Do You Know What You Know?

Introduction

Cataloguing unstructured information is a chronic problem, that if not adequately addressed can be terminal for your organization. Today we have many easy-to-use and accessible tools to create and publish information electronically. Examples are: the ubiquitous suite of Microsoft office products like Word and PowerPoint, Adobe Portable Document Files (PDF), Web Pages (HTML files), e-mail, news feeds, and the like.

Lack of information is no longer the problem—but lack of time to correlate, categorize, analyze and act on the information is a crucial problem. The information is there, hidden in reports and e-mails and published on the corporate Web site. We are placed in the position of being unable to find applicable and pertinent information to make timely business decisions. This comes at a time when the agility to quickly make fast, informed decisions is increasingly critical to survival and prosperity. As the volume of opportunities increases, the duration of the time to act on each opportunity decreases. The information-based economy is in danger of drowning in a sea of irrelevant, unstructured data.

A new segment of software has emerged to help with the task of combating “infoglut.” For example, there is software that enhances the performance of search engines, text mining, natural language search
applications, ontology, summarization and taxonomy. Taxonomy software correlates and groups unstructured information from a myriad of sources. Taxonomy software also helps us in the automation of this process. The software’s fundamental challenge is to understand the concepts and ideas that group like documents together and separate unlike documents. Think of taxonomies as computer-generated card catalogs that allow us to locate, retrieve and cross-reference information in our digital libraries.

Taxonomy software can reduce our reaction time to make informed and timely business decisions based on knowledge and information contained within the unstructured data of an organization’s digital documents. This software helps us form ideas from information we didn’t know we had while revealing relationships and correlations that were submerged or lost in the depths of the ocean of information overload. This happens on an individual basis and a community basis. We work in groups and we must be able to communicate with the constituents of the group so we can have actionable knowledge. Productivity comes from seeing connections, evaluating importance, recognizing context, understanding the implications and understanding the correlations of data and information.

Although taxonomy software cannot stem the tide of “infoglut,” it can help us find the information we need to survive and prosper in the new knowledge-based economy – to truly know what we know.

Irrelevant Information and Infoglut

An ancient Greek philosopher and ascetic, Diogenes searched for an honest man and never fulfilled his quest. You may feel like Diogenes when trying to search for and retrieve information from your enterprise portal or the Web to make a business decision.

Let’s use the example of searching for information about chips. A search on the Web for “chips” (using the Google search engine) returned 2,430,000 references. Even if only 1% of these documents were relevant, that is over 24,000 documents, which is way beyond the capability of most of us to wade through. Some of the documents contained information about – chocolate chips, potato chips, buffalo chips, wood chips, poker chips, an old TV series (ChiPS), or integrated circuits. “Chips” is one of those words that have multiple ambiguous meanings like “java,” “can,” “branches” and “boot” (noun or a verb).

People can distinguish concepts from each other based on context and the specific meaning of a word as it should be applied to the situation. Computers cannot. Other methods must be used by computers to give us the results we want. But before we explore these methods, let’s look at some examples of how people think versus how computers “think.”

Context

The definition of context is that which surrounds, and gives meaning to, something else. People explore concepts; computers primarily search for key words. Relevancy is entirely subjective to the individual who is performing the search. Only each individual can judge how relevant a particular bit of information is to what they are attempting to discover. The document may be too technical or out of date or too general for your needs. Context is the determining factor. Machines can’t distinguish between “John Smith to marry Mary Jones” vs. “Reverend Billy Graham to marry Bruce Springsteen.” Only people with the proper context can know that Reverend Billy Graham will perform the ceremony and not be the recipient of Bruce’s ardor.

“Customers” may be a descriptor for an employee in an HR document about medical benefits, but “customers” may mean something different in the context of the sales department. If a given document mentions “customers” where should it be categorized?

In the above example about “chips,” you might have been looking for integrated circuit “chips” and not a recipe to make chocolate “chip” cookies. If you had only searched in a category such as computers or electronics, you would have found fewer documents, but more precise and relevant information. If you have categories and hierarchical structures of information you will be able to narrow the search field and find relevant information faster.
**Ambiguity**

The beauty of the English language is that it has many words to describe the same thing. The corollary is that the same word may have different meanings. “Chips” is one example of an ambiguous word. “Java” is another. Java could be an island, a cup of coffee or a computer programming language. “Kick the bucket” is another phrase that could have multiple meanings depending on the context. Consider the sentence, “Joe kicked the bucket and the water spilled out.” The question is did Joe “buy the farm” (drop dead) or did Joe violently place his foot in contact with a container of water. The context of the surrounding phrases in the rest of the document will clarify this. So the question becomes, how do you categorize the documents that has words like “chips” or “java” or phrases like “kick the bucket.”

**Browsing vs. Searching**

To search effectively you must know the terms you want to use before you see what is in the collection of documents. Key word search however assumes you know what you are looking for and that is often an erroneous assumption. Knowledge workers are not always exactly sure what they are looking for but, “when they see it, they know what it is.” More than 25% of the day is involved in searching for information on the knowledge workers computer system. About 70% of that time is spent browsing for information. 75% of the people surveyed during a Yahoo market research project preferred browsing to searching.

From our previous example about “chips,” you may not have known there were many types of computer “chips,” such as processor “chips,” application specific (ASIC) “chips” or memory “chips.” If there were categories of computer “chips” such as: processor, ASIC and memory “chips”; you may find all the information you need is about flash memory “chips”. It is much easier to discover information about a particular subject if you see it in the context of related information. Browsing encourages associative thought. Browsing in categories can guide you through the information discovery process.

**Report Scope**

The objectives of this report are to:

- Inform about the current state of the market for taxonomy software
- Explain with examples how these technologies are applied to real life business issues
- Report the findings of a market survey of over 450 organizations about taxonomy
- Understand how those organizations are planning to implement taxonomy technology
- Analyze trends of taxonomy software technology, development and implementation
- Describe the technological landscape of the taxonomy software market
- Compare the technologies’ similarities and differences

**Taxonomies**

**Why Now?**

The first question we have to ask is, “Why the current high interest about taxonomy?”

Delphi Group’s research on user experiences with corporate Webs reveals that lack of organization of information is in fact the number one problem in the opinion of business professionals. These professionals may include the customer service representatives, the sales team, the financial

---

2 Delphi Group Research
services professionals, the R&D engineers and the health care professionals that work in our organizations. If these professionals are spending 25% of their time or more looking for information, then this results in an opportunity cost and represents a runaway expense item in many organizations. Our ability to create information has substantially outpaced our ability to retrieve relevant information. Delphi calls this information explosion “digital sprawl.” The impact of this situation has dramatically affected the way we work.

We now have many tools to aid in the creation of electronic information. It is easy to create an electronic report or presentation and convert those to PDF files for easy viewing and sharing. E-mails are the ubiquitous mechanism for communication within business. An estimated eight billion e-mail messages will flood U.S. corporations each day in 2002 say leading industry sources. Proposals, resumes and contracts represent other sources of unstructured documents. Portals, Web sites, visible and invisible, and intranets have also made it easy to place and share information. Web pages are increasing at the rate of 7 million pages per day.

Since it is easy to create, share and store information, the rate of unstructured information is now growing exponentially and turning into what one taxonomy vendor calls knowledge asset “landfills.” There are 250 megabytes of information for every man, woman, and child on earth. The percentage of unstructured data to the total amount of data is estimated at 85% and is growing. One of the defining challenges of this era of enterprise computing is just this: How do we find the relevant and pertinent information to do our jobs and make informed business decisions? The answer is at once obvious and elusive. We must harness the computer to help fix the problem it has helped us create.

---

**Evolution of Taxonomy Technologies**

People are natural categorizers. We tend to group similar documents into categories by conceptual subject matter. We have developed filing systems for our paper documents. Examples of these are found in every office. We have filing cabinet filled with paper documents that are grouped together by some system:

- Categorized by type e.g. contracts, marketing collateral, or invoices
- Alphabetized by company or individual’s last name
- Arranged by date
- Grouped by department e.g. legal, HR, sales, accounting

We do this so that when we can find the relevant information to complete the actions or make the decisions – in short to run our business. We call this “actionable information.”

We have structured databases to keep track of our transactions and customer lists. This data is in rows and columns that tell us who to contact, how much to pay, how many we sold, etc. But this structured information doesn’t tell us what will likely happen or why. The “why” is usually buried in our “corporate landfill” of associated e-mails, reports, or presentations.

The first step was search and retrieval software. Unstructured documents were “scanned” and indexed and key words along with their frequency of occurrence were placed into databases. Search engines were developed so that we could scan the index and find the documents and create pointers or hyperlinks to the documents that contained the key words that we thought were relevant. The simple search engines and indexers soon developed into complex ones that could support Boolean logic such as “and,”

---

2 Invisible refers to dynamically generated Web like Amazon.com catalog pages, Wall Street Journal Archives. – Invisible Web, Chris Sherman

3 UC Berkeley

4 UC Berkeley SIMS How much information.
“or,” “not” type of qualifiers. These functions were expanded to include “wild card” searches and “fuzzy” searches like “near.” When sources were relatively limited these tools helped you find the information you wanted.

Now, however, instead of returning too small a list of relevant documents, the search engines return too many choices, as in our previous example of searching for “chips” with 2,400,000 possible documents. These are brain-dead lists that insult our intelligence. When Boolean logic searches turned out to be too complicated for most people to use, natural language query software helped develop key word searches that reflected our vernacular and intentions when searching for information. The same problem soon developed – too much recall, and too many references, even when ranked by relevancy percentages.

Another mechanism soon evolved, that of metadata, or “data about data.” With this approach, when a document is created the author is prompted to list the creation date, title, subject matter, synopsis and a few keywords. Since authors are very close to the document as the creator, the relevancy and the consistency of this information were questionable.

The next step in retrieval technology was a methodology called link ranking. The more the Web page in question is linked to other systems and searched by other systems, the more probable it is important to more people. Each time a link is clicked, the click is recorded in the database and the client is redirected to the proper site, creating a simple click thru monitor that records the popularity of links. The flaw here is that this becomes a popularity poll or self-fulfilling mechanism; the more your site is visited the more likely it is to be visited.

The limitations of these various approaches were recognized by a number of “information and language” experts and a new generation of technology has emerged. Taxonomy software was originally developed to help speed the process of creating hierarchical structures.

**Is “Taxonomy” a Misnomer?**

The word taxonomy evolved from the life sciences where a plant or animal is placed in a single spot describing its hierarchical relationship to other plants and animals. One of the interesting ironies that developed during the research phase of the project was the realization that the description of this category of technology was somewhat misleading. Since we are talking about semantics and language here, there is an inherent problem with using the word taxonomy to describe this type of technology. When applied to digital information, taxonomy is a systematic classification of a conceptual space. When categorizing a document, it could and might be placed in multiple categories depending on the context.

For example at a pharmaceutical company the research scientists may place a particular document about a new drug in a category based on its chemical composition. The marketing and sales people may want to place that same document in a category called “competition.” The same document can be classified into two or more categories by different groups or under different contexts or conditions. (You can see where the analogy breaks down: scientific taxonomies are fundamentally rigid, while categorization is personal and subjective and sometimes even arbitrary).

**Using Categories in the Search Process**

People search for information in two different ways. The first is the process you use when you know what you are looking for. You know the answer, now you need to find more information about the subject. Keyword search with Boolean logic and traditional search engines are good for this type of approach.

Many times people don’t know the answer they are looking for when they begin the search for information. In fact, as discussed above, they often don’t know what they don’t know yet. Let’s use the example of doing research on categorization technologies. Here is an example of a possible hierarchy of categorization information. Each of these categories can also be called a node.
As you start the search process for categorization software you soon realize there is one alternative, the purely manual approach to categorization. There are several reasons, however, why an exclusively manual system may not be as effective: the chief drawbacks are that human classifiers are expensive, inconsistent, and obviously not scalable to the same degree as automated approaches. Although these considerations were probably not foremost in your mind when you started your search, they certainly can support your argument in favor of the purchase of a software system.

The methodologies used in the automatic process of categorization are likewise not something you would know about at the beginning of your search. By learning more about the various methods of automatic categorization (i.e., rules, example learning, semantic analysis and clustering analysis) you can better understand how each of the methodologies may be applied to your particular situation.

The point we are making here is that when you started your quest, you were not thinking about cost-justifying a software solution, nor were you aware of the various methodological approaches to the problem. This information was discovered either during or after your search—maybe. Wouldn’t it be better to have such information at your disposal at the beginning of your project?

Visual, hierarchical arrangements of subject categorization trigger associations and relationships that are not obvious when searching for keywords. This distinction is important and implies yet another reason why categorization can be critically important to the productivity of knowledge workers.

**Browsing Process**

As you browse for relevant information about your particular subject, you will find three more characteristics that describe the process: browsing is dynamic, interactive and iterative.

Browsing is dynamic. Information changes all the time. In today’s world, virtually any search on a complex topic becomes a hunt for a moving target. For example, my search for “chips” today now yields 2,490,000 hits—60,000 more instances of information than just a few days previously. Versions will change, articles will be removed, information added, etc. Delphi’s research shows that at least 10% of enterprise information changes on a monthly basis. This is a conservative finding—Delphi Group believes that the dynamic, volatile nature of information sources is the number one reason that knowledge workers have difficulty finding the information they are looking for.
Browsing is an interactive process. As you navigate a well-designed interface to information, you will automatically be directed to other relevant topics. If you search and browse through information about categorization software, for example, you will find reviews, analysis, white papers and commentaries with information about other technologies, companies or related topics of information that may be worth investigating.

Browsing is an iterative process. Repeating the process refines your focus while broadening your knowledge. Accessing relevant information and interrelated ideas and concepts supports a fundamental change in your activity—from simply searching, to finding and discovering.

**Taxonomy and Search**

One purpose of taxonomy is to aid in the retrieval of relevant information. An intrinsic benefit of the hierarchical structure of categorization is that links and summaries of information are rendered in the context of their unique “parent-child” relationships. Relevant information is more likely to be found when specific content filters are employed. For example, if we had had a general category like “computers” in our search for “chips,” we would not have wasted any time with false returns from the “recipes” category.

**Benefits of Taxonomy**

Finding relevant information quicker is the key benefit, especially when it provides immediate access to the right information that allows the user to take effective actions. Equipping enterprise knowledge workers with the tools to make faster and better-informed decisions is a strategic imperative in today’s economy. Jakob Nielsen, the guru of usability, estimates that poor classification costs a 10,000 user organization $10M annually.

To paraphrase a quote, “to search via a computer without a taxonomy system is like trying to find your way around an unfamiliar country without a map.” Taxonomy helps delineate the conceptual relationships that exist within and between various topics contained in the multitude of unstructured data within various enterprise documents.

The benefits are:

- Discovering information you didn’t know you had
- Avoiding duplicate efforts within large organizations where independent groups “reinvent the wheel” over and over again
- Not repeating the same mistake
- Reports are better prepared if the author really expects to be read.\(^5\)
- Provide overview as well as details about a subject
- Demonstrate relationships
- Reduce complexity

**Taxonomy Software Integration With Other Applications**

Taxonomy can impact many aspects of your organization. As organizations implement various software solutions to manage their knowledge assets, taxonomy can dramatically increase the effectiveness of such solutions. All the software features in the world won’t matter if they don’t facilitate “just-in-time” knowledge retrieval. Software applications such as portals, content management systems, knowledge management systems, search and retrieval software, personalization software, data extraction, and data mining can all benefit from taxonomy. Many taxonomy solutions are sold with Application Programming Interfaces to integrate into these existing applications.

