Uncovering and leveraging the insights hidden within the overwhelming amount of text-based information available to a company is a critical success factor in today’s information economy. Yet, until now there existed no way, short of manually reading (or said more appropriately, skimming) the documents to determine whether they contain the knowledge gems you seek. The challenge with this lies in the fact that it is an arduous and time consuming task and as a result, only a small fraction of the available information is uncovered as we have been forced to settle for “good enough” research rather than a more exhaustive review of the sources.

How many market opportunities are missed? How much time is wasted on duplicate efforts? How many customer issues go unnoticed? All this because people read too slowly? Yes.

The purpose of LexiQuest Mine is to automate the process of reading documents to uncover their content. After all, in a thirty page publication, there may be only two or three sentences that are relevant to your present effort, yet those two or three facts may be the difference between success and failure. However, until now, the only means of discovering these salient points was to read the entire 30 pages. No more. Linguistics - powered LexiQuest Mine will do the reading for you and within seconds present you with a graphical display of the main concepts found in the document. From here, you can quickly find the elements of interest and navigate to that segment of the document.

The true value of Mine, however, does not lie in scanning one document, but in scanning thousands. Imagine being able to, in a matter of minutes, review and cull out the key concepts found in a decade’s worth of corporate research? Or a year’s worth of call logs from your customer service center? Or a week’s worth of newsfeeds about your markets, your products, and your competitors? Keep ahead of the information flood, learn what you’ve been missing, and apply the knowledge faster.

Based on 24 years of research into computational linguistics, LexiQuest Mine can understand language the same way we do. Capable of handling over 250,000 pages of text per hour, Mine can make quick work of development and patent research, competitive intelligence, and customer feedback. Mine can work on almost any text format including HTML, XML, MS Office, PDF, and e-mail and is available for use in English, French and German.

While there are a few competing products in the market for text or unstructured data mining, none can match the performance of LexiQuest Mine because none employ the depth of capability that can only come from natural language and computational linguistics. If the system cannot understand the words it is scanning, then you will still be missing key concepts, you’ll just be missing them in more documents.

The growth and importance of unstructured (text-based) information

The age-old knowledge management problem is getting the right information to the right people at the right time. This information is essential to make informed business decisions. This type of information (unstructured data) was once of a less time sensitive nature as people stored most business critical and transactional information in databases. However in the last few years an increasing amount of essential business information is being held in unstructured and semi-structured formats. Letters, emails, documents, news feeds, and presentations etc. are forming the backbone of most company’s information systems. Accurate retrieval and organization of this information is becoming an enormous challenge for all large companies.
In the last six years companies have tried to simplify their IT structures by making browser based portals a core part of their user’s desktop, giving them access to the Internet and to Intranets containing large centrally held repositories of information. The typical intranet contains hundreds of Web servers, file servers, specialized repositories and hundreds of thousands of documents. Most analysts agree that the information accessible via these networks is experiencing phenomenal growth and is expected to triple by the end of 2003.

While corporate content is physically more accessible, it is not necessarily more organized or easier to locate. Basically there are too much data in too many formats - both internally and externally. Without organization, intranet users only search a limited set of sources and lack easier, more intuitive ways to locate the information they want. Corporations want to prevent their users from being deluged with useless information vis-à-vis their own professional environment, manage user authorizations and limit what is available to an end-user to what he really needs to carry out his work. Content quality management and terminology management are emerging needs today.

The major problem with the management of unstructured data is that there are no rules for writing text so that a computer will understand it. The result is that for every document or piece of text the language, and therefore the meaning, varies greatly. The only way to accurately retrieve and organize unstructured text is to be able to analyze the language to understand it’s meaning.

Understanding the language of the content and efficiently organizing it and accurately retrieving it is the job of the LexiQuest range of products, the LexiQuest Knowledge Suite.

**LexiQuest's core technology - Natural Language Processing**

The capability to understand human language is provided to computers through the power of linguistics, commonly referred to as Natural Language Processing (NLP). All the traditional methods: (key word searching, inverted indexes, boolean searches, statistical, probability algorithms, concept agents, neural networks and pattern recognition) do not provide any level of “understanding” of the text or of the concepts represented by the queries.

These systems are only based on the comparison of the character strings in both queries and text. This produces poor results, with a lot of noise (irrelevant results) and silence (accurate results are not found). For example in a query like “reproduction of documents”, the word “reproduction” has to be expanded to a synonym like “copy” or “duplication”. If this is not done, silence will be generated and relevant information will be overlooked by the system. But if a non-linguistic based system tries to perform this kind of synonymy, there is a good chance that it will also expand to “birth”, generating noise, which provides irrelevant information to the requestor. On the other hand, with a question such as “reproduction of cats”, the system should expand on “birth”, but not on “copy”.

In non-linguistic systems such as neural networks, the inability to manage the tradeoffs between noise and silence has pushed developers to try another approach. To compensate for their lack of understanding they statistically observe co-occurrences. The basic idea is; “if a text is about reproduction of documents, then it will likely use the word “copy”. This approach is far from satisfactory. To be effective, the text has to be quite long to provide a reasonable sample, and even then there is no guarantee that all the possible ways to phrase the concepts in the query will occur in the text.
Statistical and probability techniques are by essence limited. They are not context sensitive. For example there is a good chance that a two-word query will just give back all the texts where one of these two words occurs frequently. While a statistical system is parameterized, it has a maximum level of quality beyond which it cannot evolve. On the contrary, linguistic systems are knowledge sensitive: the more information there is in their dictionaries, the better the quality. As a result, LexiQuest Technology has no limits to the quality levels it can provide.

