ESpotter achieved an average precision of 81% and recall of 62%. After using some user customization, such as additions of new lexicons and patterns, the average precision and recall were improved to 92% and 82%, respectively.

Inputs Required

Web pages.

Software

The solution exists as a browser plug-in, as can be seen in the figure below. A demo of the solution is available at http://kmi.open.ac.uk/people/jianhan/ESpotter/.

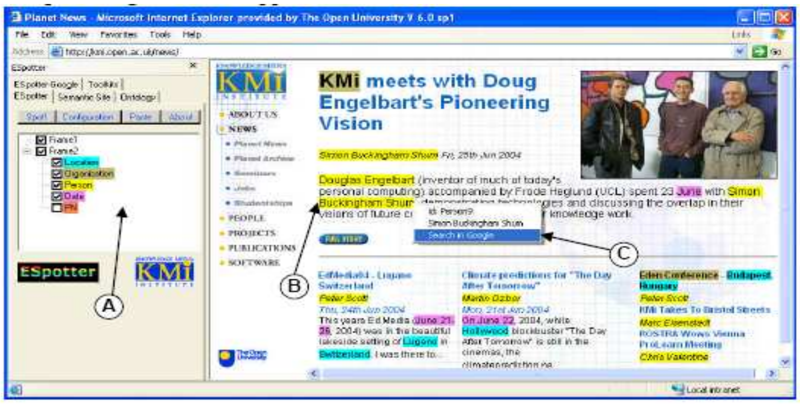

Planet News - Microsoft Internet Explorer provided by The Open University V 6.0 sp1
File Edit View Favorites Tools Help
Address http://kmi.open.ac.uk/news/
ESpotter
ESpotter Google | Toolkits |
ESpotter | Semantic Site | Ontology |
Spot | Configuration | Paste | About
Frame1
Frame2
Location
Organization
Person
Date
PN
ESpotter
KMi
ABOUT US
NEWS
PEOPLE
PROJECTS
PUBLICATIONS
SOFTWARE
KMi meets with Doug Engelbart's Pioneering Vision
Simon Buckingham Shum Fri, 25th Jun 2004
Douglas Engelbart (inventor of much of today's personal computing) accompanied by Frode Hegland (UCL) spent 23 June with Simon Buckingham Shum
Id: Person9
Simon Buckingham Shum
Search in Google
A
B
C
Local intranet

Fig. 1. ESpotter highlighting a page on the KMi Web site [4], showing (A) the ESpotter toolbar (B) entities highlighted according to their types on the Web page (C) services provided for entities (such as search for them in Google).

Information Extraction Algorithm

Named Entity Recognition (NER) discovers proper names of various types (such as “Tom Smith,” a person name, or “Lehigh University,” an organization name). No labeling is required in this solution. Normally, two techniques are used in NER: *lexicon matching* and *pattern matching*. A *lexicon* contains pairs of entities and their type, e.g., [Lehigh University, Organization]. A *pattern* consists of the formal description of the content structure of a type of entities, which is usually described in regular expressions. An example of a pattern would be the following: “a word starting with capitalized letter and followed by ‘University’ is an ‘Organization’ type.”

ESpotter is a NER browser plug-in (see above figure from (Zhu et al., 2005)). It differs from previous NER systems in two ways: (1) lexicons and patterns are adapted to domains on the web. (2) lexicons and patterns are adapted to individual users.

1. Domain Adaptation

On the Web where domains can be changed by just a click, a NER system should be able to adapt to different domains quickly. There are two types of domain adaptation. The first type exists where the same entity has different meaning for different domains. For example “Magpie” is a type of “bird” on the Royal Society for Protection

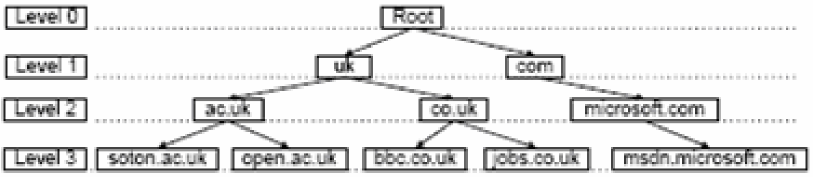


Fig. 3. A Domain Hierarchy Defined on Domain URIs.

of Birds Website, but a type of "Project" on the Knowledge Media Institute Website. Some entities have a higher possibility to appear within certain domains. For example, UK postal addresses are not likely to appear on Websites other than UK ones.

ESpotter uses a domain hierarchy (see figure above (Zhu et al., 2005)) to perform domain adaptation. The hierarchy consists of different level domains links between domains on two adjacent levels. Domains are represented by their Unified Resource Identifiers (URI). A lexicon or pattern is defined on the domain hierarchy, and a precision between zero and one is assigned to each in each domain. The higher the precision, the higher the possibility that it is related to the given domain.

2. User adaptation

First, users can add new lexicons and patterns to ESpotter, and assign their precisions on different domains. Second, users can customize ESpotter to fulfill their task at hand. They can set a "precision" threshold to control the precision and recall of NER. A higher threshold will result in a more accurate NER, but it will also miss more entities. Users can also select the types of entities they want to discover; for example, the solution could be directed to only identify "Person" types. They can also modify or delete current lexicons and patterns and give feedback on the NER results (which could be used by ESpotter to make adjustments or modifications). Then, the lexicons and patterns can be used to recognize entities. Although not explicitly discussed in the solution description referenced, the patterns are evidently manually generated. ESpotter is however able to automatically calculate the probability of lexicons and patterns on different domains. After the entities are recognized, they are highlighted according to their type, with one color for each type. These colors can also be configured by the user.

Knowledge Engineering Cost

We conclude the KEC as high, since apparently ESpotter uses manually generated lexicons and patterns. Manually generating rules is much more difficult than labeling data and leads to higher KEC. The development of this solution is also more involved that a simple prototype, and the design work also leads to a higher KEC.

Summary Table

Category: Academic	
University Name: the Open University **University URL**: http://www.open.ac.uk/	**Location**: Milton Keynes, UK
Solution Name: ESpotter	
Domain Scope: general (web browser)	**Application Type**: NE IE
Knowledge Engineering Cost: high	**Financial Cost**: n/a
Input Requirements/Preparation Required: Web pages	
Named Entity Extraction **Algorithm Name/Group**: knowledge based named entity recognition **Labeling**: n/a **Labeling Supervision**: n/a **Model Generation**: manual **Model Generation Supervision**: n/a **Process Description**: Lexicons and patterns are used to recognize named entities. But, since the lexicons and patterns are manually crafted, it is termed "knowledge based." This solution added some domain adaptation to the normal NER methods, and lexicons and patterns are given different probabilities based on the domain.	
Solution Output: named entities (10 types)	

Application to Law Enforcement: moderate	
Is performance evaluation available? yes	**Solution/demo available**? yes

Sources

Zhu, Jianhan; Uren, Victoria and Motta, Enrico (2005). “ESpotter: Adaptive Named Entity Recognition for Web Browsing.” Proceedings of *Workshop on IT Tools for Knowledge Management Systems at WM2005 Conference, Kaiserslautern, Germany,* pp. 505-510. April 11-13, 2005. Online. http://kmi.open.ac.uk/people/jianhan/zhuetal_KMTools.pdf. Accessed January 10, 2006.

3.2.8 University College Dublin: Boosted Wrapper Induction

Solution Introduction and Domain Scope

This solution was developed by researchers at the University College Dublin in Dublin, Ireland. This solution sets up an information extraction system which performs information extraction in machine-generated or highly-structured text. It has applicability in many domains. The solution has been categorized as an information extraction solution.

Output/Results

The output is a wrapper, which could be understood as a set of patterns. These patterns can be used to extract desired fields/attributes. For example, the pattern ([< a hret =”], [http]) represents the beginning of a URL, and pattern ([. html], [“ >]) represents the end of a URL. These two patterns together can be used to extract the URL.

Application to Law Enforcement

Moderate. This solution is not specifically designed for the law enforcement field. But, as with many other IE solutions, it could be used in a law enforcement setting since named entity extraction is needed to change narrative police reports into structured data. However, it cannot be directly used as changes are needed to allow the system to be used in this capacity.

Evaluation

This solution was evaluated on 16 information extraction tasks defined over eight distinct document collections. These collections included seminar announcements, Reuters articles detailing corporate acquisitions, job announcements, and various kinds of Web pages. Three of the domains tested with narrative textual data, and five of them employed structured textual data. Cross validation was used to perform the experiments, which means the sources were divided into a training set and a test set, both of which were manually labeled. A wrapper was learned from the training set and tested on the testing set. Precision, recall and F-measure were the metrics used to evaluate the performance.

The effect of boosting was also evaluated to see how it impacted performance. It was learned that the boosting rounds required by the system to reach its best performance differed according to the task. Some tasks, such as extraction from seminar announcements, required up to 500 rounds. This solution is compared with other four systems: two rules learners (SRV and Rapier), an HMM algorithm, and the stalker wrapper induction algorithm. The results showed that this solution attained better precision than the others and still maintained a good recall. Details can be found in Freitag and Kushmerick (2000).

Software

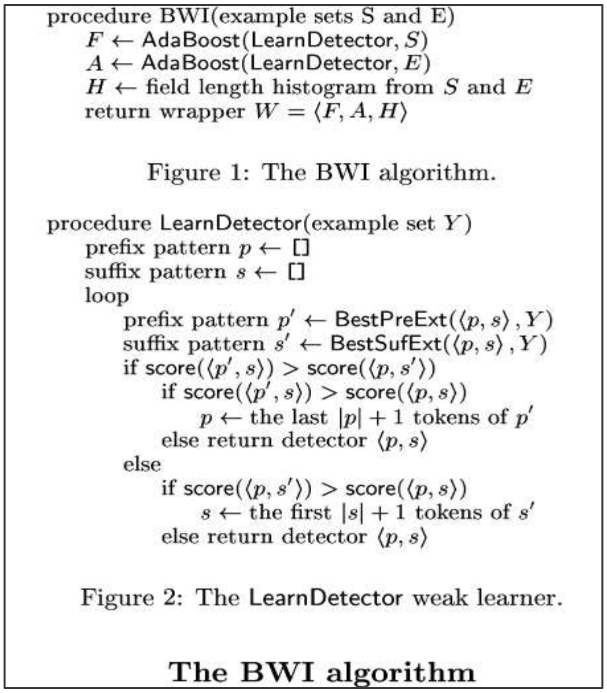
n/a

Inputs Required

Input could be textual data, structured data, or semi-structured data. HTML documents are an example of semi-structured data that can be processed by the solution.