---

\(^5\) There is a product marketing manager that places a peach cobbler recipe in his market research reports. He places the recipe in a relatively arcane section of the report. At the end of the recipe, he offers to whomever finds the recipe a bottle of Don Perignon champagne. In his career, producing many reports, he never has had to buy the champagne.
Taxonomy Is Not Just For Enhancing Search

Let's apply taxonomy to other applications. Taxonomy technology is currently making search engines more efficient by, as we've seen above, limiting a search to a subset of relevant topics. Browsing topics categorized by hierarchical relationships will allow new insights and correlations when searching for information. Imagine the impact of applying this technology to other applications that must access and manipulate masses of unstructured data.

One taxonomy vendor's customer is a professional services organization. This organization is constantly developing proposals for its customers. Since the organization is national in scope, at any given time there are many offices developing proposals for many customers—often in the same vertical industry. The professional services organization has an existing-proposal generating application. They are now using a taxonomy classification tool to discover:

• Proposals covering similar industries that have been successfully presented to similar customers
• Sections of proposals that can be reused
• Research data from previous proposals that are pertinent to the new customer
• The correlation between different proposal patterns or approaches and the success or failure of the proposal
• Up-to-date info from Internet news feeds of particular relevance to prospective customers

In another example, imagine categorizing unstructured data for the benefit of your CRM application. You might readily identify patterns and correlations that bridge seemingly disparate occurrences that are in fact tied together by some previously unknown common factor.

Finally, consider the case of a data mining application that discovers, for example, that sales of rain gear declines in the northeast while increasing in Arizona. Data mining is notoriously good at telling you the "what" but not the "why." By browsing a taxonomy-enhanced knowledge base, however, you may identify related e-mails or reports that point to the existence of an annual summer “monsoon” season in Arizona corresponding with an unusually dry summer in the Northeast. Now you have a “why” to substantiate a business decision to increase rain-related inventory in a desert region.

Determining Your Categorization Requirements

These questions will help you quantify how a taxonomy system can be applied to your organization. The primary strategic consideration for any firm is what kind of resources to commit to the process and in what proportions.

Suggested Questions

• Can employees, partners and customers find the information they want?
• How are you currently categorizing your information?
• What is the diversity of the documents?
• How many documents will you categorize?
• Are there isolated areas that have a minimal number of documents? (e.g. small intranets within the organization that don’t have enough documents to develop categories for them)
• Will this number grow over time or remain constant?
• What rate of growth do you expect—linear or exponential?
• What file types and formats are the information stored in? (e.g., text, PDF, PowerPoint, HTML, etc)
• How volatile and dynamic is the information? Does it change hourly, daily, monthly?
• When does it change?
• What is your official document publishing process and policy?
• What are the life cycle parameters?
• Who categorizes information now?
• Who sets up the categories?
• How available are they for this task?
• How do categories get updated and expanded?
Market Survey: Results and Analysis

During the first week of February 2002, Delphi conducted an extensive survey of approximately 450 end user organizations on the subject of categorization and taxonomy management. Several dozen questions were asked in regards to the business issues surrounding the evaluation, planning and implementation of taxonomy technology.

The objectives of the survey include:

• Validate the extent of the unstructured data problem faced by knowledge workers in today’s organization
• Determine the relative importance of the business issues surrounding the retrieval of information from unstructured data sources
• Understand the scope of the problem and the perceived impediments associated with job performance and unstructured data.
• Confirm the characteristics of unstructured information sources in terms of size, volatility, and language
• Verify if there are classification processes and policies in place today
• Ascertain if there are taxonomy software projects underway or pending, their relative importance, and the proposed budgets for implementation and maintenance of taxonomy
• Find out who will be responsible for defining and then maintaining the taxonomy software
• Clarify how the taxonomy software should be configured and deployed
• Discover to what extent the market recognizes the leading providers of taxonomy software

Survey Summary

Enterprises know they have a serious and growing problem with unstructured data, and that the problem is dramatically impacting their ability to make rapid and effective business decisions. Current systems are not adequate. Organizations planning on developing a taxonomy strategy remain unsure about how to do so. This presents a significant opportunity for the technology providers seeking to fill this need.

Profile of Respondents

The survey of 450-plus respondents represents a fair sampling of enterprise organizations, with over half the organizations having revenues of over $100 million. 73% are located in North America. The respondents were either executives, IT, or LOB or had project management responsibilities. Respondents’ role in determining taxonomy software was primarily as a sponsor or project lead, or involved in defining need or specifications.

Survey Methodology

Individuals identified by Delphi’s analyst team were contracted directly and asked to answer a series of structured survey questions. The survey format was primarily multiple-choice, with either single or multiple answers possible depending on the question. Respondents were also provided the opportunity to volunteer more detailed and otherwise restricted answers (i.e., “fill-in-the-blank”) which are quoted at various points throughout this report. The results from the survey are tabulated and graphed in the sections to follow. Listed below are the profiles of the respondents, the questions, the results and, finally, Delphi’s analysis of this market survey.

Survey Limitations and Risks

The survey resulted in over 450 respondents. While this number is sufficient to develop quantitative and qualitative trends, variations may be found within individual deployments or taxonomy initiatives. While respondents represents a valid cross-section of enterprise-class organizations, the population's preestablished interest in this or similar technology may distinguish this group as more knowledgeable and aware than a similar group of randomly selected enterprise respondents.
Survey Results
“Infoglut” and Knowledge Worker Efficiency

A number of questions were asked to validate the extent of infoglut faced by survey respondents and to verify the environmental causes.

“Infoglut,” the business problem of too much information, is real, growing and recognized as an important issue in today’s enterprise. The exponential rise of unstructured data is having a major detrimental impact on the efficiency of enterprise organizations. Most respondents to this survey (with job descriptions such as executives, IT management, Line-Of-Business managers and project managers) spend more than two hours a day (25% or more of an 8-hour day) searching for information to perform their jobs. More than 60% agree that finding information was a difficult process, and much of the time they cannot find the information they need to do their jobs.

Question #1 - “Finding the information I need to do my job is difficult: agree - disagree”

More than 60% of the respondents agreed or strongly agreed that finding information was a difficult process and over 50% of the respondents were spending 2 or more hours each day searching for information. This is consistent with Delphi direct client experience and various other surveys, which of shown knowledge workers typically spend 20% to 30% of their time searching for business-related information, the majority of which is stored electronically and should otherwise be easily identified.

The issue of “search time” is one of the fundamental symptoms of infoglut, and is at the heart of how most organizations measure the business impact of taxonomy management. While the positives impact information accessibly on productivity is at times overrated (i.e., simply because information is accessible is no guarantee that work is being done), there is no denying the negative impact resulting from the lack of information. Simply put, very few individuals are able to generate business value through the act of searching for information – this frequently required task is what economists refer to as a “transaction cost” or otherwise a drain on productivity. And it is what the rest of us would call “a waste of time.”
Question #3 - “The biggest impediment to finding the information I need to do my job is:”

As respondents looked and searched for information, 61% of the time they had a 75% chance or less of finding the information they needed. The two main impediments to finding the information they were seeking:

- bad tools 28% of the time
- volatility of the data 35% of the time.

The two most frequently cited reasons for this are “Bad Tools” and “Data Changes” or the concern that information is changing too fast. Even with the wide array of available search tools and content management solutions, knowledge workers either can’t find the information they need to do their jobs, or they are spending an inordinate amount of time looking for that information.

Document Turn-Over and Data Change

Another area of focus for the survey was an examination of information volatility, consistent with the finding that “data change” was the most frequently sited impediment to finding information. Over 95% of the respondents reported that more than 10% of their documents change on monthly basis. In a separate series of questions, more than half (59%) stated that they have over 50,000 documents in the corpus of information contained in their environment. Compounding the problem, the rate of information volatility is certain to accelerate. This study shows that respondents have access to at least 50,000 documents within their immediate organization— with often 10,000 or more of those documents changing on at least a monthly basis.

Question #4 - “What percentage of unstructured data changes at least monthly?”

Current Software and Manual Systems Not Adequate

Enterprise organizations are looking to taxonomy software to help solve this problem. Most of the organizations surveyed had some type of system for classifying information within their organization, but most had no formal policy for tagging the documents and left the classifying up to the author. The survey’s
central message—that finding information is extraordinarily difficult and time-consuming—points to the fundamental inadequacy of current approaches.

It’s hardly surprising, then, that even where there is some type of software in place to help with classifying unstructured data, most (over 55%) respondents felt it was necessary and important to put a taxonomy software system in place. Over 90% plan to have a taxonomy strategy in place within the next 24 months.

The objective of this set of questions is to understand if there are systems currently in place for categorizing documents and who within the organization is responsible.

**Question #5 - “Does your organization provide a system for classifying the information you work with?”**

For a slight majority, 55%, there is some system for classifying documents used in the organization. Although not a true taxonomy system, over half were using some type of software or a combination of manual approaches and software for classifying documents.

In contrast, only a third of respondents had a system in place for tagging documents.

“Question #6 - What type of system for classification of documents?”
Enterprise Organizations Demonstrate Interest; But Unclear on Timing and How To Implement Taxonomy

This next set of questions explores specific plans for implementing taxonomy software within the respondents’ organizations.

Question #7 - “Is anyone in your organization working with taxonomy software today?”

Over 75% of the respondents either don’t have or don’t know if they have taxonomy software being used or proposed within the organization. Yet 55%—a clear majority—feel it is important to have a taxonomy system.

The time frame for 90% of the respondents is to have a taxonomy strategy within the next 2 years or less, indicating the likelihood that many organizations will follow-through on the priority placed on taxonomy, as indicated by Question #8. The delta between those who indicate someone working with taxonomy software and those who have an enterprise strategy in place (about 2:1 respectively) indicate that the deployment of taxonomy management software is still very much in the experimental or pilot stages, a notion consistent with other finds throughout the study.

Question #8 - “How important is taxonomy software to your organization’s business strategy?”

Question #9 - “When will your organization develop a taxonomy management strategy?”
There are a number of taxonomy-related issues that the enterprise is unclear about. This is shown by both the diversity of answers and the preponderance of “I don’t know” responses. The data from this survey shows that organizations are particularly uncertain about:

- Who will be responsible for determining enterprise strategy for choosing, implementing and maintaining taxonomy software.

- The costs of acquiring or maintaining the taxonomy software. (This is especially troublesome given that most respondents have a role in determining needs and/or specifications for the software.)

- The topology of the vendor landscape. No vendor is recognized as a leader in this technology segment. The vast majority of respondents could not answer that question nor could they identify any company in this space in any consistent way. This obviously translates to a significant opportunity for one or more vendor organizations to take on this leadership challenge.

### Definition of Terms Used in Discussions on Taxonomy

#### Recall, Relevancy and Precision

Recall is inversely proportional to precision. The more precise you are about defining what you want to look for, the fewer documents are recalled. Using our “chips” example, 2,430,000 documents found represents “recall,” answering the basic question of “How much is out there?” The subset of 24,000 documents indicates relevancy. The 10 documents found regarding the newest Intel processor reflect precision. The ultimate measure of a taxonomy’s success or failure, precision addresses whether or not you found the answer to your question.

#### Hierarchies

As discussed before, building hierarchical relationships is something that humans do inherently. Every piece of information we gather gets placed in a virtual cubbyhole in our brain. That information is not useful until we can relate it to something else. Since many of us are visually oriented, the process of building graphic representations of hierarchical relationships helps us build those relationships faster, allowing us to more readily envision how items are related to each other. Hierarchies also allow us to see both the big picture (with just the main topics), and the details and relationships contained in each “nested” subtopic. The number of nested levels varies greatly according to subject matter, user preference, and the built-in parameters of a given system. The question of how many
levels of information are either necessary or pragmatic is open to debate among systems designers.

**Bottom-Up or Top-Down**

Methods that are used to build categories vary. One approach is to take the details and start placing them in “buckets.” This approach is termed “Bottom-Up.” Returning to our “chips” example, a Bottom-Up approach would start by separating a subset of documents into either the integrated circuit category or the food category. If you run into too many documents falling into different kinds of food categories, you may, for example, need to divide that group into both potato chips and chocolate chips. The Bottom-Up approach thus expands categories iteratively as part of the classification process.

The inverse process is called “Top-Down,” and is how people traditionally categorize. In this approach, domain experts analyze the subject matter at hand, and determine that the documents will fall into, for example, 10 general topics. Each document is then examined and placed them in its appropriate, preexisting category.

**Clustering**

Clustering is a technique for partitioning documents/words into subsets of similar documents/words based on the identification of common elements between the documents/words. Each document can be considered a “bag of words”; clustering essentially groups the similar words contained in each bag.

**Pattern Matching**

Pattern matching is the process of looking for groups of words that are often grouped together. One example is “Business Unit Manager” as a title. (Note: this is also a “noun phrase” that could be processed using semantic analysis.) Other recognized patterns include frequency of words used in a document, placement of words, proximity of words to each other, and clusters of related words. Pattern matching is inherently language-independent.

**Controlled Vocabulary and Thesauri**

A controlled vocabulary is a finite set of terms. The typical application for a controlled vocabulary is for lists of allowed values in structured record systems. A thesaurus is a particular type of controlled vocabulary that represents a formalized description containing a finite set of terms and relations between terms, and frequently also contains information on how the terms are to be applied. The following topics are included in a thesaurus:

- preferred terms
- non-preferred terms
- semantic relations between terms
- how to apply terms

Thesauri are usually in vertical industries such as legal, medicine, and pharmaceutical. An example might be a music thesaurus: Soprano: Broader terms = vocalist, singer; Narrower terms = lyric soprano, coloratura soprano; Related terms = mezzo-soprano, treble.

**Example-Based**

Another approach to classifying unstructured data is to develop a subset of documents that preestablish categories defined by a set of reference content. These “training sets” can be automatic or supervised. The software analyzes new documents in comparison to the training set and searches for similar concepts and ideas. This approach is also referred to as “machine learning.”

A limitation of the example-based taxonomy method is that the resulting classification is totally dependent on the breadth and precision of the training set.

**Neural Networks**

Artificial Intelligence as applied to a computer system is modeled after the neurons (nerve cells) in a biological nervous system. A neural network is designed as an interconnected system of processing elements, each with a limited number of inputs and outputs. Rather than being programmed, these systems learn to recognize patterns. Neural networks are an information
processing technique based on the way biological nervous systems, such as the brain, process information. Composed of a large number of highly interconnected processing elements, a neural network system uses the human-like technique of learning by example to resolve problems. The neural network is configured for a specific application, such as data classification or pattern recognition, through a learning process called “training.”

Spiders

Spiders are automated processes used to feed pages to data extraction and parsing engines. It’s called a spider because it “crawls” over the data. Another term for these programs is crawler.

Taxonomy Market Landscape

Segmenting the Market

The market for classification software is exceptionally dynamic. It is evolving and changing on a daily basis. An overview analysis like this one, therefore, can at best hope to take a snapshot in time. There are many companies supplying this software. The companies discussed in this section are market leaders and are fairly representative of the myriad of approaches and technologies offered within the market. The organizations featured in this section are listed alphabetically, and chose to participate in this project with Delphi.

Methodology Algorithms

This section will deal with the technologies underneath the technologies. As you will see there are many approaches to tackling the problem of building automatic or semiautomatics taxonomies. We will examine each of the technologies from an overview perspective and talk about them in more detail as we examine the individual products that utilize them. These technologies are based on various algorithms using statistical analysis, semantics and neural networks.

