LexiQuest

Categorize

SPSS®

Managing and leveraging the overwhelming amount of text-based information available to a company is a critical success factor in today's information economy. Yet, the task of organizing and categorizing this information into a structure that makes sense to the average knowledge worker has largely been a time-consuming and tedious manual effort. The resulting bottlenecks and general frustration with the process has led many companies to abandon the challenge, leaving cluttered and confusing websites in their wake.

The purpose of LexiQuest Categorize is to automate the process of organizing documents into logical taxonomies reducing the need for human intervention, eliminating bottlenecks and ensuring prompt, efficient delivery of information to its intended audience.

Whether the consumer is navigating your intranet, surfing your website, or receiving "alerts" informing them of newly published content, the speed and ease with which they can find the required information can often mean the difference between growth and decline for your business. Based on 24 years of research into computational linguistics, LexiQuest Categorize can understand language the same way we do. Using a prescribed set of documents, Categorize "reads" the information and intelligently "learns" the types of information you expect to see in each category. You can then customize, refine and confirm its results. Once you are satisfied, simply add your new documents to its repository and let it do the rest.

Capable of handling over 250,000 pages of text per hour, Categorize can make quick work your existing documents and then can be run each day to handle the increasing volume of new information. Categorize can work on almost any text format including HTML, XML, MS Office, PDF, and e-mail and can be used to catalog websites, research documents, analyst reports, legal briefs, etc. ensuring you access to the broadest possible sources of information while maintaining an easy to navigate sense of order. While there are a few competing products in the market for automatic categorization, none can match the performance of LexiQuest Categorize 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 the accuracy of the placement suffers and without accuracy, you are simply automating your existing chaos. It may be faster, but it's not better.

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: multi-lingual 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 theses issues.

Categorization: a cure worse than the ailment? Not anymore…

Companies are faced with ever increasing, unmanageable volumes of unstructured data. Maintaining order in this environment is a never-ending struggle. Users demand immediate access and an easy to navigate means to access the information. These goals are often in conflict as the length of time necessary to manually effect an accurate categorization of the incoming documents is too long and therefore unacceptable. The result is that this step is often bypassed in favor of speed. The result is a cluttered and confusing user environment full of loosely correlated documents. In such an instance, the user becomes a microcosm of the overall organization as they personally do not have the time to review all the available material they are forced to accept only the limited pieces they can quickly find. The outcome is that crucial information is often overlooked, and opportunities are lost.

The desired state is one in which information was made available in a logical structure (either through a navigable taxonomy or improved search functionality) without the burden of delaying its release. Until now, the barrier to achieving this state lay in the fact that human intervention was required in order to ensure accuracy of the catalogued documents. This requirement was borne of the fact that systems could not understand the meaning behind the documents they were attempting to organize resulting in a high degree of errors that was often no better than the chaos that had previously existed. Now, with the advent of true linguistics analysis, LexiQuest Categorize has broken through this barrier by enabling systems to understand language the same way people do enabling a high degree of accuracy without the need for time consuming human review (see Figure 1).

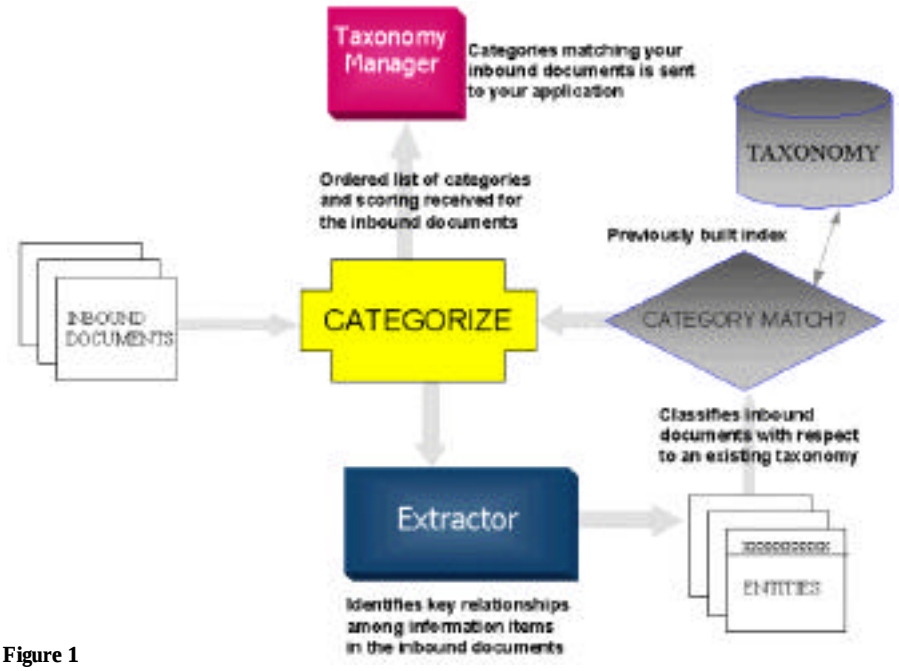

Figure 1

How does it work?

Typically, documents are submitted by end users via an independent application developed by one of LexiQuest's strategic partner or systems integrator. Categorize extracts textual information from the submitted documents and uses this information to categorize them. To do this, Categorize matches the entities extracted from the submitted documents against descriptions of categories organized according to a taxonomy. A taxonomy is a predefined list or tree of categories that Categorize can recognize. When matching categories are found for the submitted documents the application allows the company to either apply meta-tags or simply forward the documents to the customer-facing application (intranet/website/portal taxonomy navigator etc).