Information Extraction Algorithm

In this solution, the information extraction task is treated as a classification problem, so it aims to learn the classifier or *wrapper*. Documents are treated as sequences of tokens, and the task of information extraction is to identify fields in the *sequences* which consist of one or more distinguished token subsequences. Identifying the fields require the identification of both the beginning and end of each field, which are referred to as the *boundary* of each field. In this context, "field" is equivalent to "attribute." For both semi-structured and structured data, wrapper induction generates simple but highly accurate patterns. For narrative data, however, a small set of simple rules is not sufficient. More rules must be generated and then combined to produce good results. A *pattern* is a sequence of tokens. A *boundary detector d=<p,s>* is a pair of patterns where *p* is a prefix pattern and *s* is suffix pattern. *d* matches a boundary *i* if *p* matches the tokens before *i*, and *s* matches the tokens after *i*. A *wrapper W=<F, A, H>* is composed of two sets (*F={F*$_1$*,…,F*$_T$*}*, which identifies field-starting boundaries, and *A={A*$_1$*,…,A*$_T$ *}* which identifies the field end boundaries) and a function *H:[-∞,+∞]→[0,1]*, which reflects the probability that a field has length *k*. In this solution, boosting is used to improve the performance of the wrapper induction algorithm. *Boosting* is a method to improve the performance of a machine learning algorithm by repeatedly applying the algorithm to the training set and modifying the training example weight each time to emphasize the examples on which the algorithm has done poorly in previous steps. Learning a wrapper *W* includes determining *F, A* and *H* from the example sets *S* and *E*. BWI (Boosted Wrapper Induction) is the algorithm used to learn a wrapper. BWI repeatedly invokes an algorithm to learn the boundaries of fields, and repeatedly re-weight the training examples. This is presented in the above figure (Freitag and Kushmerick, 2000).



```
procedure BWI(example sets S and E)
    $F \leftarrow$ AdaBoost(LearnDetector, $S$)
    $A \leftarrow$ AdaBoost(LearnDetector, $E$)
    $H \leftarrow$ field length histogram from $S$ and $E$
    return wrapper $W = \langle F, A, H \rangle$
```

Figure 1: The BWI algorithm.

```
procedure LearnDetector(example set $Y$)
    prefix pattern $p \leftarrow []$
    suffix pattern $s \leftarrow []$
    loop
        prefix pattern $p' \leftarrow$ BestPreExt($\langle p, s \rangle, Y$)
        suffix pattern $s' \leftarrow$ BestSufExt($\langle p, s \rangle, Y$)
        if score($\langle p', s \rangle$) > score($\langle p, s' \rangle$)
            if score($\langle p', s \rangle$) > score($\langle p, s \rangle$)
                $p \leftarrow$ the last $|p| + 1$ tokens of $p'$
            else return detector $\langle p, s \rangle$
        else
            if score($\langle p, s' \rangle$) > score($\langle p, s \rangle$)
                $s \leftarrow$ the first $|s| + 1$ tokens of $s'$
            else return detector $\langle p, s \rangle$
```

Figure 2: The LearnDetector weak learner.

The BWI algorithm

Knowledge Engineering Cost

We conclude the KEC for this solution is high, since it requires manually labeled training sets, and no methods are used to reduce the labeling effort. The model generation is however automatic.

Summary Table

Category: Academic	
University Name: University College Dublin **University URL**: http://www.ucd.ie/	**Location**: Ireland
Solution Name: Boosted Wrapper Induction	
Domain Scope: general	**Application Type**: IE
Knowledge Engineering Cost: high	**Financial Cost**: n/a

Input Requirements/Preparation Required: Textual data, structured data or semi-structured data	
Named Entity Extraction **Algorithm Name/Group**: a wrapper induction algorithm refined by boosting **Labeling**: manual **Labeling Supervision**: n/a **Model Generation**: automatic **Model Generation Supervision**: supervised **Process Description**: Wrappers, which are normally used on highly structured data, are used in this solution as the information extraction learning technique. In this solution, it is used to do information extraction from structured textual data, and boosting is used to improve its performance.	
Solution Output: a wrapper, which is a set of patterns	
Application to Law Enforcement: moderate	
Is performance evaluation available? yes	**Solution/demo available**? no

Sources

Freitag, Dayne and Kushmerick, Nicholas (2000). "Boosted Wrapper Induction." *Proceedings of the Seventeenth National Conference on Artificial Intelligence and Twelfth Conference on Innovative Applications of Artificial Intelligence.* Online: http://citeseer.ist.psu.edu/cache/papers/cs/23958/http:zSzzSzwww.cs.ucd.iezSzstaffzSznickzSzhomezSzresearchzSzdownloadzSzfreitag-aaai2000.pdf/freitag00boosted.pdf. Accessed January 19, 2006.

3.2.9 University of Alicante: DRAMNERI

Company Introduction and Domain Scope

This solution was developed by researchers at the University of Alicante in Alicante, Spain. It is a free software application to perform named entity recognition. This multilingual solution could be used in many domains. Although it only performs named entity recognition and not named entity extraction, it is still categorized as an information extraction solution as it performs the most critical steps in information extraction.

Output/Results

The named entities are recognized.

Application to Law Enforcement

Moderate. This solution was not specifically designed for law enforcement field, but it could be used to aid in the conversion of narrative police data into a structured format. While the system recognizes entities, it does not extract them. It could serve as a starting point to a more applicable law enforcement solution.

Evaluation

This solution was not directly evaluated. Rather, it was used in an information extraction system whose domain consists of notarial documents as well as in a Question Answering (QA) system. In the QA system, the solution was used between an IE module and a QA module. When the IE process returns documents that do not contain an entity which belongs to the same category as the query the documents must be filtered. By applying the solution in filtering, a 26% reduction in data and a 9% increase in performance was achieved.

Software

The solution, DRAMNERI, can be downloaded from http://www.dlsi.ua.es/~atoral/#Software . This software has a free license.

Inputs Required

Plain text

Information Extraction Algorithm

Named entity recognition (NER) is a subtask of named entity extraction. NER includes identifying and categorizing entity names. Normally, there are two ways to perform NER: one is based on *domain knowledge*, while the other uses a *supervised learning algorithm.* The former requires the use of a gazetteer (dictionary) and rules, whereas the latter needs a labeled training data set. A knowledge-based model can always obtain good results for specific domains, since the gazetteer can be adapted very precisely and the manually crafted rules are normally more accurate. However, it requires domain-specific knowledge. Changing the domain normally requires revising the rules and gazetteers – an effort which requires a great deal of time and energy.

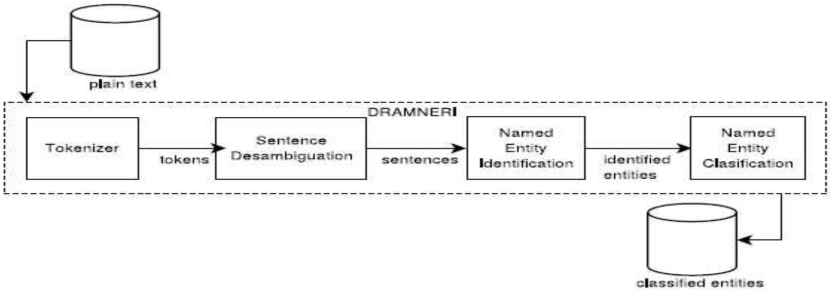


Fig. 1. DRAMNERI architecture

This solution is knowledge based, but almost all possible parameters are customizable. For example, the gazetteer, entity types, length of context, etc. can all be modified. DRAMNERI completely relies on domain knowledge; its rules and dictionaries are manually crafted. The disadvantage is that the KEC is very high, and requires a domain expert. The advantage is that it is very flexible and can be used in many domains. A high-level diagram of the solution is outlined in the above figure (Toral, 2005).

The built-in tokenizer can efficiently and correctly punctuate common texts in languages with Latin coding. (Other free tokenizers available could also be used to perform this work.) A sentence is divided into segments using an algorithm based on the method described in the EXIT system (Muñoz and Palomar, 1998). With the context information, rules and dictionaries are used to decide if the token is an end of sentence.

The Named Entity Identification (NEI) module identifies the named entities in the text for each sentence. Regular expressions are used to do this. Tokens that match any NEI regular expression joined by prepositions are identified as *generic entities.* Both the maximum number of prepositions between two tokens that match any NEI regular expression and the preposition list are configurable. For example, if *'de'* (Spanish: "of") and *'la'* (Spanish: "the") are in the preposition list, and the maximum number of prepositions between identified tokens is one, then the string *"en la Universidad de Alicante"* is identified as *"en la <ENTITY> Universidad de Alicante </ENTITY>*. But *"Pedro de la Viuda"* is identified as *"<ENTITY> Pedro </ENTITY> de la <ENTITY> Viuda </ENTITY>"* instead of *"<ENTITY> Pedro de la Viuda </ENTITY>*."

The Named Entity Classification module assigns a category to each of the entities detected in the previous step through the use of either internal or external evidence. This is done through an analysis of the entity itself and its left and right context. Trigger gazetteers are used to classify the entities through the use of external evidence. These dictionaries allow the entity class to be identified based on the appearance of a special word or words before or after an entity. Classification using internal evidence can be performed by gazetteers and rules. If an entity is in gazetteer, then its class is already known. Rules may also contain elements which refer to a gazetteer, and each rule is linked to an entity category.

Knowledge Engineering Cost

The KEC for this solution is very high, since both the rules and the dictionary are manually crafted. Since the solution is totally knowledge-based, it does not need labeled data, and no learning algorithm is used.

Summary Table

Category: Academic	
University Name: the University of Alicante **University URL**: http://www.ua.es/	**Location**: Alicante, Spain
Solution Name: DRAMNERI: A Free Knowledge Based Tool to Named Entity Recognition	
Domain Scope: general	**Application Type**: IE
Knowledge Engineering Cost: very high	**Financial Cost**: n/a
Input Requirements/Preparation Required: plain text	
Named Entity Extraction **Algorithm Name/Group**: manually crafted rules and dictionaries **Labeling**: n/a **Labeling Supervision**: n/a **Model Generation**: manual **Model Generation Supervision**: n/a **Process Description**: This solution depends on human-derived domain knowledge. The named entity identification rules are manually crafted, and the external information (such as a dictionary) is manually created. No labeled data is needed, and no learning algorithm is used.	
Solution Output: identifying named entities	
Application to Law Enforcement: moderate	
Is performance evaluation available? yes	**Solution/demo available**? yes

Sources

Muñoz, R. and Palomar, M. (1998). "Sentence Boundary and Named Entity Recognition in EXIT system: Information Extraction System of Notarial Texts." *Proceedings of IV Int. Conference on Artificial Intelligence and Emerging Technologies in Accounting*.

Toral, Antonio (2005). "DRAMNERI: A Free Knowledge Based Tool to Named Entity Recognition." *Proceedings of the 1st Free Software Technologies Conference. A Coruña, Spain.* pp. 27-32. July, 2005. Online: http://www.dlsi.ua.es/~atoral/publications/2005_fstc_dramneri_paper.pdf. Accessed January 19, 2006.

3.2.10 University of Arizona: Extracting Meaningful Entities from Police Narrative Reports

Solution Introduction and Domain Scope

This solution was developed by researchers at the University of Arizona, Tucson, AZ and is described in "Extracting Meaningful Entities from Police Narrative Reports" (Chau, 2002). The solution aims to automatically identify meaningful entities from police narrative reports, such as persona name, vehicle name, address, etc. It is primarily designed for deployment in the law enforcement field, but it can be tailored to serve the needs of other domains to extract named entities.

We have categorized this solution as an information extraction solution, since it only extracts named entities from narrative text; no relations are extracted.

Output/Results

The output is named entities. This solution can extract five types of entities: person, address, vehicle, narcotic drug, and personal property. For example for a sentence "Tom robbed a mobile phone from a person, and he drove a red Ford car," "Tom" is person name, "mobile phone" is the personal property and "red Ford" is the vehicle.

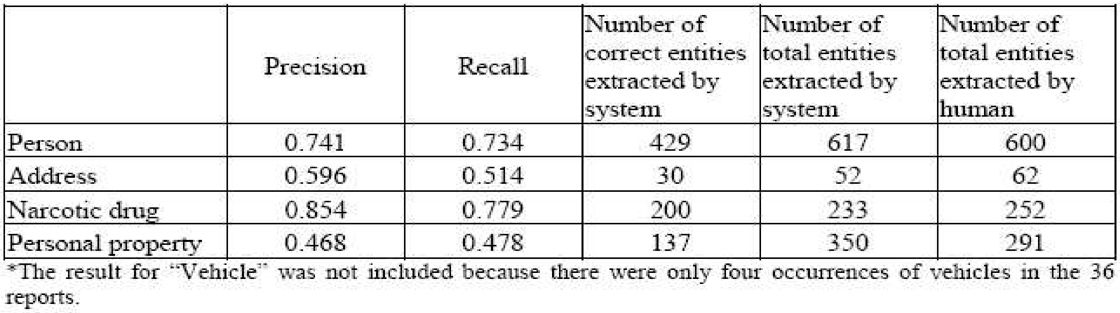
Application to Law Enforcement

Extensive. The solution has been specifically designed to be used in law enforcement applications. In the law enforcement field today, the amount of unstructured digital textual data is increasing. It is very helpful and important for crime investigation if the named entities can be extracted from the textual data. This solution is designed to meet this requirement and is currently being utilized as a first step in the CopLink® solution.