There are a number of algorithms that technology vendors customize, optimize, combine and patent in order to categorize digital documents. For a variety of reasons each vendor has chosen a particular algorithm method or combination of methods. This is a list of each of the methods that are discussed in more detail in the following sections:

- Rules-based
- Bayesian
- Linguistic and Semantic
- Support Vector Machine
- Pattern Matching and Other Statistical Algorithms
- Neural Networks

To help understand these methodologies, consider the example of the various types of engines used by car manufacturers today. There are 4-, 6-, 8-, or 12- cylinder internal combustion gasoline engines. There are also diesel, electric, hydrogen or natural-gas powered engines. Increasingly, a car’s power plant might be a hybrid of any of these. Some of these engines are more suitable for the race track than a commute to work, but all will get you to the grocery store and back. These engines have many pros and cons associated with them, and their use depends on a design perspective and what performance characteristics the engineers want to provide to the users.

Similarly, there are many approaches to building and populating a taxonomy. No one method or even a combination of methods seems to yield particularly superior results. Many of the companies profiled below are very technology-centric, and spend much of their marketing effort trying to convince us of the advantages of their approach or methodology. The bottom line is to understand how these differences affect system performance in the only environment that matters—your unique data environment.
Keyword

Let’s start with a brief analysis of search and retrieval as we know it today (that is, without an accompanying classification solution). We will then have a basis for understanding how other technologies try to improve on its weaknesses.

Search engines look for keywords in the title, synopsis or abstract, body of a document, or the meta-tag (i.e., “data about data”) section. A program “crawls” through the subject document, and each instance of a keyword is put into an indexing database that describes how many times it found that particular word and where that word was located within the paragraph, page or document. The “crawling” software looks for unique words but does not include prepositions and conjunctions such as “to,” “in,” “or,” “and,” etc.

The results are analogous to a book index alphabetizing all the words in the book. This is obviously not the most efficient way to locate pertinent information—you may as well just read the book itself.

Advanced functions of search engines, however, allow one to apply Boolean logic to the search process. For example, multiple words can be searched by looking for instances where all the words are present, or where some words are present and others are “not.”

Some search engines also “stem” the word. “Stemming” is process of extracting the root of the word and ignoring plural versions or other modifications of the word. For example, the root of jumped, jumping and jumps is jump.

Accurate spelling and unambiguous terms (not like our “chips” example) are required for keyword searches to be successful. The correct spelling requirement applies to the document as well as the search criteria.

Because of the proliferation of unstructured data, the problem we now face is not finding information about a particular subject, but finding relevant information. In a recent Dartmouth College study, it was found that one in five or about 20% of all Web pages are less than 12 days old. When you consider that Internet spiders take 3 to 4 weeks to index information, the implication here is that you are automatically by default of the process missing 20% of the most current data.

Keyword search systems are actually a subset of rules-based systems.

Rule-Based

An example of “rule-based” taxonomy could be that all documents that include the terms “San Francisco,” “Chicago,” “New Orleans” be listed in a category called “Cities, USA.” The rules break down when an example like “Cambridge” is used. Is this Cambridge in Massachusetts or in England?

Taxonomies were developed to aid in the search for information. This has been done manually in libraries for hundreds of years. Various systems were developed to apply particular “rules” on how such taxonomies have traditionally been built. Rule-based taxonomy classifies documents based on the existence or absence of pre-above. Rule-based classification requires experts to create and maintain a rule for a document to be included in a given category. Experts organize concepts into categories using “If-Then” rules. These rules can support complex operation and decision trees and are very accurate. Rule-based systems have their supporters, then, because these systems can precisely define the criteria by which a document is classified. The rule measures how well a given document meets the criteria for membership in that topic.

Besides the content of documents, rules can be applied to metadata and even business policies: for instance, a rule might specify that only PDF documents created since January 2000 should be

---

6 Mohomine
7 Verity http://www.verity.com/techbuzz/content.html
included in a particular category. Thus rules are a powerful and flexible means for automatically classifying content based on not just content itself but the metadata that describes the content’s business context. The down side of rule-based system is that expensive human domain experts have to write and maintain the rules. Other examples of rules are source of document, age, size and document type.

a) Statistical Text Analysis and Clustering
This technology observes and measures co-occurrences of words. For example, “Java” used in connection with Starbucks probably relates to a document about coffee instead of a programming language. Relative placement of words is important. Words in the first lines of a document are likely more important than information contained in the copyright section. Statistical analysis and clustering also look for word frequency, placement and grouping, as well as the distance between words in a document. Pattern analysis improves precision by resolving ambiguous or multiple meanings.

b) Bayesian Probability
The Bayesian approach attempts to learn the probabilities of words for a given category. An example of Bayesian probability applied is that if a given document contains the words “apples” and “oranges” it is more than likely this document is about fruit, which leads to the assumption that other fruit nouns such as “grapes” or “tangerines” will occur.

Applying a Bayesian algorithm sorts documents by examining the terms, words and phrases contained therein. Bayesian probability uses statistical models from words in training sets, and uses pattern analysis to assign the probability of correlation. This is one of the more common methods applied to building categories and taxonomy structures.

c) Semantic and Linguistic Clustering
Semantic analysis depends on a particular language and dialect. Documents are clustered or grouped depending on meaning of words using thesauri, custom dictionaries (e.g. a dictionary of abbreviations), parts-of-speech analyzers, rule-based and probabilistic grammar, recognition of idioms, verb chain recognition, and noun phrase identifiers (e.g. “business unit manager”). Linguistic software also analyzes the structure of the sentences identifying the subject, verbs and objects, like you did when you first studied grammar in grade school. Then sentence structure analysis is applied to extract the meaning. Stemming or reducing a word to its root also helps linguistic or semantic clustering.

d) Support Vector Machine
Support Vector Machine (SVM) is a refinement of taxonomy-by-example. These algorithms are derived from statistical learning theory. SVM’s calculate the maximum “separation,” in multiple dimensions of one document from another. Each document—essentially a collection of words and phrases that together have meaning—can be represented as a vector. The direction of the vector is determined by the words (dimension) it spans. The magnitude of the vector is determined by how many times each word occurs in the document (distance traveled in each dimension). As this iterative method continuously analyses documents, it separates them into either the “relevant” side or the “irrelevant” space. By repeating the process it categorizes those documents that are “relevant” into like categories, but more importantly learns how they are different.

Combining Methodologies
Of course, no single taxonomy methodology, algorithm, or technology is superior to another for every possible application. The trend by more and more taxonomy software companies is to combine multiple methods to categorize the corpus of documents to increase the accuracy and the relevancy of grouping similar documents.

Mohomine White Papers
Stages of Taxonomy

There are four stages of the taxonomy process. To help clarify this process, consider the example of a traditional book library.

1. Development of the taxonomy structure—the equivalent of the Dewey Decimal System

2. The process of categorizing the content and placing the pointers to the documents in the hierarchical structure—analogous to putting the books on the right shelves

3. Presentation of the information or the interface that helps you find the information—the card catalog would be the library equivalent

4. Monitoring new input and maintaining knowledge assets—analogous to the daily work of librarians; classifying new books, writing up new card catalogs, updating periodicals, and putting returned books back on the proper shelves.

Process

There are a number of subtleties that are not readily apparent when first reading through the descriptions of how these various taxonomy software solutions operate.

One major distinction often lost on the first time investigator is the differentiation between building the taxonomy structure and then populating the resulting hierarchical tree with pointers to the source of the information—in other words, “populating the tree.” These are two separate and distinct operations and one should not assume that these processes are happening simultaneously or concurrently. Rather, these are serial steps.

Below is a diagram showing the general process or workflow of taxonomy software as it categorizes unstructured data. Vendors have their variations on this, but this schematic will give you an overview of how the software basically works. Some vendors leave out steps and other add steps in between these depending on the design approach and architecture of the taxonomy process.
API or Standalone Application

As with many new technologies, a subset of the vendors have developed standalone applications that come complete with end-user browser-based clients and MS Windows-based clients. Users and/or administrators can point the software at the body of documents to be classified on hard disks, servers, intranet sites, portals and Web sites. The taxonomy engines residing on servers perform the categorization process and usually populate a database of metadata.

Taxonomy software can be an enabling technology. It aids and enhances other applications that the user interacts with through a GUI. Most of the vendors support this concept and offer an API to integrate into other applications. Since many enterprise applications are custom-built this is an important consideration.

A number of vendors view this technology as eventually being an even more fundamental component of the information infrastructure. Just as relational databases are a fundamental infrastructure component of applications such as accounting, CRM, and other enterprise applications, taxonomy software will be the infrastructure component that correlate unstructured data. This design philosophy positions taxonomy software as a core module in a suite of products that work on the unstructured data within an organization.

Another series of vendors mentioned in this report have added and tightly integrated the functionality of taxonomy into the search and retrieve application. Here are various ways we can group them:

- Application With GUI
- API Integration OEM
- Integrated Within Search Applications

The companies reviewed here run the gamut from Mohomine—which markets just a development framework for the OEM market—to Quiver, which markets an application suite as a complete end-to-end solution for the enterprise. Most companies offer both approaches, and depending on their orientation emphasize either a strongly developed GUI as an application or a more extensive API for integration.

Company and Product Profiles

The next section details the companies that participated in this project with Delphi Group. The company sections are organized in the following way:

- Introduction – short description of the company and its origins
- Technology approach – a basic description of the underlying technology of the software
- Product – a description of the functions and features of the taxonomy product
- Vision or customer study – a statement by the executives of the technology company relative to where they see the market heading or an example of a taxonomy implementation
- Assessment – a short evaluation of the technology and product approach by the Delphi Group

The companies listed in this section are arranged alphabetically based on the name of the organization. This arrangement does not imply a ranking or endorsement by Delphi Group.
Vendor Assessment Report: Autonomy

Introduction

Autonomy does not require a mutually exclusive decision between automation and manual control. Combining automatic processing with a variety of human controllable overrides removes Autonomy from the purely automatic category of taxonomy vendors. Users can add rules, rename or modify the taxonomy structure to suit their needs. Autonomy’s holistic approach to resolving the business issues associated with all types of unstructured data is one of the reasons the company is growing and profitable.

Autonomy is one of the veterans in the business of developing a software infrastructure to manage unstructured data. Their enterprise corporate customers are companies such as General Motors, PepsiCo, Lucent, Reuters, and the U.S. Department of Defense. They were founded in 1996 and have been an OEM supplier of taxonomy software to vendors like BEA Systems, Business Objects, Sybase and Vignette who have embedded Autonomy’s core technology into their products.

Technology Approach

Autonomy’s strength lies in a unique combination of technologies that employ advanced pattern matching techniques (nonlinear adaptive digital signal processing), utilizing Bayesian Inference and Claude Shannon’s principles of information theory. Bayesian inference is a way of inferring an idea from a set of evidence, and Shannon’s theory allows you to rank the importance of ideas. In the case of text, for example, the words ‘black’, ‘white’, ‘sea’, bird’, ‘flightless’ being present provide strong evidence of the idea ‘penguin’ being expressed. Each word on its own is a fairly weak indicator of the idea to the extent that a single word is ambiguous (even the word ‘penguin’ is ambiguous – English chocolate bar, Batman character, Italian air conditioner, book publisher, bird). The strength of this method is that the presence or absence of individual words doesn’t significantly change the probability of the idea being expressed. This makes matching in the idea space more reliable and accurate than in the word space.

Because it does not rely on key words, it can work with any language independent of slang or regional variations. It treats words as abstract symbols of meaning, deriving its understanding through context of their occurrence rather than a rigid definition of grammar.

Autonomy’s Classification Server is part of a wider infrastructure, the Intelligent Data Operating Layer (IDOL). IDOL provides an integrated platform that integrates content through an understanding of it. Because Autonomy’s technology understands the content it can do much more than simply classifying it. IDOL allows content to be cross- referenced, retrieved, classified and compared with other content. In addition, it profiles users based on an understanding of the content they are working with. These profiles of users are technically identical to profiles of documents or taxonomy subjects, meaning that anything that can be done with content can also be done with users’ profiles (e.g. a taxonomy of users and their interests can be created as well).

The IDOL infrastructure does two fundamental things: aggregates content from all repositories regardless of format or type (such as speech, video, text, XML etc); and builds a mathematical conceptual understanding of each piece of content for later use. It also allows each item to have an unlimited amount of metadata (source, date, author, subject heading, etc) to assist in later manipulation. A subject understanding can then be created by example from one or more pieces of content, or explicitly by sets of rules linked to metadata or keywords. Each of these would be a taxonomy element, or category, which is indexed into a server. As new information comes in, it is matched against the server’s conceptual index (while still using all legacy boolean constructs and keywords a user may wish to define) as one operation (a highly scalable operation – each document in effect ‘queries’ the categories), and a decision is made based on the most closely matching categories. This enables the user to exercise whatever level of individual control they
may want over a taxonomy component, while allowing them to make it a completely automatic process.

**Product**

Autonomy Classification Server provides a completely automated access to all of the information within an organization. It aggregates the information from its respective repositories while respecting all of the underlying security and organizes it into a taxonomy, which is either manually defined or automatically generated, or a blend of the two. Autonomy’s infrastructure technology deals with a static starting point of information as well as automatically dealing with ongoing changes in the information. In addition, it identifies new threads or themes of information.

Autonomy Server had numerous modules:

- Automates categorization, cross-referencing, hyper-linking and presentation of information
- Dynamic updates: can be retrained on the fly eSummarize documents and recommend related article via hypertext
- Provide an easy-to-navigate visual interface for searching providing a unified view.

After structuring the data into taxonomies, the Classification server tags the data automatically, adding relevant metadata tags based on an understanding of the content. Using these metadata tags a hierarchical tree can be generated pointing to the relevant documents.

**Clustering**

The Classification Server delivers the ability to automatically cluster information. Clustering is the process of taking a large repository of unstructured data, agents or profiles and automatically portioning the data so similar information is clustered together. Each cluster represents a concept area within the knowledge base and contains a set of items with common properties.

Autonomy has two types of automatic clustering:

1. Hot News Clustering – identifies the main topics of information present within the knowledge base
2. Breaking News Clustering – compares clusters from the previous periods and compares them to the current one. By identifying new clusters allows the automation of breaking news.

**Cluster Mapping**

The ability to visualize clusters is aided by two applets that identify the relationships between information clusters in one time period or between successive periods and sets of data. These are displayed as either a spectrograph or a 2D cluster map.
Customer Quotes

“There is a lot of useful information out there, but the key to using it is to first make sense of it. This entails understanding the content, sending it to the right person and linking it to other pertinent information and people. Speed is another critical factor, thus all this should happen in a non-labour-intensive manner - that is, completely automatically. That is what Autonomy, uniquely, now enables us to do. It is the engine of our KnowledgeNavigator.”

Jean-Pierre Krause, of Zurich Risk Engineering Group’s Head Office

“Autonomy enables our software engineers and contractors to retrieve accurate and personalized information, which helps them design our earth- and star-observing platforms. The amount of information in our organization is akin to the number of stars in the universe – and we encounter more and more each day. Autonomy’s technology automatically processes growing volumes of data and easily integrates with other products. The implementation has been smooth and straightforward throughout.”

Steve Nauss, Associate Head of NASA’s Computing Environment & Technology Branch for Goddard Space Flight Center’s Information System Center

“Novartis is a world leader in pharmaceuticals, healthcare, agribusiness and nutrition. That means we handle a vast amount of complex information daily, both from internal and external sources. Until now, that has cost us a great deal of time and money in manual processes, but Autonomy’s technology can automate the whole process from start to finish.”

John McCulloch, Manager, Executive Information System, Novartis

“Autonomy is a very powerful solution. It’s easy to maintain, requiring little to no maintenance. We’re now able to focus our resources towards growing our business while saving time and money.”