In addition, they are some key tasks that statistical methods cannot address, such as: multilingual access, summarization, intelligent dialog management and knowledge extraction, for example performing a “relevance feedback” function. These functionalities will in the near future make the differences between “good” and “poor” information retrieval systems. A key advantage for LexiQuest is that only linguistics can bring a credible solution to these issues.

Text mining: finding more than just a needle in the information haystack

Companies are faced with ever increasing, unmanageable volumes of unstructured data. As a result, most large organizations are not aware of all the content contained in their intranets, corporate libraries or knowledge management repositories. In addition to the daily growth of these internal information warehouses, manual attempts to monitoring external information sources such as websites, news feeds, journals, analyst reports and often miss critical information. The issue to date has been that traditional search engines take a “top down” approach to information location, that is, the user must know precisely what they are looking for and then structure their query in such a way as to be brought directly to that element. With an efficient search system, you can indeed find a needle in a haystack. But if all you are looking for are needles, you will miss the pile of gold coins and the horde of diamonds that are also hidden in the same haystack.

In today’s information economy, what is required is a system that “inventories” all of the content within large volumes of text and provides you with an easy to navigate representation of the content so that you may determine what is important. LexiQuest Mine does just this. By taking a bottom-up, mining approach, key concepts and the relationships between them are quickly identified and presented to the user in a graphical map. The map enables the user to quickly see links between concepts, uncover the nature of a term (product, person’s name, organization etc), and then structure a search query to be brought immediately to the source document(s) where more information is available.

How does it work?

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 linguistic 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 (see Figure 1) 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 existing corpus.

There are very few, if any, products in the market that have the identical function of LexiQuest Mine and those that claim to perform the same level of analysis typically lack one or more of the following attributes:

- **Linguistics** Using NLP technology LexiQuest Mine is able to recognize and extract compound words, phrases and idioms that would typically be treated as individual words by other products. This has a dramatic effect on the overall accuracy of the systems and is the core distinguishing difference with other mining products.

- **Term Typing** LexiQuest Mine is built using LexiQuest’s proprietary lexicon dictionaries and as a result, concepts can be identified as being a unique term, a product, a person, an organization or a location enabling the user to quickly determine a concept’s relevance to their particular information need.

- **Volume/Speed** LexiQuest Mine is suited to analyzing extremely large volumes of data very quickly (250,000 pages of text per hour) ensuring you never have to miss a key insight simply because you didn’t have time to look through all the available sources.
Sample applications

<table>
<thead>
<tr>
<th>Department</th>
<th>Application</th>
</tr>
</thead>
<tbody>
<tr>
<td>Investment research</td>
<td>Ability to quickly review thousands of daily analyst reports and company press releases to identify key strategy points or market shifts. Trend analysis enables the user to track emerging issues or opportunities for a firm or industry over a period of time.</td>
</tr>
<tr>
<td>General scientific and medical research</td>
<td>Used for reviewing secondary research materials such as patent reports, journal articles, and protocol publications. Identifies associations that were previously unknown (such as a doctor associated with a particular product) offering avenues for further exploration. Saves valuable time in the drug discovery process.</td>
</tr>
<tr>
<td>Competitive and market intelligence</td>
<td>Quickly scan new information (PDFs, websites, analyst reports, market research reports etc.) published on product announcements, merger discussions, price promotions, customer preference etc. for your own firm and competitors. Mine call logs from customer service centers or customer e-mails to identify product preferences or potential quality control issues.</td>
</tr>
<tr>
<td>General</td>
<td>Mine is for anyone who routinely needs to review large volumes of documents to identify key elements for further exploration.</td>
</tr>
</tbody>
</table>

Customer experiences

LexiQuest Mine has been deployed at a variety of customers to handle different needs. At Astra Zeneca, for example, it is helping research scientists in their drug discovery labs to quickly read through and absorb the content in thousands of current and archived patent documents and journal reports in order to better leverage past efforts and keep abreast of the latest findings in their area. This ensures that valuable secondary research is not missed while dramatically reducing the time spent reviewing these materials.

At Peugeot, their competitive intelligence team is employing LexiQuest Mine to quickly sort through information published on hundreds of websites and periodicals in order to monitor the activities of their competition while identifying and tracking emerging market opportunities. By broadening the number of sources that are regularly reviewed, Peugeot is able to track developments across a wider spectrum of their market, identify competitive advantages and then exploit them to drive revenue.
Technical environment

LexiQuest Mine is composed of the following:

- **Database Manager**  Available from any Web browser, the database manager provides the functions for managing LexiQuest Mine’s databases. It also allows remote administration of the LexiQuest Mine environment
- **LexiQuest Mine**  Available from any Java-compliant Web browser, the application is accessible from the user’s desktop and provides the concept extraction, graphical interface, trending analysis and access to customizable administration features
- **Database Server**  Server application that performs statistic processing for dynamic clustering. Essentially serving as the engine which drives the identification of relationships between the concepts.
- **LexiQuest Base for Text Mining**  Utilizing the core LexiQuest Language Recognition technology, Base supplies the lexical information for the linguistic analysis of documents in French, English, German, or Spanish.
- **Search Engine**  Allows documents to be retrieved from the LexiQuest Mine maps and attached terms. These documents may be in a variety of formats (XML, PDF, MS office, txt, html etc.). Can be used to retrieve documents from the existing knowledge base or can be used to develop queries submitted to public search engines such as Google.

About SPSS Inc.

SPSS Inc. (Nasdaq: SPSS) headquartered in Chicago, IL, USA, is a multinational computer software company providing technology that transforms data into insight through the use of predictive analytics and other data mining techniques. The company's solutions and products enable organizations to manage the future by learning from the past, understanding the present, as well as predicting potential problems and opportunities. For more information, visit [www.spss.com](http://www.spss.com).