Evaluation

This solution was tested on the narrative reports from the Phoenix Police Department. The testbed consisted of 36 reports randomly selected from the Phoenix Police Department database for narcotic related crimes. A human expert manually labeled these reports, identifying all the entities. Those narrative reports were also noisy. In addition to containing typos, spelling errors, and other problems, they were also written entirely in uppercase letters (which excluded the use of letter case information in analyzing the text). This solution has been demonstrated to be robust on noisy data.

Three-fold cross validation was used to evaluate the system, and precision and recall were used to measure the performance. Although the results were not as good as those reported at the MUC conference, the results for the extraction of person name and narcotic drugs should still be considered good, especially considering the noisy data. However, the extraction for address was not as good as expected. One reason for this result can be attributed to mistakes made in the manually-crafted lexicons. The extraction for personal property is poor, which is expected, since it is difficult to identify. The results can be seen in the table below (Chau et al., 2002).



	Precision	Recall	Number of correct entities extracted by system	Number of total entities extracted by system	Number of total entities extracted by human
Person	0.741	0.734	429	617	600
Address	0.596	0.514	30	52	62
Narcotic drug	0.854	0.779	200	233	252
Personal property	0.468	0.478	137	350	291

*The result for "Vehicle" was not included because there were only four occurrences of vehicles in the 36 reports.

Table 1. Experimental results

Inputs Required

The input is narrative police reports, which could be noisy compared with other types of textual data, such as news articles. These sources often contain spelling errors, typos, etc.

Information Extraction Algorithm

The system employs a neural network. It combines lexical lookup, machine learning, and minimal hand-crafted rules. There are three major components:

1. Noun phrasing: a modified version of the Arizona Noun Phraser (Tolle & Chen 2000), this component extracts noun phrases from documents based on syntactical analysis. These noun phrases form the candidates for named entities.
2. Finite state machine and lexical lookup: a finite state machine is used to process the noun phrases identified from the first step. It compares each word in the phrase, as well as the words immediately before and after this word, with the words in the hand-crafted lexicons. Based on the comparison, a binary value (0/1) is generated to indicate whether a match occurs.
3. Neural network: for the feedforward/backpropagation neural network, the input is a phrase's corresponding binary values and the scores generated by the finite state machine. The generated output is the prediction for the phrase's most probable entity type.

The system operates in a *training state* or a *testing state*. In training state, the system identifies lexical rule patterns based on a training dataset. The training dataset is input-output pairs. The learned lexical patterns are stored as synaptic weights in the neural network. In the testing state, the system extracts phrases from test data and predicts the entity type for each phrase. As previously mentioned, this solution can extract five types of entities: person, address, vehicle, narcotic drug and personal property.

Knowledge Engineering Cost

We categorize the KEC as high since the solution needs labeled data to train the lexical rule patterns, as well as some hand-crafted rules. It also utilizes some hand-crafted lexicons. Compared to a system which only needs labeled data, the KEC is high.

Summary Table

Category: Academic	
University Name: University of Arizona **Lab URL**: http://ai.bpa.arizona.edu/index.html	**Location**: Tucson, AZ, USA
Solution Name: Extracting Meaningful Entities from Police Narrative Reports	
Domain Scope: law enforcement	**Application Type**: IE
Knowledge Engineering Cost: high	**Financial Cost**: n/a
Input Requirements/Preparation Required: Textual data, which could be noisy	
Named Entity Extraction **Algorithm Name/Group**: a neural network which also combines finite state machines, hand-crafted rules, and hand-crated lexicons. **Labeling**: manual **Labeling Supervision**: n/a **Model Generation**: hybrid **Model Generation Supervision**: supervised **Process Description**: First a Noun Phraser is used to extract noun phrases from documents. Then a finite state machine is used to process the noun phrases identified from the first step. A binary value (0/1) is generated for each word. Finally, a neural network uses the binary values to predict the phrase entity type.	
Solution Output: named entities	
Application to Law Enforcement: extensive	
Is performance evaluation available? yes	**Solution/demo available**? no

Sources

Chau, M., Xu, J. and Chen, H. (2002) "Extracting Meaningful Entities from Police Narrative Reports." *Proceedings National Conference for Digital Government Research, Los Angeles, CA.* Online. http://www.diggov.org/library/library/pdf/chau2.pdf. Accessed January 5, 2006.

3.2.11 The University of New South Wales: Information Extraction Using Two-Phase Pattern Discovery

Solution Introduction and Domain Scope

This solution has been developed by researchers at The University of New South Wales, Sydney, Australia. This solution aims to discover patterns for information extraction. These patterns can then be used on the previously unseen data to extract required information through pattern matching. It is suitable for many domains, but is limited to semi-structured documents. It is categorized as an information extraction solution, since it extracts named entities from semi-structured documents but does not determine relationships or links.

Output/Results

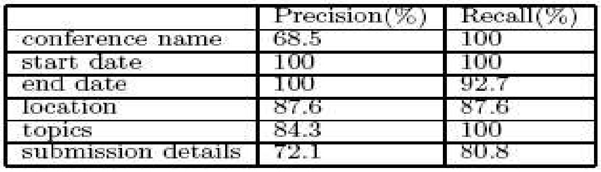
The output is patterns which are used to extract the desired information from previously unseen data. For example, the pattern "<city>, <country>, <conference>" describes how conference information is listed.

Application to Law Enforcement

Limited. The solution's input requires partially-structured (semi-structured) documents and is not able to handle plain text data. As most law enforcement documents are digital plain text data instead of semi-structured data, this solution is not well suited for use in law enforcement applications. However, in cases where semi-structured data requires analysis, this solution can be used in law enforcement.

Evaluation

Testing was performed on a sample of 178 PSLNL (Partially-structured, large-natural-language) documents, which were randomly selected from dbworld postings. By using techniques described below and in the authors' previous work, items were extracted from the data and automatically fitted into the slots of the schema. Then items were manually extracted from the data, and the two results were compared. The performance is presented in the above figure from Ma and Shepherd (2004).



	Precision(%)	Recall(%)
conference name	68.5	100
start date	100	100
end date	100	92.7
location	87.6	87.6
topics	84.3	100
submission details	72.1	80.8

Table 2: Data extraction results

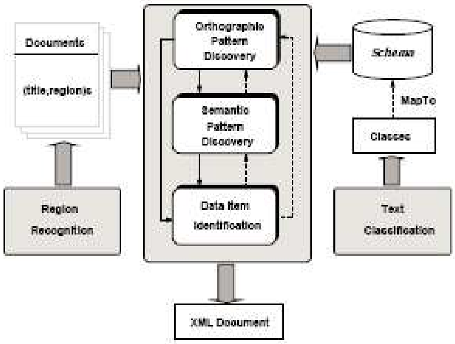


Figure 1: IE via two-phase pattern discovery

Inputs Required

Input must be PSLNL (Partially-structured, large-natural-language) documents, such as conference information and seminar announcements.

Information Extraction Algorithm

First documents are classified, and different user-defined schemas are assigned to classes of documents. Then, a region classifier is used to identify contiguous regions in the document which appear to be based on particular kind of information ("semantically-coherent regions" (Ma and Shepherd, 2004)). After the document has been classified and the regions have been identified, the pattern discovery technique is applied. Therefore, the input to the pattern discovery technique is the text of a region, the region title, and the schema of the document. The system is illustrated in the above figure from Ma and Shepherd (2004).

There are two types of patterns. First, *orthographic patterns* (OPD) are discovered, which determine the structural features from an identified region of a document. Then, *semantic patterns* are discovered, which are used to map the fields of orthographic pattern to the user predefined template (schema). This assigns a class to each field of the orthographic pattern.

The process of OPD can be described as follows:

1. *Identify names using dictionaries, dates and address via generic patterns.*
2. *Identify boundaries by searching for delimiters.*
3. *Partition the source into a sequence of interleaved delimiters and text slots.*
4. *Map each text slot to its representative tag (lexical category).*
5. *Reduce the line tag pattern to {tag}∪{delimiters}+;*
6. *Identify the most frequent-occurring pattern* P_m *in the region.*

Semantic pattern discovery aims to relate data item slots to the data components in the user-defined schema. A semantic pattern consists of a sequence of roles and orthographic delimiters. For example "<city>, <country>, <conferenceDate>" corresponds to the orthographic pattern "<SC>, <SC>, <Date>". After the patterns are discovered, they are used as a basis for identifying individual data items in text and for mapping them to an instance of the schema.

Knowledge Engineering Cost

It is difficult to estimate the KEC of this solution, since the method whereby the two types of patterns are discovered is not detailed for this solution. Nonetheless, the patterns are either manually crafted or learned from training data, which means data needs to be labeled. In addition, this solution uses some predefined dictionaries and also requires labeling of the data to perform testing – both of which increase the KEC. As a result, the total KEC for this solution is high.

Summary Table

Category: Academic	
University Name: The University of New South Wales **University URL**: http://www.unsw.edu.au/	**Location**: Sydney, Australia
Solution Name: Information Extraction Using Two-Phase Pattern Discovery	
Domain Scope: general	**Application Type**: NE IE
Knowledge Engineering Cost: high	**Financial Cost**: n/a
Input Requirements/Preparation Required: semi-structured documents	
Named Entity Extraction **Algorithm Name/Group**: Combination of orthographic pattern discovery and semantic pattern discovery **Labeling**: manual **Labeling Supervision**: n/a **Model Generation**: automatic **Model Generation Supervision**: supervised **Process Description**: Documents are first classified and assigned a user-defined schema. Then, a	

region classifier is used to identify contiguous regions in the document. After these steps, the pattern discovery technique is applied to discover patterns. These patterns can be used to extract information from previously unseen data.	
Solution Output: pattern which could be used to extract information from previously unseen data	
Application to Law Enforcement: limited	
Is performance evaluation available? yes	**Solution/demo available**? no

Sources

Ma, Liping and Shepherd, John (2004). "Information Extraction Using Two-Phase Pattern Discovery." *SIGIR'04 Sheffield, South Yorkshire, UK.* July 25–29, 2004. Online. http://delivery.acm.org/10.1145/1010000/1009107/p534-ma.pdf?key1=1009107&key2=5107656311&coll=GUIDE&dl=GUIDE&CFID=61576327&CFTOKEN=9622293. Accessed January 10, 2006.

3.2.12 University of Sheffield: Adaptive Information Extraction from Text by Rule Induction and generalization

Company Introduction and Domain Scope

This solution was developed by researchers at the University of Sheffield, UK. It extracts information from text and could be used in many domains. This solution has been categorized as an information extraction solution.

Output/Results

The output is tagging rules, which are used to extract named entities. A tagging rule is composed of a pattern of conditions and an action inserting an SGML tag into the text.

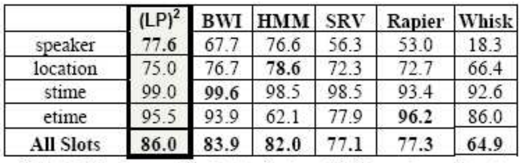
Application to Law Enforcement

Moderate. This solution was not specifically designed for the law enforcement field. But as with many other IE solutions, it could be used in law enforcement to aid in the transformation of narrative police data into a structured format. Changes would be required to the solution, however, to make this possible.