Bill Gafney, Web Specialist at Boston.Herald.com

Assessment

Autonomy is one of the veterans in the taxonomy software market segment. Working daily with real problems encountered at their installed base of over 500 enterprise customers, gives Autonomy the pragmatic experience of making taxonomy technology work to solve real business problems. Their experience in the enterprise is reflected in the integrated support for over 28 different applications with security-aware plug-ins.

Realizing that building a taxonomy structure and classifying unstructured data is only one of the tools that will help businesses deal with the ever increasing amount of unstructured data generated within and from outside the organization. Autonomy also applies the same technology to understanding the expertise of individuals within the organization. Any enterprise that chooses Autonomy as a technology provider for understanding the capabilities and expertise of individuals within the organization, will be working with a company that not only knows what it knows, but that recognizes how to apply what it knows to real business issues.
Vendor Assessment Report: Convera

Introduction

As an industry pioneer, Convera has been creating technologies for the search and retrieval of text-based information for more than 20 years. Today, the company is marketing and developing intelligent information indexing, retrieval and organization technologies that allow enterprises to take all forms of content—text, images, audio and video—and transform it into valuable information that can be shared among employees and business partners across language barriers.

Convera's categorization solution provides a modular platform for secure organization and search of rapidly changing information across repositories, across languages and across media/file types.

Convera's modular architectural approach allows IT department’s efficiency in deploying and managing a categorization project while their cross-lingual, concept-based categorization solution allows knowledge workers to find the information they need as this information is dynamically organized by how they conceptualize the knowledge.

Technology Approach

Convera’s categorization technology leverages established semantic models that mirror the way people associate different words and concepts. Convera’s approach means that information organized by a given taxonomy reflects how people think about the ideas embodied in related documents. The result is that user expectations are met as they navigate a taxonomy, browsing their way towards the information they need, ultimately finding the information where they would expect to find it. Convera has modeled how they think about ideas in their language, relative to a given taxonomy. This approach is able to respond dynamically to changes that occur in document repositories as well as in the manner in which people think about information.

Convera’s approach also means that organizations need not rely upon the vagaries associated with authors keyword-tagging documents for allocation to particular categories. Authors may well understand a particular subject but often lack “the big picture” when it comes to how a particular piece of information relates to other topics. Convera’s categorization approach frees authors and organizations from this burden by providing virtual domain expertise through the use of synchronized taxonomies and semantic networks in various languages. Convera’s solution also provides categorization of documents across languages so that in multilingual environments users can intuitively navigate repositories of information that contain documents in multiple languages.

Convera has recognized the amount of confusion in the marketplace that surrounds the topic of information organization and has responded by bringing to market a solution that includes not only search and categorization engines but also a categorization personalization feature, domain-specific taxonomies, and related domain-specific semantic networks for most languages. Convera’s objective is to reduce the customer’s confusion and risk by providing a single stop for both information organization and search solutions.

Convera’s basic technology includes the following items:

- Categorization to support browsing and search
- Profiling (personal categories) to alert users to new relevant documents and changes in documents
- Industry specific semantic networks that support concept-based categorization and search
- Pattern search through Convera’s proprietary Adaptive Pattern Recognition Processing (APRP) for recall
- Synchronizers that access content in a wide range of repositories to find and update all the content in the enterprise lowered.
Concept-Based Categorization

Concept-based categorization accounts for the fact that there may be several different words that can be used to express essentially the same concept. It is often the case that industry-specific jargon is developed to more describe objects, concepts or processes that might otherwise be expressed through the use of simple layman's terms. Concept-based categorization enables the controlled expansion of terms within a category rule set e.g., a search for “international commerce” will find documents that contain terms such as foreign trade, import, export global mercantilism and free trade. By using a concept-based rules engine, the administrator can take advantage of subject matter expertise to support a more accurate rule set that is automatically updated and improved with subsequent versions of the semantic network. In other words, a standard taxonomy can be imported and then modified by subject matter experts to match your organization’s needs.

Concept search also does what we naturally do in conversations with each other: it clarifies the meaning of query words through analysis of surrounding words (e.g., the word tank when surrounded by words such as military and vehicle is more likely to be a fighting vehicle and less likely to be a container for holding fuel).

The core support behind concept-based categorization is the standard RetrievalWare Semantic Network, a collection of approximately 500,000 English words that expands to over 1.6 million semantic relationships and idioms that are organized by concept. Although a program can use surrounding words and linguistic analysis to guess at the meaning of a term in a document or query, it is important to provide a category administrator the choice to select the meaning of a term. RetrievalWare's PowerSearch feature allows users to control word expansion of query terms in the Semantic Network, which means they choose which, and how many, related terms to include in the category rules. Users can also choose specific meanings for their query terms. For example, when using the term for banks in a category’s rule set, an administrator can easily instruct RetrievalWare to use only those other meanings associated with financial institutions and not those dealing with the bank of a river or a turning aircraft. In addition it is possible to select individual terms within a meaning to sharpen the rules even more.

Support of Languages and Subject Domains

RetrievalWare includes modular support for more than 25 languages and many domain-specific semantic networks through a Plug-in architecture. This feature acts like a multilingual subject expert that not only understands the concepts and terms in a subject area, but also knows how they are used across languages.

Concept-based categorization that takes advantage of cross-lingual search and domain-specific semantic networks reduces the need for authors or other people to tag the documents because this feature automatically provides these relationships. For example, if a category rule uses the term “central” meaning in or near a middle position, not only could documents containing middle, medial, midway and other relevant English terms be included, but also it could include documents in other languages using terms such as middelpunt (Dutch); milieu (French); Mittelpunkt (German); medio (Spanish) and mezzo (Italian). If the medical domain-specific term bronchus was used in a category rule, other English terms such as windpipe, respiratory tract and bronchial tubes could be automatically be included in the rules, along with terms from other languages such as bovenlip (Dutch); arrière-gorge (French); Atemwege (German); bocado de Adán (Spanish) and bocca (Italian).
This architecture allows Convera’s RetrievalWare to not only reduce the need for tedious and inconsistent tagging of documents but also provides cross-lingual and cross-domain categorization and search. For example, when an English-speaking financial analyst wants to research the potential risk of making a loan to a German pharmaceutical company, there may be relevant documents created in other languages containing technical terms that could impact his decision. With cross-lingual and domain-specific categorization, the analyst may find articles about this company in English, German, Italian and French newspapers that discuss recent loan defaults and technical advances.

Security

One very important aspect of categorization, particularly in intranets, is security. Now that most modern information management solutions provide easy access to enterprise information, there is an increased need to control access to sensitive data. And, the security system must be able to keep pace with changes to access control lists (ACLs). RetrievalWare provides a security infrastructure that respects the native security of the disparate document repositories accessed by RetrievalWare through a single log-on.

RetrievalWare addresses the issues of authentication, distributed, cross-repository and document/library level security for categorization and search. RetrievalWare’s cross repository, library and document-level security ensures that only the documents that a given user should see are displayed, thus preserving the total security of the system.

Products

Convera RetrievalWare is an enterprise-class cross-lingual multimedia categorization and retrieval solution for organizations that have large amounts of information they need to make searchable. RetrievalWare can index, organize and securely retrieve contents from a wide range of repositories and document types, including file systems, over 200 document types, groupware systems, Web pages, XML, relational databases, leading document management systems and even video and scanned paper documents. All of this information can be browsed and searched simultaneously from any Internet-enabled computer running either Internet Explorer or Netscape. RetrievalWare allows users to use simple natural language queries as a basis for categorization and retrieval without concern for search syntax, word usage and other details for content in over 25 languages.

Tools for integration and modification of RetrievalWare

RetrievalWare offers a series of toolkits that enable RetrievalWare to also be used by System Integrators, OEMs, corporate developers, and Convera’s Integration Services Group to modify, extend, or embed RetrievalWare for more custom solutions. The RetrievalWare SDK contains toolkits to enable the integration of RetrievalWare servers as well as the creation of custom client interfaces or modifications to out-of-the-box interface clients. Convera’s RetrievalWare runs on most major operating platforms (Microsoft windows based and UNIX) and uses industry standard Web servers, RDBMS and browsers. Although almost any standard business computer fulfills the minimum system requirements, each customer’s requirements and computing environment are unique and therefore may need additional resources to achieve the desired level of performance and functionality.
Vision

To efficiently provide users with knowledge appropriate for their needs and roles, based on all available information, regardless of location, language or form, and through a system that interacts naturally with the user at his, her or its convenience, anticipating the user’s needs when appropriate and desired.

- Provision of domain-specific taxonomies and semantic networks
- Support for multiple languages in categorization and in search
- Robust security provisions that recognize and respect access restrictions associated with particular repositories and/or documents
- Support for Multimedia information such as image, audio and video files

Assessment

Convera has long been a pioneer in this market and their products reflect the upcoming requirements of the enterprise as the next generation of search and categorization software emerges.

The latest release of their product incorporates many of the future trends Delphi Group sees emerging in this market segment such as:

- Provision of domain-specific taxonomies and semantic networks
- Support for multiple languages in categorization and in search
- Security provisions that recognize and respect access restrictions associated with particular repositories and/or documents.
- Support for Multimedia information such as image, audio and video files

Convera’s advantage is that it embeds multiple information management tools into a unified suite of products that integrate with other business applications in the enterprise. Ultimately, Convera’s software makes employees and customers more productive by helping them manage, locate and access information – regardless of format, location or language.
Vendor Assessment Report: Entopia™

Introduction

Where does the knowledge reside in the enterprise?

The answer is a seemingly intuitive one – primarily in the minds of the individuals who form that organization. Yet most software applications are designed to address the organization’s needs and not the needs of the individuals. Entopia™ has developed software that leverages the concept that the individual’s knowledge is the basis of the enterprise organization’s explicit and tacit knowledge.

Rather than focus on taxonomy as the end, Entopia software uses taxonomy as a means to transform information into knowledge and ultimately into business action. Individuals classify documents into categories as a by-product of the process of collecting and creating the information. When the user finds the information they need, they categorize it either on their own computer system, a shared system of files and folders or the enterprise system data repository. Entopia software automatically suggests the category based on the semantic engine’s transparent analysis of the document, but the individual controls where the information is placed in the knowledge tree and who has access to this view of the document and its relationships to other text data.

Technology Approach

Entopia’s design goal is that the taxonomy structure generation and population of that hierarchy should be a by-product of an individual’s normal work process. As relevant documents are found or created, Entopia software starts a transparent process of extracting key concepts from the document to be the basis of search and a taxonomy. This process is called dynamic semantic profiling.

Entopia’s dynamic semantic profiling process is a two-stage process. The first stage consists of a semantic engine that reduces the collected content to its bare concepts, while a metadata builder adds contextual information and pairs it with the reduced content. The coupled content and context is added to the knowledge base into an invisible taxonomy. The entire first stage occurs in the background of user activity, including adding files, comments, queries, etc. The second stage is a dynamic profiling of users and content that occurs when an individual taps into the knowledge base. User and source profiles are generated at the time of the query, guaranteeing the most up-to-date and relevant results for the user.

The first stage of Entopia’s dynamic semantic profiling process is centered on a series of linguistic algorithms as well as language-independent statistical techniques. This process starts with a rule-based contextual linguistic parser that removes words without discriminatory semantic content, taking into account the context in which they appear. Then the software uses a morphological engine in order to regroup terms in equivalent classes. The document that has now been reduced to its essence is fed to two Neural Networks, which have been designed and trained to extract key concepts of documents and adjust their relative importance to represent the document. The extraction is not only based on the presence of individual terms but also looking at relationships between them, constrained within grammatical and linguistic rules. These key concepts are amassed within the invisible taxonomy.

The second stage initiates when a user queries the system to capitalize on the collected information. At the time of the query, the system recognizes the user on the basis of all work done up to that moment. The invisible taxonomy adjusts to present relevant results personalized to each user by taking into consideration the structure of the query, the user’s unique profile, and the visual taxonomies.

On the visual front end, the user is prompted to place the information that is collected into one of three hierarchical zones:

1. Their personal zone
2. Work group zone
3. Enterprise zone
As contributors add information, the Smart Classification feature suggests “best fit” destinations, based upon the first-stage semantic analysis. The visualization of these personalized taxonomies will be made possible through a Virtual Classification Structure. It allows a single user to personalize his virtual overlay (based on the invisible taxonomy) upon a corpus of documents which may reside in the different zones.

Products

Entopia’s Quantum is a software suite designed to help enterprise organizations improve individual productivity, manage information and facilitate the transfer of knowledge among individuals and workgroups. Entopia characterizes its software by features that fall into three modules: Collect, Collaborate, and Capitalize.

Collect

The Collect module is the front end of the suite, accessed through a browser-based “thin client” or Windows desktop client. Entopia Quantum allows users to gather and save information, from any digital source, into the visible taxonomy. These documents can be MS Office files, PDF files, Web pages, e-mails, and many more file types. Through collection, metadata is automatically generated for indexing and retrieving information and added to the invisible taxonomy.

The result is a Quantum File, or Q-File, that contains:
- The content
- Enrichments (highlights, annotations) and comments
- The semantic profile of the content and its enrichments
- Contextual metadata

Collaborate

This module is a server infrastructure design that enables users within the same organization to share information and collaborate with each other through the workgroup zone in the visible taxonomy. A host of collaboration features are offered including voice and text anchored comments and threaded discussion, e-mail notification, document check-in / check-out, and HTML publishing. In the process of collaboration metadata is also added to the enterprise knowledge base. This collaboration module encourages the synthesis between concepts and the relationships between people that uncover tacit and explicit knowledge within the organization.

Capitalize

The third module is comprised of tools to capitalize on the accumulated knowledge base initially created during the Quantum Collection and Collaboration processes. These tools include Smart Classification, the Virtual Classification System (described above), auto-summarization, and a next-generation search engine, the Knowledge Locator.

The Knowledge Locator is a search solution that allows users to find not only documents but also experts and sources of knowledge within the enterprise. Aware that different goals require different search tools, the Knowledge Locator utilizes both a familiar full-text keyword or phrase search, and a more advanced “semantic search.” To track down documents containing specific terms, only a full-text search is adequate. For “fuzzier,” more conceptual searches, one needs semantic search that yield results “about” or “around” the query. This also reduces the noise by listing only very relevant documents.
While other available search tools may stop there, Entopia’s Knowledge Locator also takes into consideration the context of the search.

Vision

Entopia’s vision is to provide the best place in the enterprise to locate information, which is also the genesis of the company name Entopia, derived from the Greek word entopizo (v. to locate or bring to light) and the more commonly recognized utopia (n. any ideal place or state).

Following Entopia’s core belief that knowledge management software will only be adopted when it helps individuals directly and conforms to their needs, Entopia’s view of taxonomies is that they must:

• Reflect individuals’ thinking first, as a visual “front end.”
• Be customizable based upon individual preference.
• Allow multiple views of the same data (by content, concept, intent, date, etc.).

Secondly, taxonomies must fulfill the requirements of an enterprise, which means that taxonomies should:

• Have an invisible “back end” allowing differing organizational views
• Be dynamic and evolve over time
• Be “soft” (objects may exist in multiple categories)
• Grow with the size of the knowledge base
• Be in tune with natural patterns of the knowledge base

Entopia believes that the user is central to any successful knowledge management system. While many organizations view taxonomies as merely search and navigation tools, Entopia believes that taxonomies are important enabling components of the knowledge worker’s daily decision process. For an individual, the process of understanding and integrating new information is significantly eased with an appropriate taxonomy; a taxonomy that is in tune with the individual, their current interest and the new information. In order to be efficient in this new role, the system has to be flexible and continuously adapt to the current needs of the knowledge worker.