The initial category descriptions are obtained using sets of learning documents which serve to "teach" Categorize what you would like to see in each category. From these, a series of terms are found and weights assigned based on the uniqueness of the term. For example, a common word like "view" would receive a low weight as it will be found in documents assigned to many categories whereas a phrase like "fuselage" would receive higher weighting due to its more exclusive use in the aerospace category (see Figure 2).

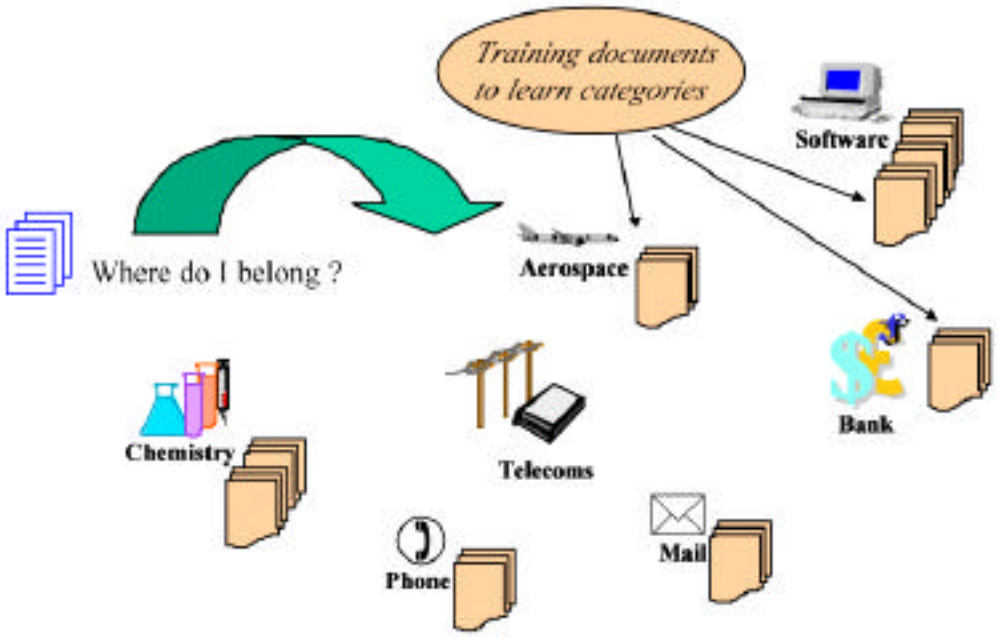

Figure 2

It is this combination of linguistics-powered term extraction and comparative weighting that places LexiQuest Categorize at the forefront of intelligent categorization. There are very few, if any, products in the market that have the identical function of LexiQuest Categorize. Most categorization or taxonomy creation products lack one of the following attributes:

- **Linguistics** Using NLP technology LexiQuest Categorize 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 categorization products.
- **Term Extractors** LexiQuest Categorize employs the same technology as our text mining tool, LexiQuest Mine and as such has the ability to extract specific types or categories of information from text. This enables the categorization to be independent of the domain being processed and accurate regardless of the industry.
- **Volume/Speed** LexiQuest Categorize is suited to cataloguing extremely large volumes of data very quickly (250,000 pages of text per hour).

Sample applications

Department	Application
Investment research	Ability to organize huge volumes of reports and industry publications which reside in the bank whether in a library or in each department. Then track news and information regarding a particular company or product and identify interesting or unusual relationships between the data.
General scientific and medical research	Easily catalogue and navigate thousands of documents on patents, web, journals, intranets, lab books
Competitive and market intelligence	Receive alerts when new information is published on product announcements, merger discussions, price promotions, customer preference etc. for your own firm and competitors.
General	Keeping taxonomies and search systems accurate and current to satisfy needs of customers and employees improving satisfaction with site and ensuring all relevant information is available to the right audience.

Customer experience

A leading business and competitive intelligence association in France was faced with an enormous challenge. Tasked with the sole mission of collecting and sharing information in the technical industry across France, they were struggling with the volume of new documents they were receiving. Each month, the organization was having to review and coordinate distribution of 10,000 to 100,000 new publications and having to manually review each new piece to determine its proper location/audience within their 1200 node taxonomy. Limited resources and the unpredictable volume of work was creating an unacceptable backlog in the process, frustrating their members and management. By applying LexiQuest Categorize, they are able to automate this process and dramatically increase the speed of their delivery to their members while improving the degree of order and accuracy across their catalogue.

Technical environment

LexiQuest Categorize consists of the following internal components (see Figure 3):

- **Categorizer Engine** The categorizer engine is the main component of the categorization system. The categorizer provides the input-output interface and handles all dialog with the internal components.
- **Extractor Module** Extraction is the operation of examining the morphology of words in a document and extracting words from the document so that these words may be subsequently transformed and used to match suitable categories for the document. The extraction process for documents is mostly a process of identifying the right subset of words relevant to the chosen subject. This component detects and selects invaluable information, and in certain situations, associates added value to the text initially submitted.
- **Taxonomy Manager** The Taxonomy Manager is a powerful graphical user interface that will permit you to create and maintain your category sets, complete "training" with sample documents, verify and measure the quality of your categorization application, and customize the linguistic resources to precisely adapt the terminology extractor to your unique needs.

Figure 3

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.**