Evaluation

This solution was tested on tasks in two languages: English and Italian. The corpus was divided into training and test sets; the learning algorithm was trained and tested on the appropriate sets. Two experiments were performed: the CMU seminar announcements (containing 485 seminar announcements) and the Austin job announcements (containing 300 job announcements). In the first experiment, the named entities to be extracted were speaker name, starting time, ending time and location. In the second experiment, the named entities were ID, job title, salary, company, recruiter, state, city and country, programming language and a few others.

F-measure (ß=1) was the metric used to evaluate the performance. In the seminar announcement task, this solution had the best performance compared to five other solutions, as presented in Table 5 (Ciravegna, 2001) reproduced below. In the job announcement task, this solution also recorded the best result, as presented in Table 6 (Ciravegna, 2001) below. (LP)^2 is the learning algorithm used in this solution.

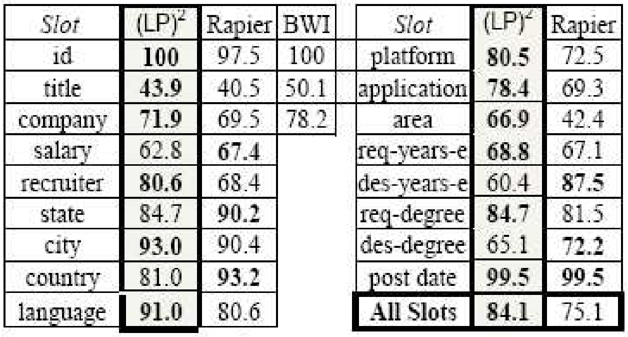


	$(LP)^2$	BWI	HMM	SRV	Rapier	Whisk
speaker	**77.6**	67.7	76.6	56.3	53.0	18.3
location	75.0	76.7	**78.6**	72.3	72.7	66.4
stime	99.0	**99.6**	98.5	98.5	93.4	92.6
etime	95.5	93.9	62.1	77.9	**96.2**	86.0
All Slots	**86.0**	**83.9**	**82.0**	**77.1**	**77.3**	**64.9**

Table 5: F-measure (β=1) obtained on CMU seminars. Results for algorithms other than $(LP)^2$ are taken from [Freitag and Kushmerick 2000]. We added the comprehensive ALL SLOTS figure, as it allows better comparison among algorithms. It was computed by:

$$\frac{\Sigma_{slot}\,(\text{F-measure} * \text{number of possible slot fillers})}{\Sigma_{slot}\,\text{number of possible slot fillers}} *100$$

Concerning $(LP)^2$ results from a 10 cross-folder experiment using half of the corpus for training. F-measure calculated via the MUC scorer [Douthat 1998]. Average training time per run: 56 min on a 450MHz computer. Window size *w*=4.



Slot	$(LP)^2$	Rapier	BWI	*Slot*	$(LP)^2$	Rapier
id	**100**	97.5	100	platform	**80.5**	72.5
title	**43.9**	40.5	50.1	application	**78.4**	69.3
company	**71.9**	69.5	78.2	area	**66.9**	42.4
salary	62.8	**67.4**		req-years-e	**68.8**	67.1
recruiter	**80.6**	68.4		des-years-e	60.4	**87.5**
state	84.7	**90.2**		req-degree	**84.7**	81.5
city	**93.0**	90.4		des-degree	65.1	**72.2**
country	81.0	**93.2**		post date	**99.5**	**99.5**
language	**91.0**	80.6		**All Slots**	**84.1**	75.1

Table 6: F-measure (β=1) obtained on the Jobs domain using half of the corpus for training.

Software

$(LP)^2$ is only a research prototype, but an industrial system, LearningPinocchio, which was based on $(LP)^2$ has been developed and used in industrial applications.

Inputs Required

The input is textual data, such as job and seminar announcements.

Information Extraction Algorithm

This solution presents an adaptive IE algorithm $(LP)^2$. $(LP)^2$ is a covering algorithm that uses shallow NLP in order to overcome issues with data sparseness. $(LP)^2$ needs a training corpus *t* to learn the tagging rules. In the training corpus, the user tags entities with SGML tags to highlight the information to be extracted. $(LP)^2$ induces symbolic rules that insert SGML tags into texts. The rules are induced in two steps:

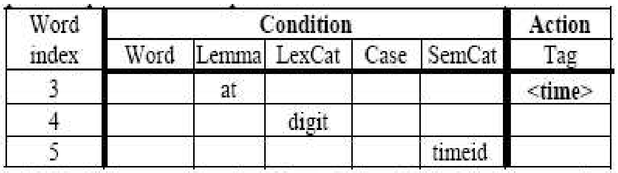
1. Tagging rules are induced by bottom-up generalization of tag instances in the training corpus. In the generalization process, shallow NLP is used.

2. Correction rules are induced. Corrections rules correct the mistakes and imprecision of the previous tagging rules.

A tagging rule is composed of a left and side and right hand side. The left hand side contains a pattern of conditions, while the right hand specifies the action of inserting an SGML tag in the text. Tagging rule induction is learned from the positive examples in the training corpus. Positive examples refer to the SGML tags inserted by the user, and the remainder of the corpus consists of negative examples. For each positive example three steps are taken. First, an initial rule is built. Then, the initial rule is generalized. Finally, the k best generalizations of the initial rule are kept.

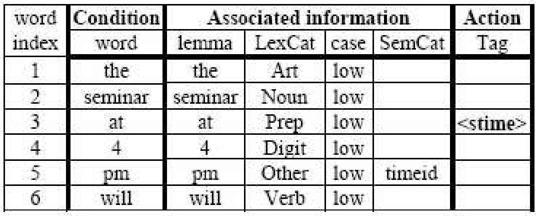
In the generalization process, $(LP)^2$ uses a shallow approach to Natural Language Processing, utilizing a morphological analyzer and a part-of-speech tagger. It also uses a dictionary predefined by the user. A lexical item is used to summarize knowledge about every word, determining such attributes as word's the lexical category (noun, verb, digit, etc.) and case. Table 1 (Ciravegna, 2001) below is an initial rule and associated information.

The purpose of the rule generalization process is to relax constraints in the initial rule pattern, "which could be done both by reducing the pattern in length and by substituting constraints on words with constraints on some parts of the additional knowledge". One generalization of the rule presented in Table 1 is shown in Table 2 (Ciravegna, 2001), reproduced below.



Word index	Condition					Action
	Word	Lemma	LexCat	Case	SemCat	Tag
3		at				<time>
4			digit			
5					timeid	

Table 2: A generalization for rule in table 1. The pattern is relaxed in length (conditions on words 1, 2 and 6 were removed) and conditions on the other words were substituted by other constraints.



word index	Condition	Associated information				Action
	word	lemma	LexCat	case	SemCat	Tag
1	the	the	Art	low		
2	seminar	seminar	Noun	low		
3	at	at	Prep	low		<stime>
4	4	4	Digit	low		
5	pm	pm	Other	low	timeid	
6	will	will	Verb	low		

Table 1: Starting rule (with associated NLP knowledge) inserting <stime> in the sentence ``the seminar at <stime> 4 pm will...".

The tagging rules may report some imprecision in the slot filler boundary detection when they are applied to the testing corpus. For example, a typical mistake is *"at <time> 4 </time> pm"*, where "pm" should be part of the time expression. To solve this problem, $(LP)^2$ induces rules for shifting the wrongly-placed tag to the correct position.

Knowledge Engineering Cost

We conclude the KEC for this solution is high, since it requires labeled training and testing data. No method is used to reduce the labeling effort. The rule generation module is automatic and does not require manual effort.

Summary Table

Category: Academic	
University Name: University of Sheffield **University URL**: http://www.shef.ac.uk/	**Location**: Sheffield, UK
Solution Name: Adaptive Information Extraction from Text by Rule Induction and generalization	
Domain Scope: general	**Application Type**: IE
Knowledge Engineering Cost: high	**Financial Cost**: n/a
Input Requirements/Preparation Required: Textual data	
Named Entity Extraction **Algorithm Name/Group**: a covering algorithm **Labeling**: manual **Labeling Supervision**: n/a **Model Generation**: automatic **Model Generation Supervision**: supervised **Process Description**: $(LP)^2$ is a covering algorithm for information extraction from text. It induces rules by learning from positive examples in the training corpus. First, tagging rules are induced by bottom-up generalization of positive examples from the training corpus. Then, correction rules are induced to correct the mistakes and imprecision of the previous tagging rules.	
Solution Output: tagging rule	
Application to Law Enforcement: moderate	
Is performance evaluation available? yes	**Solution/demo available**? no

Sources

Ciravegna, Fabio (2001). "Adaptive Information Extraction from Text by Rule Induction and Generalisation." *Proceedings 17th International Joint Conference on Artificial Intelligence (IJCAI 2001), Seattle, WA*. Online: http://www.dcs.shef.ac.uk/~fabio/paperi/IJCAI01.pdf. Accessed January 19, 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).

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).

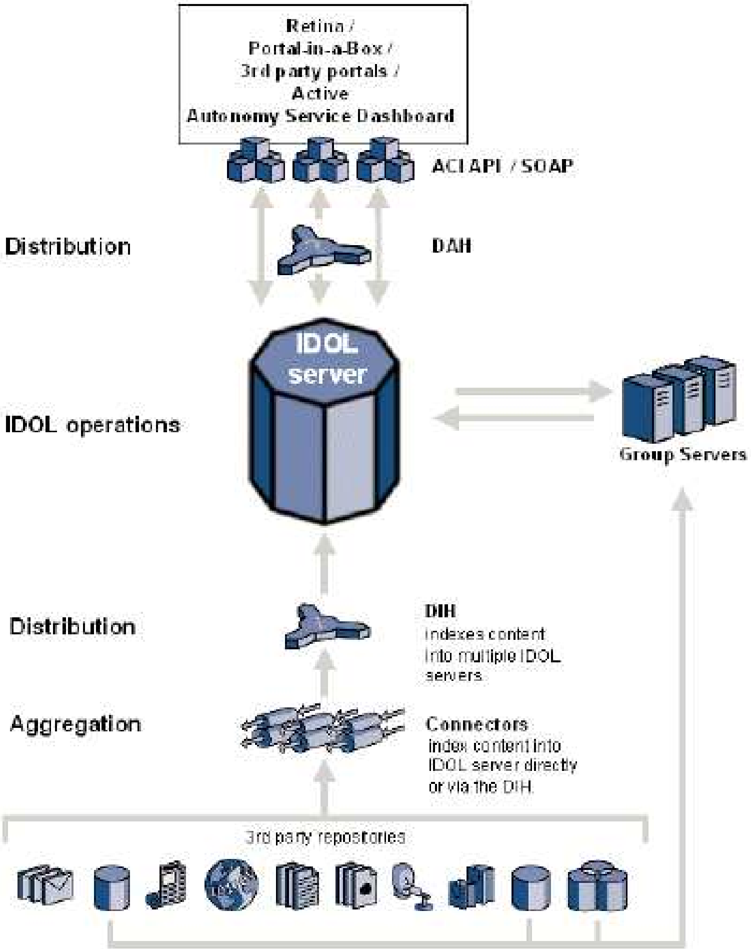
Financial

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).

Information Extraction 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.	
Information Extraction **Algorithm Name/Group**: concept extraction through the use of Shannon's Information Theory (entropy) and Bayesian Inference (probabilistic models) **Labeling**: hybrid **Labeling Supervision**: n/a	

Model Generation: automatic **Model Generation Supervision**: unsupervised **Process Description**: Concepts are labeled in text through the use of context and an automatic XML tagger/integrator.	
Solution Output: Autonomy Content Infrastructure™ (ACI™) standardizes the data format so that it can be communicated. SOAP can also be used.	
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.