Assessment

Entopia’s software is targeted at, among others, the following verticals:

• Financial Services
• Professional Services
• Pharmaceutical / Biotech
• Government
• Education

and these horizontal roles:

• Research-Oriented Users
• Cross Functional Project Leaders
• Marketing Professionals
• Industry Analysts
• Knowledge Leaders
• Corporate Learning Managers

Entopia’s product incorporates taxonomy technology in a knowledge management solution application that empowers the individual to:

• Discover information
• Convert that information to knowledge
• Expedite decisions

Rather than isolating knowledge in silos of enterprise repositories, Entopia leverages the way the individual works to collect, collaborate and capitalize on the tacit and explicit knowledge within the organization. By beginning with personal productivity, Entopia’s approach effectively captures tacit and often times disconnected information. Since it is implemented on the user’s computer and helps them become more productive while controlling their own privacy, it delivers organizational productivity as a by-product.

Since most organizational data resides on individual computers and hard drives, this information is often lost when an employee leaves, is promoted or moved. Entopia’s software brings back this orphaned information to the organization because it “selfishly” serves the individual where the organizational knowledge has always resided.
Vendor Assessment Report: Mohomine

Introduction

Mohomine’s strategy is to be a supplier of taxonomy technology that is tightly integrated into other applications. Mohomine is focusing on the OEM market segment of enterprise application vendors, infrastructure vendors, system integrators and custom applications within enterprise organizations.

Mohomine believes that the exponential rise of unstructured data and the volatility of the data requires an automatic solution that minimizes manual intervention and increases the end user’s efficiency to realize a greater return on investment.

Technology Approach

Mohomine sees classification as one of the components along with text extraction, clustering and summarization as part of an underlying framework designed to allow enterprises to control the flood of unstructured data. Through a common set of API’s these components are exposed and tightly integrated with enterprise applications. By building a common framework for these modules efficient integration can be achieved.

The method Mohomine uses for building and populating a taxonomy structure is their unique development of the neural network approach supported by seven patents currently filed. They fuse pattern recognition technology with machine learn by example techniques. There are numerous benefits to this technology. One of the primary advantages is that this approach requires few training documents to yield accurate results. The training material can be a minimum set of 8 documents to an entire custom predefined taxonomy. That means that the classifier can adapt to changes in relevance and new topics easily and quickly, thus lowering expenses associated with maintaining a taxonomy. The neural network separates unlike documents from each other by learning the boundaries between categories instead of looking for probable reoccurrences of concepts. Other cited advantages of neural network are speeds up to 300 documents per second and accuracy levels between 90-95%.

Although Mohomine promotes the advantages of automatic classification, the API’s provides tools to change the names of the nodes in the hierarchy to reflect the users needs.

Mohomine’s technology is language independent with support for Western European and Asian languages. The same engine can classify western European languages with custom linguistic processing for each supported language. This linguistic supports removes articles such as “a,” “the” and “and.” Stemming the words to their roots helps identify like concepts for example jump is the stem of “jumping,” “jumped” and “jumps.”

The original design intent of their product is to be integrated into other applications, thus the supplied API can be programmed into another application within a very short period of time.

Products

Mohomine’s Classification software, mohoClassifier 2.0, automates the process of populating customer-defined taxonomies with documents. Mohomine’s products are targeted at enterprise applications such as: Content Management, Search & Retrieval, Enterprise Information Portals, Human Resource Management Systems, Supply Chain applications and Customer Relationship Management applications.

The current product is a Microsoft based development environment with Linux and Solaris platform support under development. Mohomine provides a C++ , com, and http interface to the classifier. Java interfaces will be available soon.

Mohomine uses a technology that learns the boundaries between categories rather than just the similarities. Analogous to the Sesame Street TV game – “Two of these three things are the same and one is different. Can you tell which one is different?” Where the viewer is shown two very similar objects and a third that is different. The viewer is taught to recognize the differentiators, so too does this classifier methodology. This process is applied to text patterns within the unstructured data. This technology needs fewer training documents to “understand” and develop the categorization groups or nodes.
The process is self-teaching through a machine learning function that returns more and more differentiated categories. Because this is a form of pattern recognition, Mohomine’s approach is language independent. As the body of documents grows and changes dynamically the “machine learning” function will grow and change with documents it analyzes.

To augment the neural network methodology, Mohomine uses a linguistic processing module for each language supported. This process understands the morphology of words by a stemming process and sentence structure to provide disambiguation.

mohoClassifier’s architecture and design goals are focused at providing the customer with the fastest and most adaptable classification system possible. There are a variety of features that achieve this goal:

• Flexible Data Classification Options: mohoClassifier can be run in a Server or Client configuration. The Server configuration is best used for batch classification of many documents. The Client configuration is best used by remote clients for lower volume, single item classification. The Client configuration interface is typically used by applications and ASPs.

• Scalability: support for taxonomies of thousands of categories and millions of documents.

• Readily Trainable: been designed to be readily trainable, allowing the user to specify a custom taxonomy and requiring typically between 5-8 items per category to achieve optimal classification accuracy.

• Multiple Categories: documents can be classified into multiple categories.

• Dynamic updates: can be retrained on the fly enabling fine tuning of document classification even after your taxonomy structure has been defined. Additionally, mohoClassifier automatically supports alterations to a taxonomy.

• Multiple Languages: mC supports English and Western European languages.

**Vision**

As one of their early customers and investors, Gilman Louie, president and CEO of In-Q-Tel said “Mohomine’s technology gets to the heart of what we see as the biggest problem facing enterprises and national security organizations today — information overload.”

Mohomine’s senior management vision is to build a horizontal, enterprise-wide integrated solution for unstructured data management (UDM) unstructured, semi-structured or structured data sources. The UDM platform enables the classifying and extracting of concepts and relevant information from unstructured corporate repositories and Web sites automatically, quickly and accurately. The combination of classification and text extraction and summarization is the key for enterprises to capitalize on their core competencies. Giving the enterprise the tools and the platform to build applications that can take advantage of changing information into actionable knowledge is where Mohomine sees the future.

**Assessment**

Mohomine appears to see the potential of being the enabling technology to add structure to unstructured data. By combining complimentary technologies such as extraction and summarization with classification, enterprise applications can automatically supply actionable information to the majority of an enterprise’s users. This will help users find, file and process information in a smarter way. Increasing the value of applications such as Human Resource Management Systems, Content Management Systems, Enterprise Portals and Customer Relationship Management software by correlating the unstructured information buried inside, the value of classification technology increases dramatically. A tightly integrated classification capability will be standard functionality for enterprise applications by 2004.
Vendor Assessment Report: 
Quiver

Introduction

There is more to taxonomy than applying classification algorithms or clustering documents in a hierarchical structure. Since concepts and ideas are relative and most documents contain complex concepts, it is essential to give people the control of where and how these documents fit into a categorization schema.

Quiver’s approach to this problem is an application that allows for the hands-on control and maintenance of the process of taxonomy. In their approach they combine the strength of a manual system – accuracy, with the strength of an automatic system – efficiency.

Delphi Group and Quiver agree that taxonomy is not a onetime process. When investing in taxonomy software users should also think about investing in a taxonomy process. The value of your information is only as valuable as your ability to utilize it to take business actions.

Because of the inherent volatility and subjective nature of the documents, a key to a successful taxonomy implementation is placing the documents where users would expect to find them. If the documents are mis-classified, users will stop using the service. The Quiver system was designed for easy maintenance and adapting the taxonomy to the users and their current and ever changing needs.

Technological Approach

QKS Classifier is a truly hybrid solution to the information overload problem by providing human oversight of the decisions made by the categorization engine in an easy-to-use toolset.

Implementing their variation on the Naïve Bayesian methodology Quiver processes the text contained in unstructured documents and groups related documents in clusters of hierarchical nodes through an automated multi-step process.

1. Parsing or tokenizing converts the documents into a “bag of words” via standard vector format
2. Feature selection scores how informative that word or phrase is for that particular topic
3. By choosing 10-15 documents per topic, these documents teach the machine learning Categorization Engine about each topic and how to classify other documents that match
4. Each document is then tokenized into a hierarchy of topics. Documents can belong to one topic, many topics or no topics at all.
5. Then through a proprietary technology the software assigns a numeric relevance score based on multiple criteria. This relevancy score aids in the classification process. For example, an administrator can set a threshold for all documents with a relevancy score of 85 or higher to be automatically categorized into the machine chosen topics.

But categorization is only the first step, a workflow system was developed by Quiver to permit the user to determine how much of the taxonomy is automatic and how much is in the manual override mode. A major value proposition of Quiver’s product configuration lies in its administration and collaboration environment. The distributed workflow is all organized from a single panel for nontechnical users. This puts the decisions about how documents should be classified in the hands of the content managers and subject manager experts. Workflow Memory increases editorial productivity by remembering decisions previously made by information managers, on a document-by-document basis allowing a machine learning process to make classification more efficient.

The browser-based client houses the directory with the nested hierarchy of related topics. By simply clicking and expanding the topics and subtopics one can quickly navigate to the appropriate concept. When integrated with your corporate search engine via the supplied API, finding the relevant topics is fast and efficient.

Products & Services

The QKS Classifier™ is an application designed for managing the taxonomy process. The primary
differentiator for Quiver is their modular design, which has three components:

1. Categorization engine discussed in the previous section
2. Directory Management Toolset (DMT™)
3. Output and Display Interface

From the Directory Management toolset module via an intuitive single panel:

- Document locations are chosen, which contain the documents to be categorized,
- Training sets developed and run teaching the classifier how documents are related.
- Domain experts can then be assigned to review documents based on qualification criteria.
- Administrators can assign rights, permissions and set limits and hours for classification.

This tool is very flexible allowing multiple individuals to approve the classification of the document with optional administrative control for managing the people and the process. Following the MS Outlook GUI metaphor, this presents a hierarchical structure of folders and sub-folders and files is on the left which can be expanded and contracted with simple mouse clicks. The list of documents for review is listed in a window to the right, with metadata information managed in the panel directly below the main window. With administrator rights users can assign documents for review by your domain experts. These are the people in your organization who understand their area of expertise – be it budgets, or marketing information or technical support procedures. The toolset also allows administrators to filter out documents on the basis of age, size, source and document type.

Targeting enterprise applications such as Search, CRM, content management and ERP systems, Quiver API’s deliver standard XML output to portals, intranets, extranets and other enterprise applications. This output can be either a XML output, standalone or integrated directory view

Drawing on a deep understanding of information classification, ranking and retrieval products, Quiver’s professional services group works closely with each client to identify project objectives, advise and set an appropriate timeline and goals accordingly. They also assist in deployment and integration and ensure long-term, flexible and reliable customized solutions that fit your specific business needs. Quiver services range from taxonomy consulting to UI customization, custom reporting, training and systems integration.

Vision

Quiver’s unique approach augments advanced auto-categorization technology with an intuitive directory and workflow management toolset to provide clear visibility into classification decisions and maximum control of the end-user experience.

Quiver sees as their vision to be the provider of information management solutions to maximize enterprise leverage of corporate information assets. Quiver’s view is that technology alone cannot replace the accuracy of human expertise and contextual analysis. Quiver combines the efficiency of technology with accuracy of human oversight to deliver optimal knowledge solutions for intuitive and comprehensive access to enterprise information. Whether Quiver is deployed standalone or as a part of an employee portal, a customer extranet, or a public website, an organized view of content is crucial to successful information sharing and productivity improvement.

Assessment

To make taxonomy useful for your constituents it must apply their view of the world. A highly interactive application like Quiver will assure those who want control in the creation and maintenance of a taxonomy system to have it easily customized for your users. This approach assumes that the accuracy of the relevancy of the categorization is very important to your organization. If close is good enough, then Quiver’s approach may be overkill.

Quiver products have a solid methodology for developing a useful taxonomy with as much or as little control as your users and management wants. If what you are looking for is an application to develop and maintain your taxonomy with a distributed group of domain experts within your organization, then Quiver is a choice to consider.
Vendor Assessment Report: Semio

Introduction

Like the taxonomies they build, Semio is evolving, changing and adapting to the needs of their users. Semio was founded in 1996 and is on their third generation of taxonomy products. Currently they offer a multi-product line of taxonomy solutions that will expand in the near future.

Founded on information gathered during academic linguistic research, the Semio approach has evolved as experience was gained working with customers in real life enterprise environments. Their suite of products and services will help companies get their information sprawl under control, quickly and efficiently. Realizing that one size does not fit all, Semio has applications, API’s, interfaces and administrative and maintenance tools to get started quickly and then expand and scale to enterprise environments.

Technology Approach

Semio’s approach is to analyze the text of unstructured information via a series of linguistic patented algorithms that identify and extract concepts (key phrases and terms) in documents and statistical clustering techniques. Some of the linguistic methods are identifying noun phrases and proper names, analyzing the recurrence of words and their proximity to each other and the uniqueness of the words to a particular document. Another method would be a comparison of concepts from an imported lexicon or previously defined taxonomy.

No document training sets are required for this linguistic method, which produces a hierarchy of concepts that are very granular and multi-tiered. This approach has yielded as many as 50,000 categories with multiple levels depending on the corpus of documents within the organization. Designed to scale through distributed and parallel processing architecture, some Semio customers have over 13 million documents in their corpus that have been categorized into a multi-tiered hierarchy. The methods of categorization are governed by rules the user can see and modify.

Keeping up-to-date is no problem as the design allows for continuous processing so as documents are added, changed, and updated they can be re-indexed to keep the taxonomy relevant and current.

Pre-Built Taxonomies and Thesauri

Semio has also developed a number of taxonomies for specific vertical markets based on industry standard thesauri that can jump-start the process and provide standardized structures for the taxonomy. Some of these currently available taxonomies are:

- IT
  - Hardware
  - Software
- Healthcare
  - Pharmaceuticals
- Telecom
- HR
- Financial
  - Investment Banking
  - Bankruptcy
- Legal
- Sales & Marketing
- Geography

Vertical taxonomies under construction are: petrochemical, sports, entertainment, religion, education, sciences, Internet, insurance, construction, paper products, food, and government.

Semio has over 30 sets of pre-designed categorization templates addressing various organizations. Since no pre-built taxonomy will fit seamlessly into your organization, Semio provides an extensive tool set to modify, edit, and rename these structures and approved word lists to customize for your organization.
Products

• SemioTaxonomy™ - applications such as corporate sales portals and R&D portals, as well as integrations with Plumtree and Autonomy

• SemioMap™ - visualization of content via graphical maps of concepts and their inter-relations

• SemioSkyline™ - a viewer designed to add browsing and searching capabilities as well as information about the document. This can be customized to reflect your organization's look and feel.

• SemioTagger™ - is the categorization engine that supports over 200 file formats including Lotus Notes and Documentum and runs on Windows 2000 SP1 server with output to either XML, Microsoft SQL or SemioSkyline. This component contains the Knowledge Engineering workbench which is a module for editing the rules that govern the linguistic categorization process. Other modules are crawlers, converters, categorizers, exporters and an administration tool.

A key differentiator for Semio is a “Starter File” in the KE Workbench module in the SemioTagger engine. This “Starter File” can contain 2,000 to 10,000 predefined categories based on your organization's requirements. This “Starter File” answers the basic question of where and how do I start this taxonomy process.

Customer Study

“Semio picked up a tape of corporate data and returned it in less than two days with this same data organized in a taxonomical structure. Through this exercise we were able to demonstrate to our board that we could categorize our data with minimal human intervention,” said John Walford, IT infrastructure manager, Highways Agency.