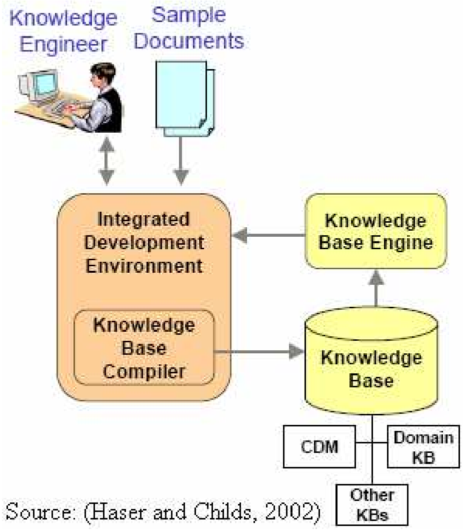
Inputs Required

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

Information Extraction 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).

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



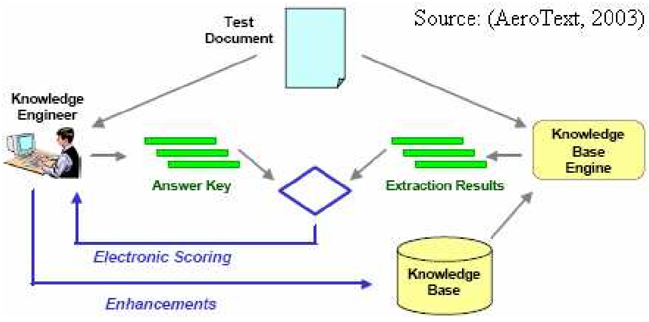
Source: (Haser and Childs, 2002)

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.

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).



Source: (AeroText, 2003)

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)	
Information Extraction **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

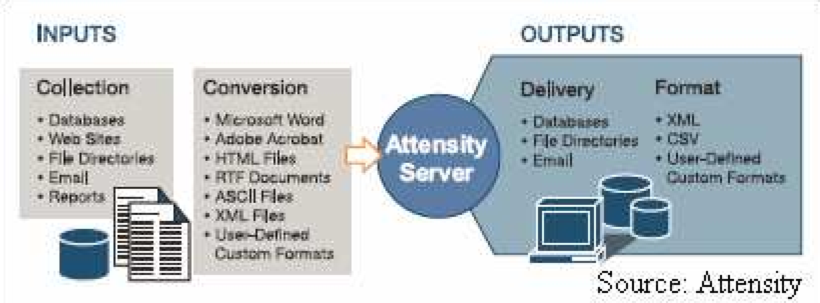
Company Introduction and Domain Scope

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



Source: Attensity

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.

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

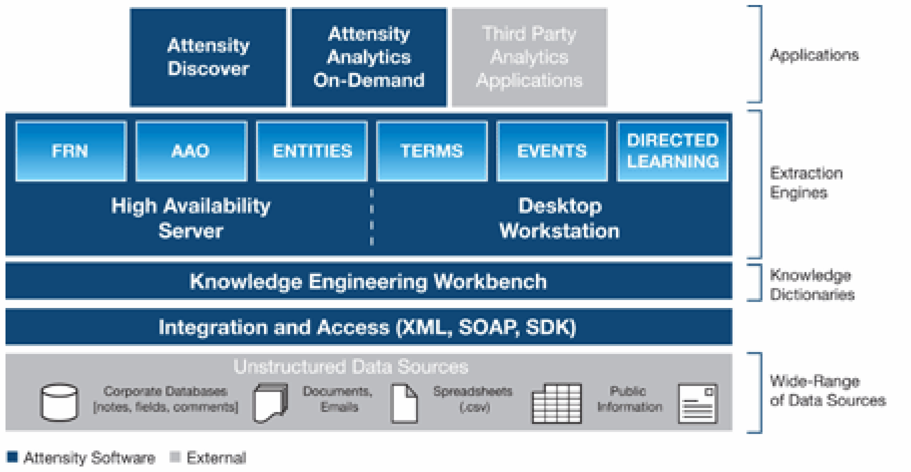
In a company white paper (Attensity 2005b), Attensity claimed that they "made a fundamental breakthrough in converting unstructured text into structured tables with 95% or better accuracy (precision + recall)." Using a 1GHz Intel CPU, the solution can process high raw text at a rate of 5MB/minute. The 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.

Information Extraction 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" (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.	
Information Extraction **Algorithm Name/Group**: proprietary **Labeling**: manual **Labeling Supervision**: n/a **Model Generation**: active **Model Generation Supervision**: supervised **Process Description**: After the labeling has been completed, the Automatic rule generation is carried out in a supervised manner.	

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

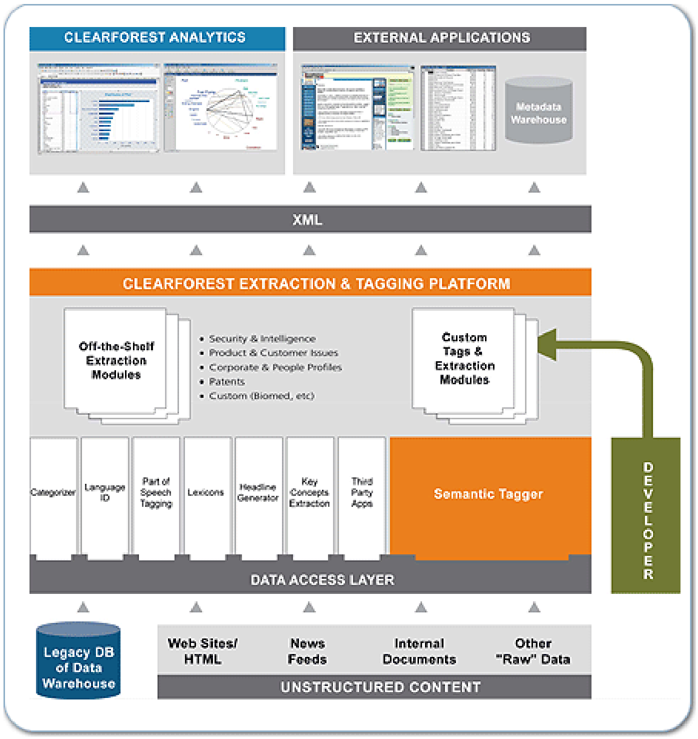
As mentioned in (Wu and Pottenger, 2005a), ClearForest participated in the 2002 KDD Challenge Cup competition in biomedical domain (Regev, et. al, 2002). During this competition, F-measure scores of 78% and 67% were achieved in the Document Curation task and the Gene Product task, respectively.

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.

Information Extraction 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. *Descriptive tags* “provide information on the meaning and type of words and phrases in the text” (ClearForest, a) by learning such entities as “person”, “company,” “industry,” or even “organ,” “medication,” and “disease.” *Factual tags* provide intrasource link analysis (Pottenger et al., 2006b). The last category of tags is *role tags*, which identify textual regions or portions within a document such as “title” or “author.”

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, a), and structural tagging (based on typographic and positional characteristics). Descriptive tags require no labeling of the data, but rather use manually crafted rules to extract the information as part of its semantic tagging process. Role tags also use information extraction technology, but require labeling of the data with a Manual/Active procedure (i.e., "learn by example"). The rule generation is then automatic and supervised.

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	
Company Name: ClearForest 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: medium/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.	
Information Extraction **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 crafted rules to extract entities from 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. (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.

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.

Output/Results

Ranges of characters, structured sets of morphemes, words, phrases, and text are all extracted with Delphes' technology.

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 Sciullo 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.

Information Extraction 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	
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.	
Information Extraction **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**: Delphes' integrated information system works to determine words' contextual purpose through the use of configurational analysis. The text is parsed and analyzed morphologically through the comparison to specialized dictionaries. Context is analyzed to form constituents and the information is indexed to provide fast retrieval.	
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: moderate	

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.

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

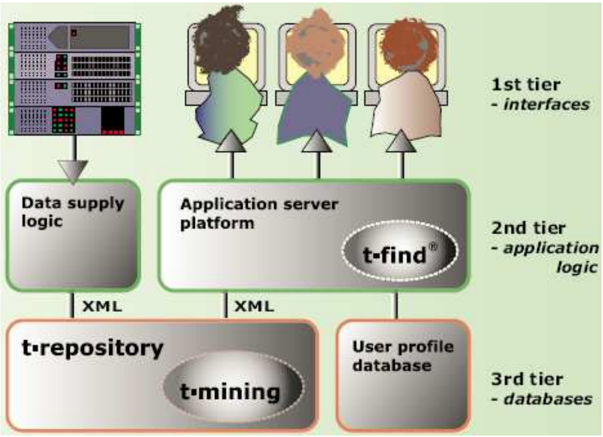
No evaluation information was available.

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.

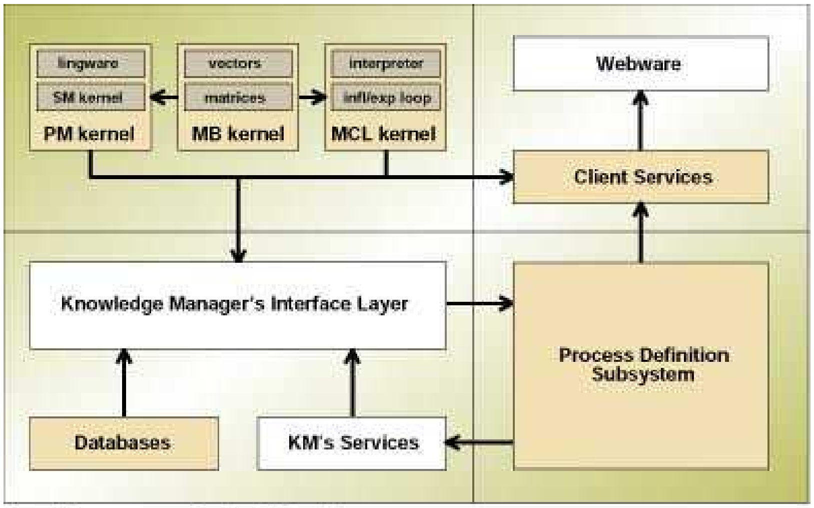
A demo of t-find® was available at http://cwi-opac.eidetica.com/ but was not active at the time of this survey. The company's language guesser has a demo at http://www.eidetica.com/services/guesser.

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).

Information Extraction 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 manually crafted rules, yet does involve interactive training coupled with the use of neural networks, the KEC appears to be medium.

Summary Table

Category: Commercial	
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)	
Information Extraction **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**: Advanced methods to extract subject keywords inside documents and titles are used such as matrices, linguistic processing techniques, and fixed keyword lists. Then, the information is stored as XML within the Eidetica database.	
Solution Output: XML-formatted data in the Eidetica database	
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

The data converted is stored within the Endeca Data Foundry and passed via XML.

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

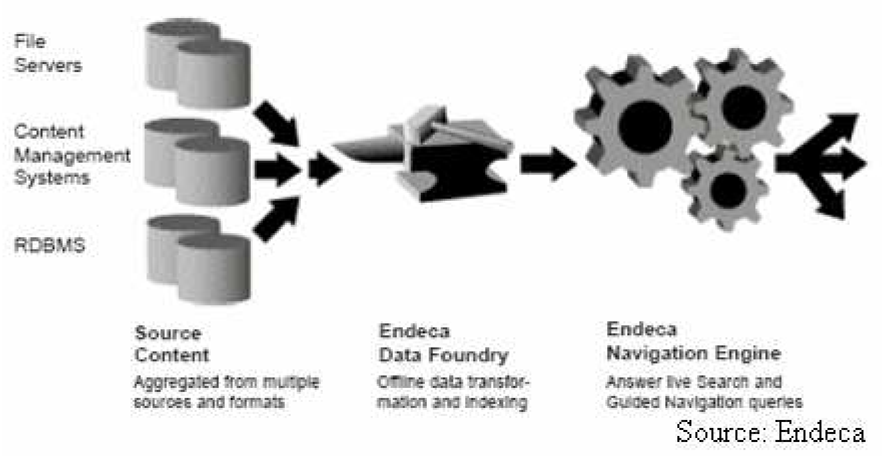
No detailed performance evaluations were found, although it is claimed that World Book experienced an increase in search speeds by a factor of 8 – 10 times (Endeca). However, "current deployments [of the Endeca Navigation Engine] scale to over a billion records, terabytes of contents, thousands of facets [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.

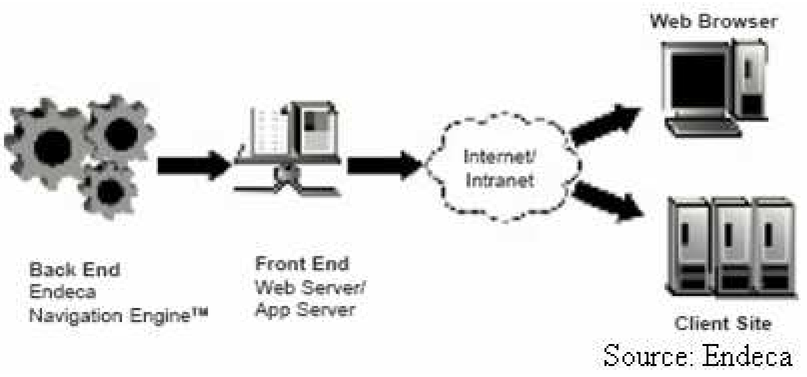
Software

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 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


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

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

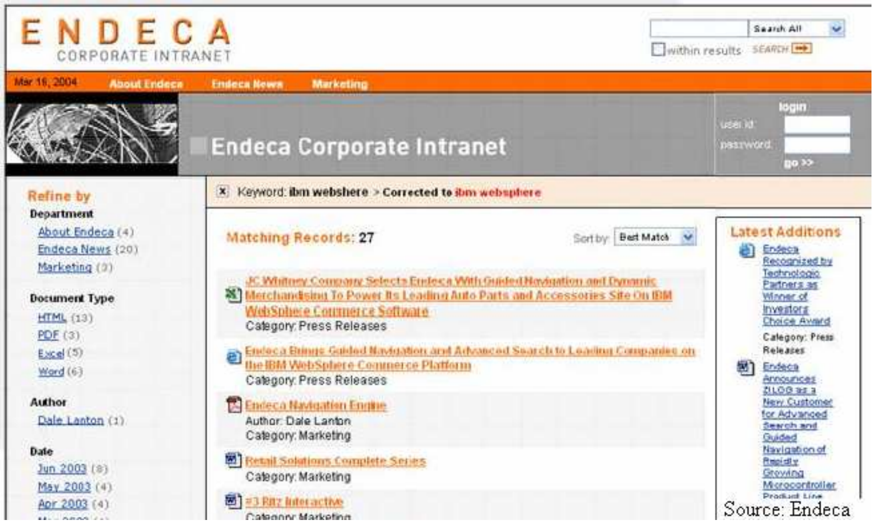
The Navigation Engine can access over 370 different file formats and supports over 250 languages. While much is done automatically, the solution can also be configured to enhance and refine search options, relevancy ranking modules, and business rules in the formation of links.

Configurations are performed using scripts (Perl, etc.), ODBC connections, as well as text and XML files.

Information Extraction 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).

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

Matching Records: 27

Sort by: Best Match

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)

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

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

Given that the unstructured information extraction primarily utilizes ClearForest's technology, the approach has a medium to high KEC. This is primarily due to the effort needed to manually craft extraction rules.

Summary Table

Category: Commercial	
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: medium/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.	
Information Extraction **Algorithm Name/Group**: proprietary **Labeling**: n/a **Labeling Supervision**: n/a **Model Generation**: manual **Model Generation Supervision**: n/a **Process Description**: ClearForest technology is utilized.	
Solution Output: The data converted is stored within the Endeca Data Foundry and passed via XML.	
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 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.

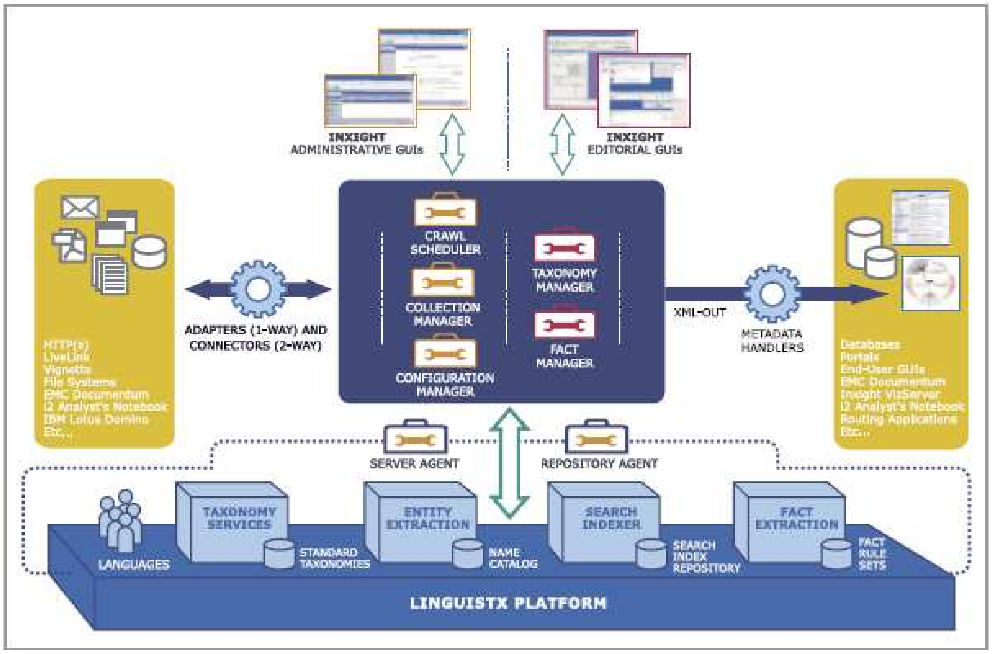
Financial

No details about the cost of the solution were found.

Software

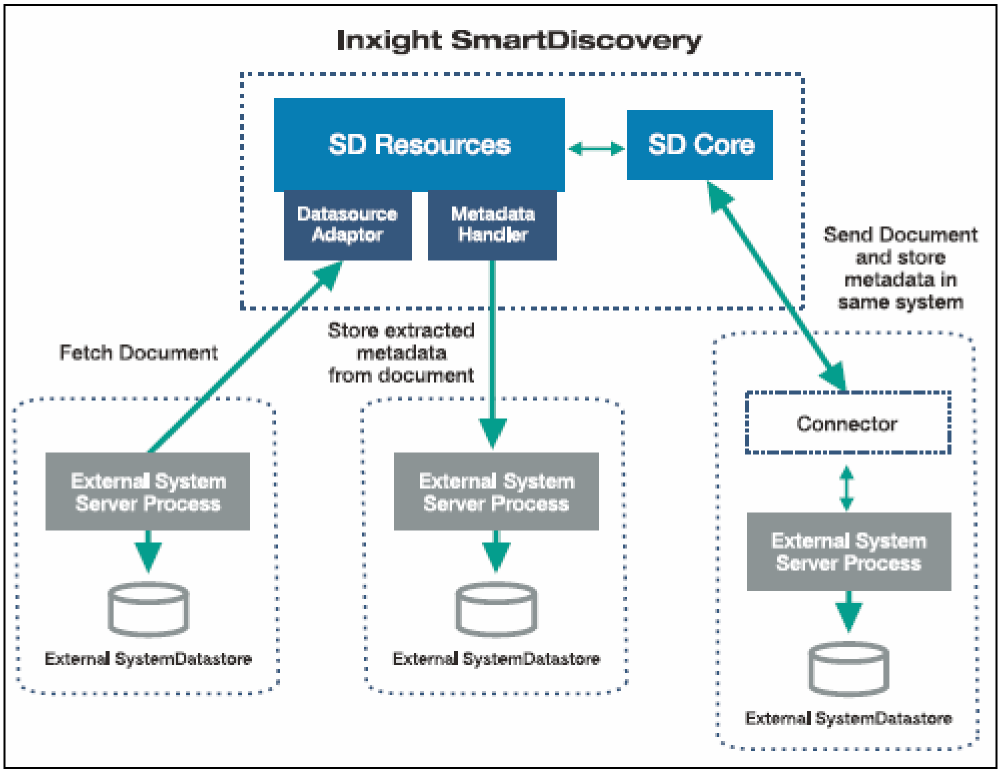
Inxight offers a suite of tools for information extraction and link analysis. The company's flagship product, SmartDiscovery® incorporates several components that are also available individually.

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 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.


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

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.

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.

Information Extraction 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).

Knowledge Engineering Cost

While the company claims that it's SmartDiscovery entity extraction component (i.e., ThingFinder) performs its work "[w]ithout training sets or manually created rules," this is true only for the end user that needs to new entity types defined. However, for the user who wishes to define their own entity types and also for Inxight's creation of the 27 entity types available in the system, the manual creation of rules is necessary. Given this, the KEC of Inxight's IE process is high.

Summary Table

Category: Commercial	
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	
Information Extraction **Algorithm Name/Group**: proprietary (LinguistX) **Labeling**: n/a **Labeling Supervision**: n/a **Model Generation**: manual **Model Generation Supervision**: n/a **Process Description**: Entity types are extracted via rules that are manually created, either by Inxight or by the user as a custom entity type.	
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.

3.3.9 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.

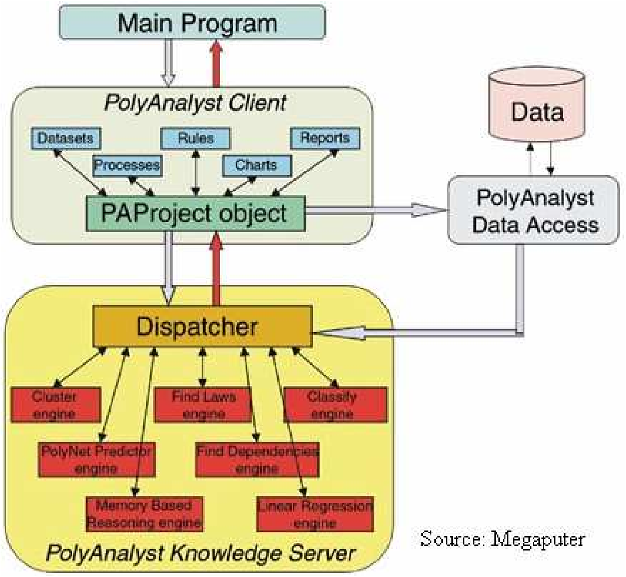
Financial

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 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.



Source: Megaputer

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).

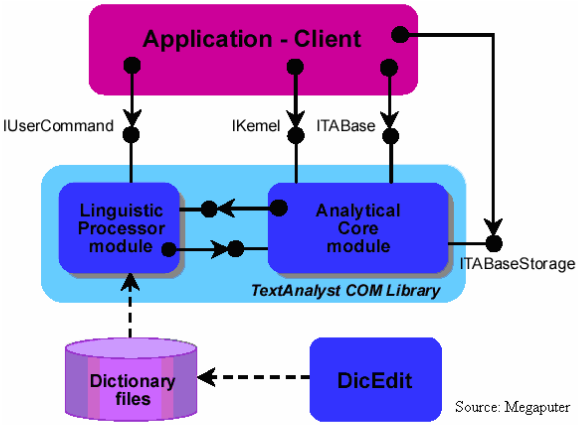
Megaputer offers 30 day demos of nearly all of its offerings at http://www.megasysdev.com/webdown/prodlist with registration. Microsystems offers the TextAnalyst SDK at http://www.analyst.ru/index.php?lang=eng&dir=content/products/.