The project started on January 2001 and the pilot was initiated in March of 2001. This was an integration with Plumtree to provide a portal for 2500 users distributed across 11 Highways Agency sites. Use of Semio indexing and categorization has stripped away and exposed bad data, highlighting what needs to be done and where. It has uncovered data previously hidden or unavailable for analysis. Semio's software has reduced the time and cost associated with knowledge delivery and has improved decision support.

Assessment

With over 100 implementations Semio understands what enterprise customers need and want to turn information into knowledge. Semio realizes that extracting knowledge from information overload is a chronic problem that could be terminal in today's knowledge-based economy.

Semio has a suite of proven products that have been refined by working with enterprise customers to resolve real business issues involving categorizing unstructured information. Addressing the evolving needs of customers Semio will be expanding their product's capabilities to address significant business issues such as security and personalization.
Vendor Assessment Report: Stratify

Introduction

Stratify is highly attuned to the context and associative meanings of words. The company recently changed its name from PurpleYogi (memorable but not relevant to their business) to Stratify. They did this to reflect the company’s focus on creating order (stratification) out of the chaos of unstructured data.

An aptly named product, the Stratify Discovery System helps customers discover knowledge hidden within unstructured data scattered across the enterprise. Consisting of HTML pages, reports, proposals and e-mails, these text documents are located on file systems, Web servers, databases and content repositories. Stratify sees its products as helping executives make better informed decisions by finding the “whys” (supporting e-mails, reports, news items) and correlating them to the “whats” derived from applications drawing on relational databases.

Stratify believes that a taxonomy of important business topics should be an integral part of every major enterprise application. By creating logical topic hierarchies and accurately classifying enterprise documents into them, Stratify provides business critical information to a variety of enterprise applications. Examples of these types of applications are: Enterprise Search Solutions, Customer Relationship Management, Content and Document Management, Corporate Portals, News and Information Aggregation and ultimately Business Intelligence applications. Access to information in an organized and contextually relevant manner will dramatically enhance the value of those applications for today’s businesses.

Technology Approach

Stratify automatically creates a hierarchy of concepts important to a given business. By applying pattern-matching algorithms to a sample set of documents, Stratify aggregates individual documents into clusters and arranges these clusters into a topic hierarchy. The software continuously adjusts the composition of the clusters, identifying outlying documents and duplicates as it proceeds. Stratify can also import an existing taxonomy or provide a pre-built taxonomy containing more than 15,000 topics to jump-start an implementation.

Any hierarchy can be enhanced using additional rules and keywords to define topics. Software tools allow nontechnical people to edit the taxonomy to meet their specific needs. Users can add documents to training sets and test the results of these changes in real time. The tools help the user diagnose and resolve problems in classification. By applying business rules and different filters, the user can specify which sources of information are more important to them. Users can add, delete and link concepts as they wish since this information is stored in a metadata repository.

Taxonomies evolve with time as topics and relationships change. Unstructured content has implicit meanings that people interpret in different ways depending on the current context and their individual interests. Stratify allows the easy manipulation and re-linking of concepts via the metadata contained in the hierarchy, instead of having to reclassify an entire corpus of documents. For example, as references to Clinton change from president to ex-president, the same documents can now be classified together and yet be found under another category as well. This is a good balance of hands-on approach and automatic functionality.

Consistent with Stratify’s belief that different categorization algorithms and methodologies have different strengths and limitations, Stratify uses a variety of classifier engines to analyze and categorize documents into appropriate topics. The Stratify product is designed so that multiple classification methodologies run in parallel. The classification analyses are compared and chosen by using the results of the individual processes.
with patented technology built into the Combiner module. This modular architecture allows Stratify to incorporate new methodologies as they are developed.

Stratify uses industry standard SQL databases for its metadata repository, exposing APIs that allow the import and export of metadata in standard XML format. The ability to easily integrate Stratify’s technology into existing enterprise applications is a feature that will make corporate IT groups appreciate Stratify’s adherence to open standards. Stratify servers are J2EE compliant and can integrate with legacy systems via APIs that accommodate XML, Java, SOAP and COM.

Another unique feature is the advanced personalization capability of the Windows Client that learns users’ interests and alerts them to new documents matching their profile. While the user is in the process of writing a document, the Stratify system classifies that document and presents related documents to the user. This happens automatically, but users can modify their interest profile as they desire. In this way, the user can be made aware of interesting documents without sifting through irrelevant search results.

**Products**

Stratify’s Discovery System has a modular architecture. Stratify packages their product as an application with Web browser and Windows clients. This application consists of the following components:

- Crawlers to collect and extract the text information;
- Hierarchy Builder to cluster the sample set of documents into relevant hierarchical groups based on common concepts, statistical patterns, key words and other algorithms;
- The Classifier analyzes documents based on independent, multiple classifiers, allocates them to appropriate topics in the hierarchy, and aggregates critical information such as document source, title, time/date stamp and the location into the metadata repository. This eliminates the need for manual tagging;
- Management Tools allow users to modify and maintain the categories;
- Metadata database and server to store the data about the unstructured data so that applications can access and view documents in standard XML format; and
- Web Access and Windows Client interfaces allow users and administrators easy access through an intuitive interface.

Stratify also provides a Classification Server that can be used to scale the system, or for use as an embedded product.

To help customers get started, a 15,000-node taxonomy is available with the application. Users can run their internal document against this pre-built taxonomy and then customize it to their environment and needs.
Customer Case Study and Testimonials

One of Stratify’s customers shows that taxonomy is not just a search engine enhancement. Infosys Technologies Limited, a leading consulting and IT services provider, needed to respond more quickly to its prospects and customers to win new opportunities and cut costs. They chose the Stratify Discovery System software to launch a data management system to organize and classify unstructured marketing documentation, such as reports, customer proposals, case studies and competitive research.

By incorporating the Discovery System into their existing set of enterprise software applications, Infosys’ sales and marketing teams are able to create proposals faster. Reusing applicable data while integrating new information makes the proposal generating process more efficient. The Stratify system also gave Infosys insightful competitive intelligence. The return on investment was made in less than five months.

Infosys is integrating the Stratify Discovery System technology into its growing Knowledge Management practice. The Infosys Knowledge Management Service focuses on linking knowledge management to overall business strategy and is holistic in character; it addresses people and processes with technology tools, such as the Stratify Discovery System, playing the role of a key enabler.

“Harnessing the intellectual capital within a company has become essential for competitive advantage and even survival,” said “Kris” Gopalakrishnan S., Deputy Managing Director and Head- Customer Service and Technology at Infosys. “The only effective implementation of a knowledge management strategy is one that captures the intellectual capital throughout the entire organization. The Stratify Discovery System is an important tool for our own consultants in enhancing client relationships and as part of the KM service offering to our clients.”

Assessment

The Stratify target market is the enterprise organization. The ability to run in a Microsoft or J2EE development environment with either Web browser or Windows clients will appeal to large organizations with mixed legacy systems. Focusing on key vertical markets such as professional services, financial services, government agencies, the pharmaceutical industry and biotech market, gives Stratify the reach to a broad base of customers while remaining focused with a consistent message and knowledge of business issues specific to those vertical markets.

Pre-built reference taxonomies tailored to specific industries appeal to customers who want to kick-start their initiative and show immediate results. Software management tools give nontechnical customers the ability to customize and tailor the topic hierarchy to their specific needs as those needs change and evolve over time. Example: do you place documents about Al Gore under politics, education or both, and then link them together?

Stratify’s use of multiple classifier methods and algorithms to achieve more accurate and relevant categories is an approach Delphi Group sees more suppliers adopting.

Delphi Group agrees with Stratify that adding structure to unstructured data helps customers make strategic decisions and discover new ways to utilize the knowledge assets within their organizations.
Vendor Assessment Report: Textology

Introduction

After your organization has categorized the information into a structure – then what?

There are a large variety of solution providers addressing the problems of unstructured text categorization and classification via taxonomy software. Textology drills a layer deeper into documents to add valuable information that helps companies identify, prioritize, and use critical information extracted from text sources. In addition to producing precise categorization, this detailed understanding allows Textology to extract key sentences and combine them to produce concise document summaries, and comprehensive meta-tag data. Textology has approached many of the technology companies mentioned in this report as potential partners and not as potential competitors. There are multiple reasons for this.

1. Textology’s approach improves the granularity of taxonomy structures by basing classification on identifying key concepts and the context in which they are used. This provides greater precision than subjects based on statistical aggregations of key words, and allows difficult categories like “explanations of market conditions” to be accurately handled.

2. Organizing documents into a formal taxonomy is only a partial solution. To achieve greater business value, important details within texts must also be identified and extracted so that they can be used to enable better business decisions. Textology’s summarization, meta-tagging and extraction capabilities provide this capability.

Getting value out of text mining and summarization is only possible if you have enough granularity and detail in the concepts you can identify. You must be able to distinguish a large number of concepts in the corpus of text, and avoid false overlaps by understanding the context in which they are being used. Having done this, you are positioned to gain greater value from documents.

For example, one of the functions Textology can provide is to summarize information linking the primary concepts contained in a document to the key sentences supporting them. This lets users quickly see if a document contains the information they are looking for, and helps them rapidly scan texts by highlighting the most relevant sentences. They can also select the concepts they are interested in and generate unique summaries based on the key sentences related to those concepts. For example, a quarterly financial report might generate a summary highlighting sales figures for one user, and a summary highlighting restructuring costs for another user.

Textology was founded as a joint venture between Elbit, Ltd and Assa-or Systems. Assa-or has produced products for the American and Israeli defense and intelligence communities. Elbit makes focused investments in the e-business and m-commerce markets with infrastructure technologies, cellular applications, and value-added services.
Technology Approach

Textology’s approach to developing a taxonomy is a hybrid method based on combining the best features of linguistic capabilities, machine learning and statistical algorithms. It leverages linguistic patterns to add context that improves the detail and accuracy of statistical techniques, and uses statistical methods to automatically identify and evolve these linguistic rules. These algorithms can be tuned for specific applications to produce optimal results. Taxonomy also give users control over whether they want to increase precision and narrow their results to see only the most relevant items, or do a broader analysis that ensures coverage of all related items. A text, thus analyzed, is then assigned to one, none or many categories based on its identified concepts and their context. Textology’s hybrid method also minimizes the need for large training sets to achieve acceptable levels of accuracy.

Extracting concepts on a fine granular level produces additional consequences such as:

• Precise text summaries with unique versions created for each concept the user is interested in

• Dynamically individualized summaries to reduce reading time

• User selectable summary views (list of key sentences, paragraphs, or simply highlighting sentences within document view)

• Granular taxonomy and easy cross-categorization to other taxonomies

• Comprehensive meta-tag information

• Customizable relevant document headlines or titles

• Detailed data extracts that allow data to populate other applications data requirements.

Products

Textology products are a family of modular components and API’s for integration into enterprise applications. The company is focused on specific customer types where analysis and knowledge extraction from high volumes of high value text is critical to the business processes and decision making, such as:

• Financial Services and News Wire Organizations

• News Aggregators

• Business News Publishing and Services

• Insurance and Legal Organizations

• Healthcare and Pharmaceutical and Biotech Firms

• Defense and Intelligence Services

This market focus allows performance optimization for targeted applications.

Java-based architecture ensures platform independence and embeddable components for easy integration into OEM products. The technical requirements are:

Windows NT or Solaris.

Typical NT server is a 4 CPU Pentium 4 with 512 MB RAM.

Textology provides connectors to standard text and Web document formats. Textology output of meta-tag data supports standard databases and applications. In addition to training and support, Textology provides professional services to install and setup Textology products, and to integrate them into customer application environments.
Vision

Textology is targeting its solutions to customers where there is a high volume of dynamic, volatile text information and high value associated with extracting detailed concepts from this unstructured text. Delivering innovative business solutions for enhancing productivity and creating business value in text critical environments is Textology’s mission. By focusing on a select group of customers, Textology can leverage their expertise and value to address critical business issues.

Example Application

Most financial newswire services use a dedicated staff to daily review hundreds of Web pages, press releases, internal documents, and other potential news sources. For each identified news item they must read it, classify it for relevancy to one or more channel topics, construct a short text summary and/or news headline, create a list of key words associated with the article, identify a list of related articles. They then enter this information along with the URL or location identity of the source document into a company specified meta-tag format. Finding related articles is particularly a time-consuming process due to the imprecise process of sifting through the large number of hits returned by entering the identified key words into an internal Web site search engine.

Staffing is limited by budget constraints, so there isn’t sufficient time to review all the potential sources completely. As a result, quality suffers from broad holes in coverage of potentially important items. New products are difficult to add because they would add additional work to already strained resources. Revenue opportunities are lost because there is no way to expand their services to include articles from additional content providers. In addition, consumers are continually asking for new ways to create even more personalized channels that require much greater detail and flexibility in the way news topics are identified.

Assessment

How much detail is too much and how much is not enough is a sticky problem to solve. Textology’s approach is to provide a very detailed classification and extraction of concepts and ideas contained within large volumes of unstructured text. Solving the problem of information overload becomes more critical when the volume and volatility of the information is very large. Textology’s solution is for those organizations who need a very detailed, granular taxonomy and the ability to extract information and summarize it to make critical business decisions. Textology solutions requires serious consideration.
Vendor Assessment Report: TopicalNet

Introduction

“Quick start” and automatic taxonomy generation are two descriptors that summarize TopicalNet’s approach. The quick start comes from a pre-built taxonomy of over 1 million categories that is an integral value of the TopicalNet solution. This extensive pre-built taxonomy eliminates the need for developing a set of training documents for the software. For some organizations trying to find the “right” training documents becomes a circular problem. You can’t find out what you don’t know because you need to use known documents for training the software. Simple installation requirements make this solution typically up and running in one day.

Quick start is an essential element to TopicalNet’s ROI argument, but so are features to allow the use of existing taxonomies within an organization. Through its mapping technology, they are able to leverage the depth and accuracy of their underlying million plus categories while expressing the classifications using the names and organization of a company’s own taxonomy.

Automatic categorization is based on TopicalNet’s extensive pre-built taxonomy. TopicalNet has over a million categories with an interconnected matrix of topic relationships. The software crawls through your unstructured data, parses the concepts and matches and places them into the pre-built categories or the companies’ categories.

Designed to work with multiple types of environments that range from a single desktop machine, to file servers, to an enterprise information system, TopicalNet is quick to deploy and quarterly updates keep it “topical.”

Technological Approach

TopicalNet’s technology uses semantic and syntactic knowledge to classify documents. The system understands that “biology,” “biological,” “bio” and “biologists” are all related. It knows that “textbooks” and “text books” are lexical variants. It also knows that “Linux operating system” and “Linux” are semantic variations.

Each quarter, TopicalNet fetches over 60 million Web pages. Specialized software analyzes these pages to extract knowledge about the relationships between words, phrases and classifications. One of the outputs from this analysis is a taxonomy of about 1.5 million richly-connected categories. Another output is a set of over 1 million semantic relationships between individual words. TopicalNet also finds and analyzes over 2.3 billion distinct word phrases in the document set. This automatically generated base of knowledge allows TopicalNet’s software to classify new documents “out-of-the-box.”

Products

TopicalNet Classifier is the core product. This is an out-of-the-box solution. A major differentiator is that little training for either the users or the system is required. This means quick deployment and quick ROI. The system is very scalable from one extreme of 60 million pages down to the files on a single PC. Currently, the TopicalNet product offers both an API aimed at custom integration into enterprise applications and a packaged user interface for smaller organizations looking for a quick return on investment.

The product consists of several modules:

- Data acquisition to harvest data from many file types and formats such as HTML, Text, MS Office products, presentations and PDF files.
- The Classifier that develops the hierarchical and interrelated categorizations and scores the documents relevancy.
- The presentation module that shows a browsable directory structure.
A unique application of this technology that the company shows is reclaiming data and knowledge assets that would normally be lost on personal computers of former employees.