Inputs Required

The "language independent" TextAnalyst is designed to work with any alphabet-based language. Currently, the solution is provided with dictionaries for English, French, Spanish, German, Italian, Dutch, and Russian (Megaputer).

Information Extraction Algorithm

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).



Source: 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.

Knowledge Engineering Cost

Given the dictionary-based information extraction algorithm coupled with the use of neural networks (and the implied need for labeled training data), the TextAnalyst has a high KEC. This is also due to the fact that the system does not provide any sort of algorithm to assist the user in the extraction of values and simply utilizes term matching.

Summary Table

Category: Commercial	
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: high	**Financial Cost**: $2,300 to $14,900; $16,000 (developer kit) (from 2000)
Input Requirements/Preparation Required: TextAnalyst is designed to work with any alphabet-based language.	
Information Extraction **Algorithm Name/Group**: proprietary **Labeling**: n/a **Labeling Supervision**: n/a **Model Generation**: hybrid **Model Generation Supervision**: supervised **Process Description**: Values are processed using a combination of linguistic and semantic text analysis and a Hopfield-like neural network.	
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.

3.3.10 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

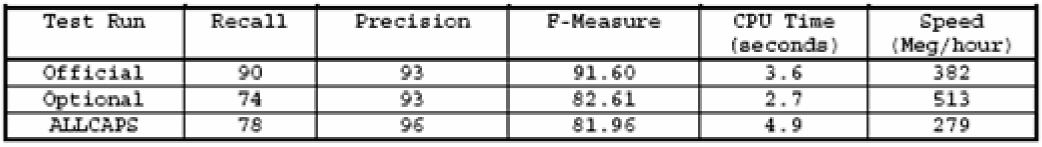
Extensive. NetOwl software originated in the 1990s for work specifically in the government domain. While the SRA subsidiary IsoQuest, Inc. was formed in 1996 to understand the market potential for their technology, government applications have remained a focus for NetOwl technology. NetOwl technology "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" (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

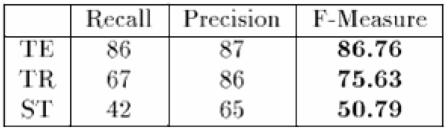
Table: NE Test Results Source: (Krupka and Hausman, 1998)

"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.).

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

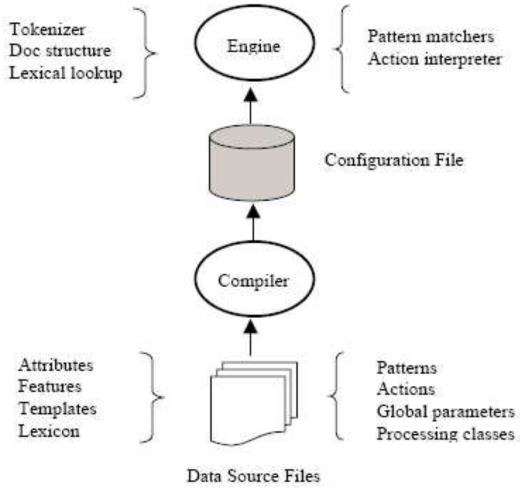


Figure 2: NetOwl Extractor Configuration Compiler

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	
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	
Information Extraction **Algorithm Name/Group**: proprietary **Labeling**: manual **Labeling Supervision**: n/a **Model Generation**: manual **Model Generation Supervision**: n/a **Process Description**: Manually-crafted rules are used to identify entities based on the use of lexicons and pattern rule bases. Then, a template is used to carry out the extraction process.	
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.

3.3.11 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

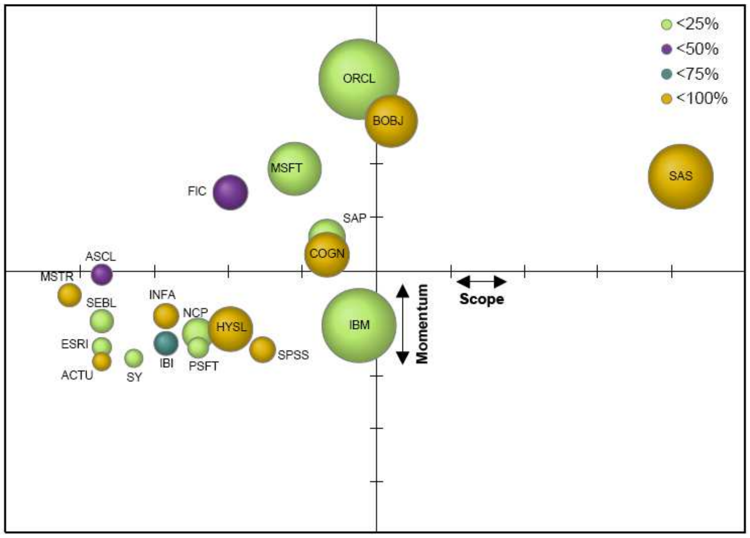
TextMiner has several types of output. The extracted and transformed data is stored within the system as an SAS dataset, while reports can be published in HTML. Additionally, process flow diagrams can also be modified, saved, and shared with others (SAS, 2005e).

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



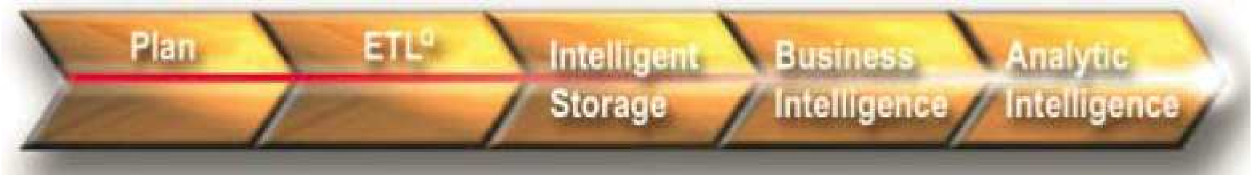
IDC conducted a survey of the business analytics (BA) software market in 2003 (Vesset and Morris, 2004) to evaluate the performance of a variety of companies, of which SAS was included. The companies were ranked on four axes: size (worldwide license and maintenance revenue of BA software), momentum (size-adjusted 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

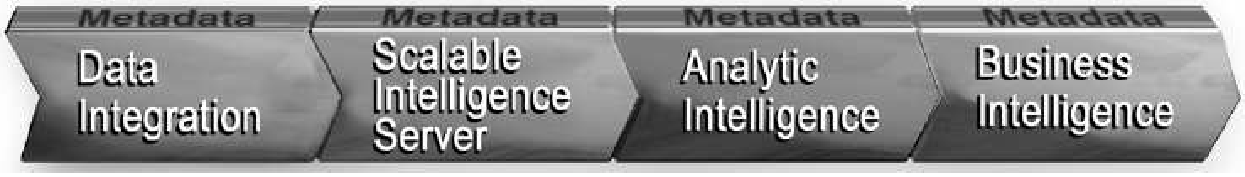
Little detail of the cost of SAS components and solutions is available. However, Charlesworth (2005) states that SAS's Marketing Optimization solution is "typically purchased by companies with in excess of 250,000 customers" and goes on to say that SAS claims "that the solution will ordinarily pay for 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

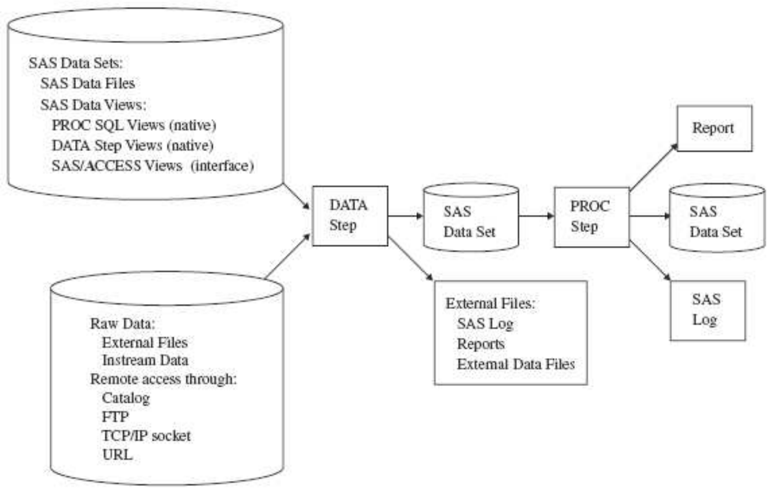
The Text Miner solution "combines a variety of information sources, including text and traditional databases" (SAS, 2005e) and can handle a wide variety of textual data formats, including PDF, extended ASCII, HTML, MS Word, and WordPerfect. Web crawling capabilities are also available. Customized routines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

Information Extraction 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). According to (SAS, 2003b), the Text Miner solution also incorporates Inxight's LinguistX and ThingFinder solutions. The extent to which SAS utilizes Inxight's technology is not apparent from the publicly available literature.

The company also goes into detail on 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 information extraction, little detail is given as to how the entities are extracted. However, as the use of Inxight technology is acknowledged, the methodology which SAS uses to extract the data is also believed to be based primarily on manually crafted rules, etc. Therefore, the KEC of the information extraction portion is high.

Summary Table

Category: Commercial	
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: high	**Financial Cost**: unknown
Input Requirements/Preparation Required: Text Miner: various textual formats in several languages	
Information Extraction **Algorithm Name/Group**: proprietary **Labeling**: n/a **Labeling Supervision**: n/a **Model Generation**: manual **Model Generation Supervision**: n/a **Process Description**: Text Miner uses a three-step process: *accessing the unstructured text*, *parsing the text and turning it into actionable data*, and *analyzing the newly created data*. A GUI allows the user to automatically extract and analyze the data through defined rules.	
Solution Output: TextMiner data is stored as an SAS dataset, while reports are HTML. Process flow diagrams can also be saved.	

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.

3.3.12 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."

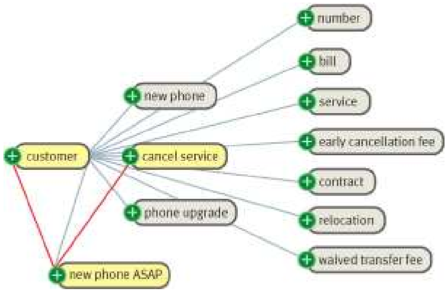
Please visit http://www.spss.com/estore/softwaremenu/index.cfm for more pricing information.

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).

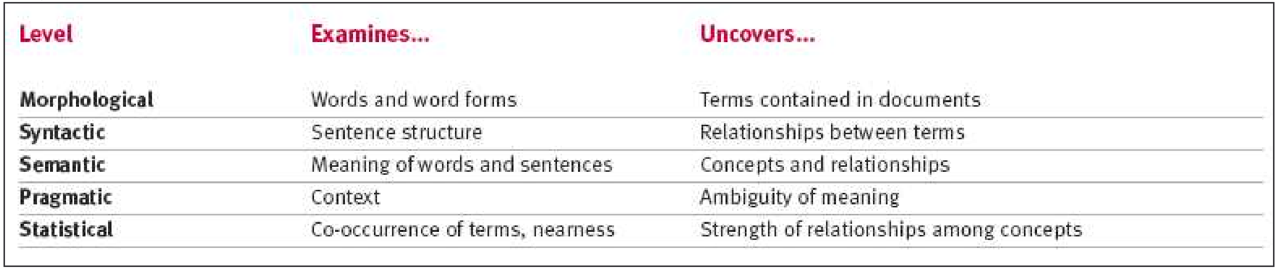
Information Extraction Algorithm

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).

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/.