TopicalNet assures their users that they have access only to documents allowed by the current security system. The system can run on a number of operating systems and environments such as Unix, Windows, Macintosh and Linux.

Vision

TopicalNet's vision is that classification adds a dimension of clarity, fidelity, and actionability to the analysis of most types of unstructured business data. The software has the capability of defining the thematic attributes to answer the question, “What is this data about?” The software facilitates collaboration between employees on disparate information by organizing the information into recognizable categories. TopicalNet's mission is to develop and market software applications with the lowest possible cost and deployment effort to be applied to a broad range of customer problems associated with unstructured data.

Assessment

Getting started quickly and scaling up to your enterprise's capacity usually are two diametrically opposed features. TopicalNet has been able to address both of these issues in one solution. Using the pre-built taxonomy eliminates the need for developing training sets and the quarterly experience with very large corpora (over 60 million Web pages analyzed) combines scalability with being current and topical. TopicalNet software has the ability to tune and/or train the system to do a better job for specific clients' content and taxonomy. TopicalNet has shifted the burden of responsibility for a successful implementation from the client to this impressive product that does the bulk of the hard work.
Vendor Assessment Report: Verity

Introduction

Verity’s three-tier portal infrastructure software integrates discovery, content organization and social network technology into corporate portals, e-commerce portals and business-to-business portals, along with a wide range of e-business applications.

Verity currently has over 1500 customers, including 80% of the Fortune 50 companies, and is the OEM search provider of over 200 leading software vendors. In the Delphi report Market Analysis: Portal Software Leaders, released in May 2001, Verity was the market leader in the portal infrastructure software category.

Technological Approach

Verity breaks content organization down into four steps:

1. **Build the Taxonomy.**
   Verity recognizes that different businesses are at different stages and levels of content organization. Some may have existing taxonomies (the skeletal navigation structure of nodes/categories), while others don’t know what categories they need. To accommodate this, Verity has the flexibility to build taxonomies in a number of ways. Existing corporate taxonomies can be imported, industry taxonomies can be used, new taxonomies can be built, imported taxonomies can be modified, and/or the corpus of information can be analyzed and relevant concepts automatically extracted.

2. **Build the Model.**
   Before a taxonomy can be populated with documents, a model defining each category must be built. Companies with limited resources may require a completely automatic solution, while others may choose to augment automatic methods with human intellect to achieve specific business goals. With Verity, models can be built using a number of methods: rules defining categories can be generated automatically, rules can be imported from existing taxonomy models or industry taxonomies, and/or domain experts can build new rules or modify imported and automatically generated rules.

3. **Populate the Taxonomy.**
   Using the model built in step 2, Verity automatically populates the taxonomy with documents during the indexing process.

4. **Use the Taxonomy.**
   Populated taxonomies can be used in a number of ways. End users can browse through them by drilling down through broader categories to more focused concepts and individual documents, or they can narrow search results by limiting queries to specific categories. Taxonomy models can also be used to enable content notification applications.
Products

Verity Intelligent Classifier is the application that enables Verity’s back-end content organization capabilities. Intelligent Classifier can be purchased as a component of Verity K2 Enterprise and Verity K2 Catalog, in conjunction with the Verity Profiling Engine for content routing and notification, or as a standalone content organization suite.

Using Intelligent Classifier, businesses can combine what Verity calls the ABCs of Content Organization:

A. Automatic Classification.
Positive and negative exemplary documents can be used to automatically generate the rules that define categories. This employs Verity’s Logistic Regression Classification (LRC) technology.

B. Business Rules.
Domain experts can manually create new rules, or modify imported rules and/or automatically created rules to enhance accuracy or meet specific business goals. The ability to easily modify and fine-tune the rules that make up the model of each category is essential to classifying mission critical information.

C. Concept Extraction.
Verity’s Thematic Mapping can be used to analyze the entire corpus of documents to reveal themes and concepts. This can be used to generate entire taxonomies, break populated categories down into subcategories or to mine enterprise knowledge for new insights. Once Thematic Mapping has identified concepts, it labels them with names that are easy for people to make sense of.

Customer Case Study – The Gale Group

Gale Group Case Study
The Gale Group is a world-leader in e-information publishing for libraries, schools and businesses. Best known for its accurate and authoritative reference content, as well as its intelligent organization of full-text magazine and newspaper articles, the company creates and maintains more than 600 databases that are published online, in print and in microform.

The Gale Group aggregates the widest range and most up-to-date periodical articles on the market. But it wants to offer even more. That’s why Gale selected Verity’s next-generation content organization technology. Verity will help Gale’s editorial team organize the company’s growing collection of content in categories for millions of end-users worldwide.

The Challenge
With as many as 30,000 new articles being aggregated into Gale’s proprietary databases on a daily basis, the company knows the success of its offerings hinge on how easy it is for the millions of students, library users, and business people who access Gale products to find the information they need. That’s why the company demands that its content is organized in the most precise manner possible so users can intuitively locate the valuable content Gale offers.

Currently, Gale harnesses the editorial expertise of more than 150 of its employees to manually classify articles from more than 7,000 periodicals into as many as 40,000 categories and subcategories. But with plans to expand the volume of content it aggregates and grow its business even further, the company realized it needed to augment its manual classification efforts with a more cost-effective solution.
The Solution

After a thorough evaluation of several classification technologies on the market, Gale selected Verity, the portal infrastructure company with the three-tier foundation of next-generation business portals. Of paramount importance to Gale was Verity's second tier, content organization, which combines the ABCs of Classification—Automatic rules, Business rules and Concept extraction.

The company will take advantage of a combination of Verity automatic classification capabilities, including our third-generation Logistic Regression Classification (LRC) technology to create the rules that define categories using both positive and negative examples. In addition, Gale Group editors will fine-tune and customize the rules created by LRC.

The Benefits of Verity

Verity will help end-users—the students, library users and business people who access content from Gale databases around the world—to navigate their way through content from an even larger number of publications.

Verity technology will benefit Gale's editorial team by helping it increase the efficiency of Gale's already stellar classification efforts. Documents that Gale’s editors previously would have read and sorted through themselves can now be handled by Verity technology. This allows the editors to either focus on classification tasks for new material, or to augment the results of Verity’s LRC technology with business rules that capture and utilize their domain knowledge to continually improve results. Additionally, Gale will be in a position to grow its business even more by expanding its offerings and strengthening its product line now that it is using Verity technology to handle more content on a cost-effective basis. This is an important step in the company’s ongoing quest to serve existing clients better—and to continue to build its impressive market-share around the world.

Assessment

Verity is a recognized market leader in search and classification, with a long history of addressing business issues surrounding unstructured and semi-structured information. Verity’s content organization technology effectively and accurately classifies the mission-critical information that is spread across most enterprises, delivering solid return on investment.
Vendor Assessment Report: Wherewithal

Introduction

There are various approaches to addressing the issues of constructing and maintaining an enterprise taxonomy. The continuum ranges from a fully automatic approach with minimal human intervention to Wherewithal’s approach which is a collaborative effort that makes every member of your team—or your entire company—contributors who leverage their own unique knowledge to create a taxonomy for your intranet. Wherewithal uses human judgment to create and maintain the taxonomy. This approach ensures the taxonomy is relevant, personal and up-to-date.

Wherewithal’s approach allows category owners to “opt in.” What this means is that the domain experts can contribute to the maintenance and construction of their category as they have time and need. This encourages a sense of internal “pride of ownership” and an external community pressure to maintain the relevancy of the category. A check and balance approach to taxonomy.

The advantages are that the maintenance and construction of the taxonomy becomes self-fulfilling. The most popular categories will get the most searches and the most positive or negative feedback based on how well the searchers believe the category is maintained. The other advantage is that this approach alleviates the IT bottleneck. The contributors to the taxonomy structure are the ones that have a vested interest in the category, not someone in the overburdened IT department.

Technical Approach

Wherewithal has developed a multi-threaded software engine that holds the structure of the taxonomy in a proprietary format designed for fast response times under very heavy load and extremely large numbers of categories and items. This is called the Collaborative Taxonomy Engine™ (CTE) and has proven its capability by being the core technology of the Xoron.com, an Internet search site.

A patent pending search algorithm—called the Hierarchical Search Algorithm™—uses the hierarchical structure itself to resolve the search. This algorithm is designed to both create the most relevant search results for users, and also to create a system by which many contributors (viz. “tens of thousands” or more) would be able to collaborate on keywords without the system getting bogged down with duplicated effort. This structure is based on the human placement of the document, page and information pointers using the Category Owner Toolbox™ that is used to create and maintain the taxonomy category.

The result of a “search” on this taxonomy structure gives multiple hits based on the keywords placed in the structure by human contributors and the relationship of those keywords to others used in the taxonomy. To ensure results are relevant, and to enable the scalable system of collaboration, results are scored using multiple criteria. These criteria include:

- Lexical “closeness” of the terms used by the searcher to the keyword found in the category item (i.e. word pattern matching—“computers” vs. “computing,” etc.)
- How many times the keyword is found within the taxonomy hierarchy—if the result is found several times within the hierarchy, it is thought of as more relevant to that search.
- Keywords found at higher levels within the hierarchy are given more weight in the scoring process. This means that contributors at “higher levels” of the taxonomy have greater control over search results than those at lower levels.
• The order of the keywords—if they are found in the same order as the user entered them, they are more relevant.

• The “closeness” of the words within the hierarchy. For example if the keywords skip multiple hierarchical levels, the score is lowered.

• The number of keywords used in the contributing categories. The fewer the keywords the higher the score. In other words, if a result is found and the item has only the desired keyword, then it is thought of as being very relevant, whereas if the item is described with a larger number of keywords, the item is less relevant to that particular search. The engine judges both the item itself and the collection of keywords within the given hierarchy for each potential result.

The Hierarchical Search Algorithm also allows categories to be “repositioned” seamlessly in other category hierarchies without changing its keywords. In other words, it frees the contributor to concentrate only on their particular subject versus worrying about that of the entire category hierarchy. This in turn allows others to use that category in other contexts.

This structure is based on the human placement of the document, page and information pointers using the Category Owner Toolbox that is used to create and maintain the taxonomy category. The system also includes a feature called SecondOpinions™, which allows users to customize their view of the taxonomy at each level.

The selected SecondOpinion is scored higher in the search results than opinions that are not selected. This way users’ choices of “opinions” in the search engine gradually customize it to show the results that are more relevant to them (but still include all possible “opinions” in case the selected one doesn’t have the necessary results).

Products
Wherewithal uses a collaborative approach. They integrate the needs of organization and retrieval, allowing users to create a searchable and browsable structure. This in turn creates a resource where end users are enabled to make faster and better business decisions. Contributors across the enterprise opt in to take ownership of categories of their favorite subjects (using Wherewithal tools and an ordinary Web browser).

The process is:
1. Each category gets an “owner” (employee contributor), and that category lives underneath a higher-level category, and so on.
2. Owners create categories, subcategories, and enter items (URLs or other kinds of data) within these categories, along with keywords for each item and category.
3. Together category owners create a knowledge base in the Collaborative Taxonomy Engine, creating a living snapshot of the company's business decisions and structure.
4. Programmers now use the CTE database to build navigation into the intranet, because this database is dynamic and changes with the business. The site is now integrated with site’s search functionality.
5. Using the search box on the intranet Web page, users can search through the collaboratively-built taxonomy, producing search results for both categories and individual items, all scored according to relevance.

The benefits are:
• A living snapshot of the company’s structure and business goals, which changes dynamically with the business
• Site structure and search functionality are seamlessly integrated
• Improving the site’s structure also improves its search results
• Search results are under the control of individual stakeholders across the organization
• A scalable system that allows for reuse of other people’s work and reduces or eliminates duplication of effort

Wherewithal has packaged the software in a suite of tools and applications for the enterprise market in a package called Wherewithal Enterprise Web Directory.

This suite consists of:
• The Collaborative Taxonomy Engine (described above)
• The Category Owner Toolbox for maintaining and creating categories within the taxonomy
• Custom Directory™ - a way to customize the look and feel of directory content
• MultiSearch™ programmable meta-search engine - a way to integrate existing enterprise search functionality into a single set of search results
• The Taxonomy Service API, an XML-based interface into the CTE

This product is designed to develop categories to make searching on intranets and portals more efficient.

**Executive Vision**

Wherewithal envisions a world in which the classification and organization of items for use by others is an integral part of everybody’s daily job. The company’s technology focus is specifically on the issues of bringing together large groups of people to collaborate on a single knowledge base. Wherewithal’s executive team says, “Leverage the Internet’s power to gather knowledge from individuals to create it’s index.” “Someday every person on earth will be a part time or full time infomediary—a knowledge worker dedicated to classifying parts of the internet or a corporate Intranet”

Wherewithal’s perspective on taxonomy can be summed up:
• Knowledge as a “non-fixed taxonomy”
  – Also called a “hierarchical matrix”
  – Also called a “hierarchy with infinite contextual variant branches”

In other words, that things are classified according to a certain context, and that things should be able to be classified in number of ways

• Indexing cannot be done “automatically”
  – It must be done by real people to have real relevance
  – Given enough people (1% of users in any given base), people will make crawlers irrelevant

Wherewithal’s value proposition is:
• A more complete and relevant corporate taxonomy means a better-organized and easier to search intranet, leading to faster searching and browsing
• The collaborative approach allows IT and management to maintain control but saves time in meetings deciding on taxonomy and Intranet structure
• Business decisions are reflected instantly to the Intranet when users decide on new structure, and the Intranet changes as fast as the business does
• Search results are under complete control of the business, vs. a random crawler
• Programmers can use a single repository for structural information versus separate databases and static HTML Web pages
Assessment

Delphi Group sees Wherewithal’s products as overcoming the common obstacles of making a taxonomy timely and relevant to the users by using the collaborative model for constructing and maintaining the taxonomy. This community effort will ensure that the documents and information that most people care about will be placed in the categories where most people will want to find them. An example will be that Bill Clinton will be moved from the category of “president” to the category of “former president.”

This technology has proven itself to scale and work well on the categorizing the entire Internet. A key challenge in implementing the product is integrating its approach with the existing corporate culture and document publishing process. The community method with its carrot and stick approach has a very good chance at keeping the flood of information controlled by the users and the authors of the information. Self-serving as well as visionary this is a technology worth serious investigation.
Controversies & Pitfalls

Taxonomy Strategy: The Time is Now!
As an organization you need to start organizing your data now. The longer you wait, the more unwieldy and overwhelming the task will be. Although manual classification works well for “small” volumes of data, how small is small and how big is big, depends on your resources, requirements and expectations. And due to the inherently ubiquitous nature of Infoglut, even heretofore “small” classification projects are becoming evermore unmanageable.

Our recommendation is that you start with an expert on taxonomy and classification – people with extensive training in the library sciences. Then, interview your internal domain experts and start them setting up agreed-upon categories.

Then decide how big a problem this is. How volatile is the information? How much time do your people spend looking for information? How much do you lose in opportunity costs because employees can’t make an informed decision quickly?

Manual vs. Automatic
There seems to be a lot of “controversy” about manual taxonomy, versus automatic, versus a hybrid of the two. Delphi believes this is a tempest in a teapot and not really a relevant issue, because in order to make the taxonomy relevant to the users, it must match their needs and unique rules for relevancy. This is a good time to reemphasize that there are two distinct steps in constructing a useful taxonomy. The first is the design of the structure of the taxonomy (e.g. the Dewey Decimal System). The second is populating the structure. The range of taxonomy software providers is a continuum of approaches. This range is exemplified from Mohomine representing a completely automatic, strictly parent-child hierarchy to Wherewithal’s and Entopia’s approach, which entail collaboration of participants regarding both the construction of the taxonomy and actually populating the resulting hierarchical structure.