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	
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/high	**Financial Cost**: various ($199 - $7,452 for components)
Input Requirements/Preparation Required: textual data	
Information Extraction **Algorithm Name/Group**: CLEM language **Labeling**: n/a **Labeling Supervision**: n/a **Model Generation**: manual **Model Generation Supervision**: n/a **Process Description**: Rules are developed in the CLEM language and users are assisted by a graphical expression builder.	
Solution Output: database entries	
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 (2005a). 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 (2005a), 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.

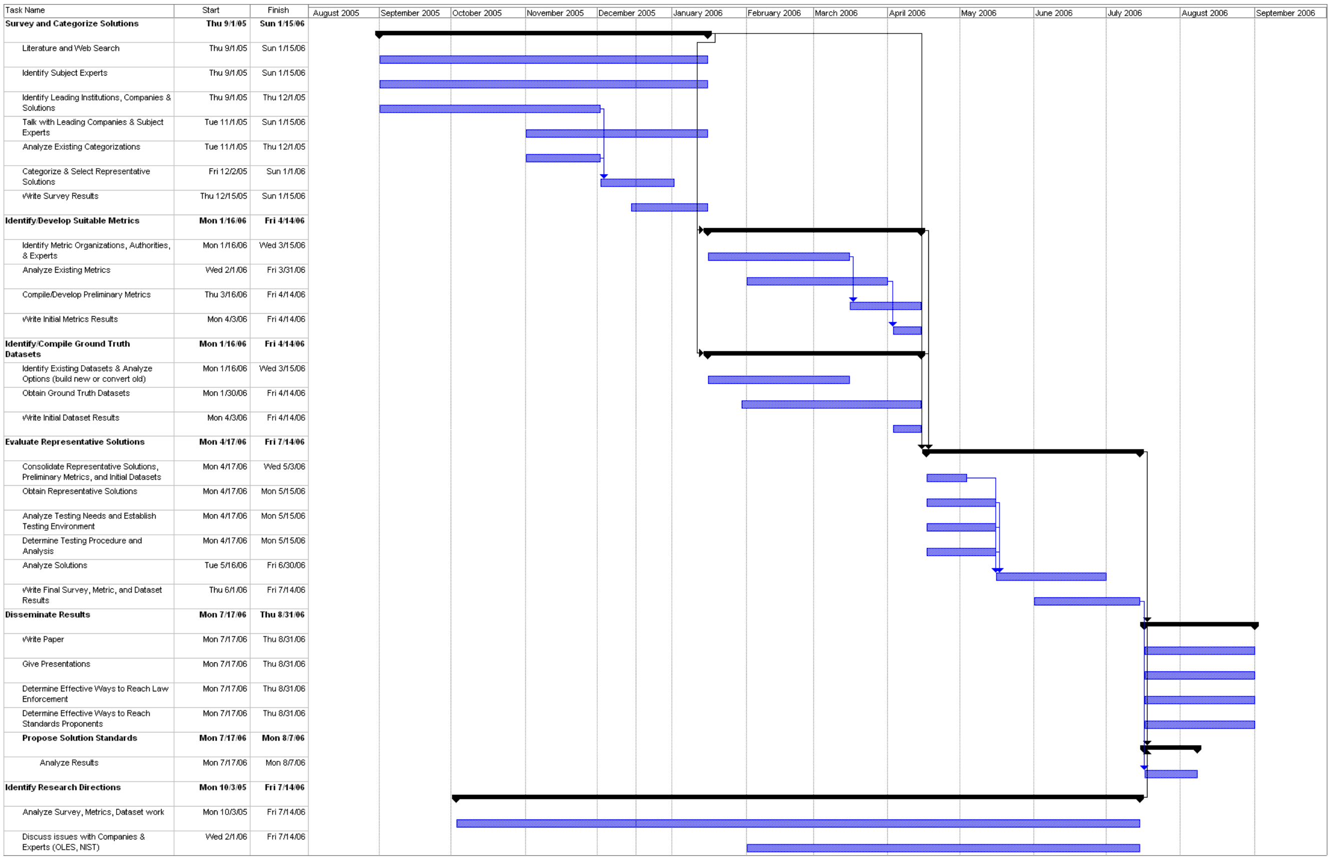
As mentioned in Section 1.2, our survey method calls for the accomplishment of seven steps. To date, we have successfully completed the first, and perhaps most extensive, portion of our survey effort: the survey of the information extraction field and the organization of the solutions into categories. 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 2005-93045-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

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.

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.

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.

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.

Borkar, V; Deshmukh, K.; and Sarawagi, S (2001). "Automatic segmentation of text into structured records." *Proceedings of SIGMOD, 2001*. Online: http://citeseer.ist.psu.edu/cache/papers/cs/26886/http:zSzzSzranger.uta.eduzSz~alpzSzixzSzreadingszSzp175-borkar-auto-classify-text-into-structured-records.pdf/borkar01automatic.pdf. Accessed January 18, 2006.

Brants, Thorsten (2000). "TnT-A Statistical Part-of-Speech Tagger." *Proceedings of 6th Applied NLP Conference, ANLP-2000. Seattle, Washington*. Online. http://citeseer.ist.psu.edu/cache/papers/cs/26650/http:zSzzSzacl.ldc.upenn.eduzSzAzSzA00zSzA00-1031.pdf/brants00tnt.pdf. Accessed January 26, 2006.

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.

Callan, Jamie and Mitamura, Teruko (2002). "Knowledge-Based Extraction of Named Entities." *CIKM'02, McLean, Virginia*. November 4–9, 2002. Online. http://delivery.acm.org/10.1145/590000/584880/p532-callan.pdf?key1=584880&key2=6419338311&coll=GUIDE&dl=GUIDE&CFID=63537399&CFTOKEN=50211284. Accessed January 27, 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.

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.

Chau, M., Xu, J. and Chen, H. (2002) "Extracting Meaningful Entities from Police Narrative Reports." *Proceedings National Conference for Digital Government Research, Los Angeles, CA*. Online. http://www.diggov.org/library/library/pdf/chau2.pdf. Accessed January 5, 2006.

Chieu, Hai Leong and Ng, Hwee Tou (2003). "Named Entity Recognition with a Maximum Entropy Approach." *Proceedings of the Seventh Conference on Natural Language Learning (CoNLL-2003), Edmonton, Alberta, Canada* pp. 160-163. Online. http://www.comp.nus.edu.sg/~nght/pubs/conll03.pdf. Accessed January 26, 2006.

Churches, Tim; Christen, Peter; Lim, Kim and Xi Zhu, Justin (2002). "Preparation of Name and Address Data for Record Linkage Using Hidden Markov Models." *BMC Medical Informatics and Decision Making, 2: 9*. Online. http://www.biomedcentral.com/1472-6947/2/9. Accessed January 27, 2006.

Ciravegna, Fabio (2001). "Adaptive Information Extraction from Text by Rule Induction and Generalisation." *Proceedings 17th International Joint Conference on Artificial Intelligence (IJCAI 2001), Seattle, WA*. Online: http://www.dcs.shef.ac.uk/~fabio/paperi/IJCAI01.pdf. Accessed January 19, 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.

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.

Diaz, Elena Viñuela (2004). *A Query System for UNESCO's World Heritage at the WWW.* April 30, 2004. Online. http://www2.imm.dtu.dk/pubdb/views/edoc_download.php/3157/pdf/imm3157.pdf. Accessed January 5, 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.

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.

Freitag, Dayne and Kushmerick, Nicholas (2000). "Boosted Wrapper Induction." *Proceedings of the Seventeenth National Conference on Artificial Intelligence and Twelfth Conference on Innovative Applications of Artificial Intelligence.* Online: http://citeseer.ist.psu.edu/cache/papers/cs/23958/http:zSzzSzwww.cs.ucd.iezSzstaffzSznickzSzhomezSzresearchzSzdownloadzSzfreitag-aaai2000.pdf/freitag00boosted.pdf. Accessed January 19, 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.

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.

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.

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.

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.

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.

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.

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.

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.

Ma, Liping and Shepherd, John (2004). "Information Extraction Using Two-Phase Pattern Discovery." *SIGIR'04 Sheffield, South Yorkshire, UK.* July 25–29, 2004. Online. http://delivery.acm.org/10.1145/1010000/1009107/p534-ma.pdf?key1=1009107&key2=5107656311&coll=GUIDE&dl=GUIDE&CFID=61576327&CFTOKEN=9622293. Accessed January 10, 2006.

Mayfield, James; McNamee, Paul and Piatko, Christine (2003), "Named Entity Recognition using Hundreds of Thousands of Features." *Proceedings of CoNLL-2003, Edmonton, Canada*, pp. 184-187. Online. http://acl.ldc.upenn.edu/W/W03/W03-0429.pdf. Accessed January 9, 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.

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.

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.

MUC7 (2005). *Message Understanding Conference Proceedings: MUC-7 Table of Contents.* March 8, 2005. Online. http://www.itl.nist.gov/iaui/894.02/related_projects/muc/proceedings/muc_7_toc.html. Accessed January 25, 2006.

Muñoz, R. and Palomar, M. (1998). "Sentence Boundary and Named Entity Recognition in EXIT system: Information Extraction System of Notarial Texts." *Proceedings of IV Int. Conference on Artificial Intelligence and Emerging Technologies in Accounting*.

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

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

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.

Platt, John C. (1999)." Probabilistic Outputs for Support Vector Machines and Comparisons to Regularized Likelihood Methods." *Advances in Large Margin Classifiers* Scholkopf, B.; Smola, A.; Bartlett, P. and Schuurmans, D., eds., pp. 61-74. MIT Press. Online. http://research.microsoft.com/~jplatt/SVMprob.ps.gz. Accessed January 26, 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). *Link Analysis Survey Status Update.* January, 2006. NIJ Proposal Number 2005-93046-PA-IJ.

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.

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.

Sang, Erik F. Tjong Kim and De Meulder, Fien (2003). "Introduction to the CoNLL-2003 Shared Task: Language Independent Named Entity Recognition." *Proceedings of CoNLL-2003*. *Edmonton, Canada*. Online. http://acl.ldc.upenn.edu/W/W03/W03-0419.pdf. Accessed January 26, 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.

SearchTools (2005). *Search Tools Product Listings in Alphabetical Order.* December 14, 2005. Online. http://www.searchtools.com/tools/tools.html. Accessed January 25, 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.

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.

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.

Toral, Antonio (2005). "DRAMNERI: A Free Knowledge Based Tool to Named Entity Recognition." *Proceedings of the 1st Free Software Technologies Conference. A Coruña, Spain.* pp. 27-32. July, 2005. Online: http://www.dlsi.ua.es/~atoral/publications/2005_fstc_dramneri_paper.pdf. Accessed January 19, 2006.

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.

Vapnik, Vladimir N. (1995). *The Nature of Statistical Learning Theory*. Springer-Verlag.

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.

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, Tianhao and Pottenger, William M. (2003). "A Semi-supervised Algorithm for Pattern Discovery in Information Extraction from Textual Data." *Proceedings of the Seventh Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD-03). Seoul, Korea, April/May, 2003*. Online. http://www.cse.lehigh.edu/~billp/pubs/PAKDD03.pdf. Accessed January 11, 2006.

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.

Zhou, GuoDong and Su, Jian (2002). "Named Entity Recognition using an HMM-based Chunk Tagger." *Proceedings of the 40th Annual Meeting of the ACL, Philadelphia, Pennsylvania*. pp. 473-480. July 2002. Online. http://acl.ldc.upenn.edu/P/P02/P02-1060.pdf. Accessed January 27, 2006.

Zhu, Jianhan; Uren, Victoria and Motta, Enrico (2005). "ESpotter: Adaptive Named Entity Recognition for Web Browsing." Proceedings of *Workshop on IT Tools for Knowledge Management Systems at WM2005 Conference, Kaiserslautern, Germany,* pp. 505-510. April 11-13, 2005. Online. http://kmi.open.ac.uk/people/jianhan/zhuetal_KMTools.pdf. Accessed January 10, 2006.