It’s critical to bear in mind that the end result of any taxonomy initiative is a human interface: concepts and ideas are inherently relative, personal and subject to change. Consequently, virtually all taxonomy vendors supply some type of tool to customize and rename the nodes of the taxonomy structure to suit individual needs. The difference between these applications and tool sets is one of degree.

One of the key advantages of an automatic system vs. a manual system is consistency. An important perspective here is to consider whether the people doing the classifying have the same criteria for assigning categories as do the users. Categorizing the same concepts into the same place is what automatic systems do well. If the automatic systems misunderstand a concept, they will at least mis-categorize all related documents and not scatter them in multiple categories.

The decision to adopt a manual, automatic or hybrid approach is a complex one. Whatever your choice, though, your organization should commit sufficient budget and human resources to maintaining an accurate, up-to-date and relevant taxonomy system. The trend in the industry is to combine machines and human processing to develop and maintain taxonomies.

Maintenance and Dynamic Information. As you become more involved in the process of designing and deploying a taxonomy, you will soon realize that this is not a onetime effort. Taxonomy is an ongoing process that requires a long-term investment—business priorities, technology, language, and human interest are in a constant state of flux. The more volatile the information, the more the need for a systematic process to keep the information categorized and relevant. New documents are constantly being added to repositories, while new versions of old

<table>
<thead>
<tr>
<th>Manual</th>
<th>Automatic</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pros</td>
<td>Pros</td>
</tr>
<tr>
<td>Accurate, Logical</td>
<td>Efficient</td>
</tr>
<tr>
<td>Controllable</td>
<td>Limited Accuracy</td>
</tr>
<tr>
<td>Inconsistent</td>
<td>Scalable</td>
</tr>
<tr>
<td>Inefficient</td>
<td>Lacks Control</td>
</tr>
<tr>
<td>Does not scale</td>
<td>Difficult to Train</td>
</tr>
<tr>
<td>Resource Intensive</td>
<td></td>
</tr>
</tbody>
</table>
documents are released, and out-of-date documents are removed from circulation. Changing strategies, evolving products, and advancing technologies all drive changes. Taxonomies may be industry- or even department-specific. Information today is by nature dynamic—consequently, categorization systems must be dynamic as well.

**Directory Building vs. Hierarchical Categories**

Another argument in this field emphasizes the fact that directories are about stored things as opposed to related concepts. Directories are virtual bins and do not necessarily reflect a hierarchical relationship. Proponents of this view would simply ask you to look at the directory structure on your personal computer. Carrying this argument forward supports the idea of customizing directories to help reflect your unique view of the world. The alternative approach is to implement a strict hierarchy with topics and subtopics organized in a strict grandparent-parent-child structure. Delphi believes that each approach has advantages and disadvantages. The choice is one of working style more than substance.

**Granularity of the Taxonomy Structure**

Although there are two distinct sides to this debate, your decision will place you somewhere along a continuum of alternatives. For the purpose of this discussion, taxonomies can be arbitrarily divided into three sizes by the number of nodes or headings and subheadings:

- Small - 1,000 or less
- Medium - 1,001 to 20,000
- Large - +20,000

There are many very large taxonomies. The proponents of large taxonomies say that more is better. Since the organization of information is hierarchical, the users can drill down to as much detail as they wish. Levels of hierarchy greater than 10 are not uncommon in implementations on this scale.

At the other end of spectrum, proponents of small taxonomies argue that more than five levels once again confront users with a kind of Infoglut, receiving too many hits on a search, too much irrelevant information, etc. The third interpretation here is that individuals or work groups can develop their own relevant taxonomies as a subset of large taxonomies.

**Librarians**

Librarians are, first and foremost, people who help you find information. Corporate librarians are experts on how various categorization schema are designed. They know how to find information—that “needle in a haystack.” The idea of an automatic taxonomy system may at first seem threatening to them. The reality is that, because of the dynamic nature of information, these experts will become more and more valuable as the amount of information expands exponentially.

**Users Needs and Personalized Taxonomies**

The needs of individual users represent another major aspect to examine. Will one comprehensive enterprise taxonomy address everyone’s needs, or will you need departmental taxonomies as well? Or will individual workers require their own unique taxonomies? Or will your environment require a blend of all of the above? As you investigate different taxonomy products, be sure to investigate how flexible the products are for generating multiple taxonomies.

**Speed, Accuracy, Robustness and Scalability**

There is no universally accepted standard for evaluating the various algorithms or software configurations in regard to speed, accuracy, and scalability. When your organization is in the final stages of evaluation and has developed its short list of vendors, Delphi Group recommends testing the different solutions against a significant portion of your unstructured data, letting your users verify that the documents are categorized quickly and accurately and on a scale that meets your needs.
Future trends

Topic Maps
“Topic maps are a new ISO standard for describing knowledge structures and associating them with information resources. As such they constitute an enabling technology for knowledge management. Dubbed ‘the GPS of the information universe,’ topic maps are also destined to provide powerful new ways of navigating large and interconnected corpora.”

The best way to describe how topic maps are different from taxonomies is to go back to our “chips” example. We may find that when we develop a topic map around chips, there are recipes for chocolate chip cookies. A taxonomy would list the various recipes and supply the pointers to them. A topic map would “associate” a white chocolate chip recipe with dark chocolate recipes and milk chocolate chip recipes. Then a topic map is formed between “chips,” recipes for cookies, and various types of chocolate chips.

The components of a topic map are:
Topics - A “subject” (or more generally any “thing”) associated with the topic is the name. There are also distinct types of names – base names, display names, and sort names.

Occurrences - These are topics linked to one or more information sources.

Associations - An association is a link element that asserts a relationship between two or more topics, e.g. “Tosca” was written by “Puccini,” “Tosca” takes place in Rome, Rome is in Italy.

For more information on this subject visit http://www.topicmap.com/, a site dedicated to the use of topic maps.

Personalization
“The right information, at the right time, for the right person” has been the mantra of knowledge management software developers for some time. Taxonomies are by their very nature volatile, contextually relevant and personal. An individual’s shifting priorities and goals can at any given time change the way they want to view information categories. If an individual is in an education mode, for example, they may want to discover information about a particular process or a new product. If that same individual is doing a competitive analysis, then they want to look at information from that perspective and context. A number of companies (like Stratify, Semio, and Wherewithal) have added personalization functionality to their taxonomy products. Delphi Group expects this capability to be incorporated in more and more products in the future.

Vertical Taxonomies
Some taxonomy providers (such as Semio) are developing and marketing vertical taxonomies geared toward a particular vertical market such as pharmaceuticals, financial, etc. The proponents of such an approach say this will “jump start” the process for your organization. Detractors, however, maintain that each organization has its own culture and its own way of categorizing.

Trying to use someone else’s taxonomy can be like wearing someone else’s shoes—shoes that are already molded to that person’s feet and will not fit another person even if the size is the same. Despite this, each of the vendors supplying pre-built taxonomies do allow customization of organization and the naming of the nodes for each of the categories. Some of these currently available taxonomies are:

- Hardware/Software
- Healthcare/Pharmaceuticals
- Telecom
- HR
- Financial/Investment Banking
- Legal
- Sales & Marketing
- Geography

Other vertical taxonomies are: petrochemical, sports, entertainment, religion, education, sciences, Internet, insurance, construction, paper products, food, and government.

Taxonomy Integrated With Applications
Delphi Group believes taxonomy software is an enabling technology. The long-term evolution of the market for taxonomy will be its integration into applications such as Enterprise Portals and Content Management. The first integration of taxonomy technology will be with search and
retrieval software. We are already seeing this trend as companies such as Semio and Mohomine develop aggressive OEM programs.

Expect to see taxonomy software increasingly integrated with applications like these:

- Search & Retrieval
- Internet & Intranet
- Portals
- Content & Document Management
- Supply Chain
- CRM & Business Intelligence

**Security**

Security issues are among the key topics on the minds of information managers. Most taxonomy applications follow the standard security model for the server they run on within the enterprise system. This approach can be summarized by this simple rule: if you have access to the area where the document is stored, you have security clearance to see the document. Delphi Group believes this type of security implementation is the minimum level allowable. Security issues relative to unstructured data in the future will involve at least three different aspects:

1. Security attributes of each individual document. When the document is created, various properties will be assigned as to which individuals, which set of clearances, which departments, etc., will be permitted to view or edit the documents.

2. Security attributes of the individual. This type of security will be based on clearance levels, membership in particular departments or assigned role(s) in the organization.

3. Security issues regarding the overall operating environment. Examples of these considerations will be factors such as what time of day access is allowed, access from inside or outside the firewall, number of documents accessed, etc.

For instance, you probably do not want all your employees to see the list of personnel files that are assigned to specific categories of diseases. If you look up oncology you certainly don't want to see John Doe's personnel file linked to that topic. Even if you don't have access to John Doe's personnel file, the fact of its potential association with sensitive topics such as an employee's medical condition is of obvious concern. A number of technology providers (including Autonomy, Convera, Semio, Stratify and Verity) are expanding their security functionality.

**Ontologies**

An Enterprise Ontology is a collection of terms and definitions relevant to business enterprises. An ontology is more than a taxonomy or classification of terms. Although taxonomy contributes to the semantics of a term in a vocabulary, ontologies include richer relationships between terms. It is these rich relationships that enable the expression of domain-specific knowledge, without the need to include domain-specific terms.

An ontology is more than an agreed-upon vocabulary, however. The terms in an ontology are selected with great care, ensuring that the most basic (abstract) foundational concepts and distinctions are defined and specified. The terms chosen form a complete set, whose relationship one to another is defined using formal techniques. It is these formally defined relationships that provide the semantic basis for the terminology chosen.

In the context of knowledge sharing, an ontology is a description (like a formal specification of a program) of the concepts and relationships that can exist for an agent or a community of agents.

An ontology often takes the form of an extremely large database of words and phrases, their meanings and their conceptual relationships. Examples of conceptual relationships are: a “commissioner” is a member of a “commission”; “good” is an antonym to “bad”; and “lumber” has substance, i.e. “wood.”

---

10 [http://www.aiai.ed.ac.uk/~entreprise.enterprise/ontology.html](http://www.aiai.ed.ac.uk/~entreprise.enterprise/ontology.html)
Taxonomies in effect are simplified ontologies. Where taxonomies generally classify categories in “broader” or “narrower” terms, ontologies can include more descriptive classifiers such as “located in” or “part of.”

Beyond Text

A few of the companies mentioned in this report (Convera and Autonomy, for example) are expanding the concept of taxonomy to embrace types of data other than text. Although somewhat governed by the limitation of accurate conversion of multimedia files (consisting of video and audio) into text, the classification engines can still perform adequately to categorize these files by their content. The normal requirements of accurate conversion from audio content to text are not as stringent when pattern-matching analysis is applied. For example, one might view a news clip about former president Clinton and understand its theme without accurately knowing each word that was spoken.

End Note

“Knowing what you know”—that is, enjoying ready access to actionable information—is one of the major factors determining success in today’s knowledge-driven economy. The ability to correlate the “whats” from relational structured databases to the “whys” embedded in the unstructured data of e-mails, reports, contracts, presentations, and Web pages will be a primary driver of innovation and competitive advantage. As in the past, when just-in-time inventory was critical to success, so today is “just in time knowledge” the critical factor.

Delphi Group's research shows that the unproductive time spent looking for information within the digital repositories of the enterprise is growing, and affecting more and more of the expensive personnel we have managing our organizations. Executives, managers and middle-office knowledge workers alike require sophisticated new tools for the delivery of relevant information.

Taxonomy software can reduce our reaction time to make informed and timely business decisions based on the knowledge and information contained within the unstructured data of an organization’s digital documents. This software helps us form ideas from information we didn’t know we had, while revealing relationships and correlations that would otherwise be lost in the ocean of information overload. Productivity and innovation come from seeing connections, evaluating importance, recognizing context, and understanding the implications of these correlations.

Although taxonomy software cannot completely stem the tide of “Infoglut,” it can help us find the information we need to survive and prosper in the new knowledge-based economy—to truly “know what we know.”
Vendor Contact Information:

**Autonomy, Inc.**
301 Howard Street, 22nd Floor
San Francisco, California 94105
Phone: (415) 243-9955
Fax: (415) 243-9984
E-mail: info@us.autonomy.com
http://www.autonomy.com

**Convera**
1921 Gallows Road, Suite 200
Vienna, Virginia 22182
Toll-free: (800) 788-7758
Phone: (703) 761-3700
Fax: (703) 761-1990
E-mail: info@convera.com
http://www.convera.com

**Entopia, Inc.**
1301 Shoreway Road, Suite 302
Belmont, California 94002
Toll-free: (866) ENTOPIA
Phone: (650) 632-0101
Fax: (650) 802-6709
E-mail: info@entopia.com
http://www.entopia.com

*continued next page*
Vendor Contact Information:

Mohomine, Inc.
5120 Shoreham Place, Suite 200,
San Diego, California 92122
Phone: (858) 362-3000
Fax: (858) 362-3099
E-mail: mohoinfo@mohomine.com
http://www.mohomine.com

Quiver, Inc.
2121 El Camino Real, Suite 300
San Mateo, California 94403
Phone: (650) 653-3577
Fax: (650) 653-3578
E-mail: info@quiver.com
http://www.quiver.com

Semio Corporation
1730 South Amphlett Boulevard, Suite 101
San Mateo, California 94402
Phone: (650) 638-3330
Fax: (650) 638-3339
E-mail: info@semio.com
http://www.semio.com

Stratify, Inc.
501 Ellis Street
Mountain View, California 94043
Phone: (650) 988-2000
Fax: (650) 988-2159
E-mail: info@stratify.com
http://www.stratify.com
Vendor Contact Information:

**Textology, Inc.**
150 Post Street, Suite 400  
San Francisco, California 94108  
Phone: (415) 710-6589  
Fax: (253) 369-0253  
E-mail: info@textology.com  
http://www.textology.com

**TopicalNet, Inc.**
800 W Cummings Park, Suite 2900  
Woburn, Massachusetts 01801  
Phone: (781) 932-8400  
Fax: (781) 932-2558  
E-mail: info@topicalnet.com  
http://www.topicalnet.com

**Verity, Inc.**
894 Ross Drive  
Sunnyvale, California 94089  
Phone: (408) 541-1500  
Fax: (408) 541-1600  
E-mail: info@verity.com  
http://www.verity.com

**Wherewithal, Inc.**
2995 Woodside Road, Suite 400  
Woodside, California 94062-2446  
Phone: (415) 705-0726  
Fax: (415) 723-7498  
E-mail: info@wherewithal.com  
http://www.wherewithal.com/
About This Document:
The product-specific information contained in this document is intended to provide an overview of a specific product and vendor at the date of publishing. Facts presented have been verified to the best of our ability with the vendor and actual users of the product where indicated, however, Delphi cannot insure the accuracy of this information since products, vendors, and market conditions change rapidly. Delphi Group makes no implied or explicit warranties, endorsements, or recommendations in this report nor should such warranties be inferred from its contents. A complete assessment of your specific application, the method of implementation for a given product or technology, and the current state of that product must be considered in order for a recommendation to be made on any product's suitability for your purpose, needs and requirements.

Delphi Group is a leading provider of business and technology advisory services to Global 2000 organizations. With offices established around the world, Delphi has assisted professionals across disciplines and industries at nearly every major national and global organization and branch of government. Its clients and subscribers include more than half of the Global 2000.