Enterprise Content Categorization –
The Business Strategy for a Semantic Infrastructure

KAPS Group
Knowledge Architecture Professional Services

Table of Contents

About the Author

Tom Reamy is the Chief Knowledge Architect and founder of KAPS Group, a group of knowledge architecture, taxonomy and e-learning consultants.

Tom has 20 years of experience in information architecture, intranet management and consulting, and education and training software. He has published articles in various journals and is a frequent speaker at knowledge management conferences.

When not writing or developing knowledge management projects, he can usually be found at the bottom of the ocean near Carmel, CA, taking photos of strange creatures.

Introduction: Elements of Text Analytics

Text analytics is a broad set of capabilities that is still being defined, but holds the simple promise of adding a sophisticated language and semantic component to a whole range of processes within business and government organizations. These text analytics capabilities, particularly enterprise content categorization, can potentially solve all those information overload problems that enterprise search and enterprise content management (ECM) promised to solve, but didn't.

What's in a Name?

Text analytics, while not new, has taken off in the last couple of years. As with any dynamic field, there is still quite a bit of confusion about what it is, how to best do it, and how it relates to other linguistic or semantic elements. This paper will address these questions and describe the overall business and strategic context of text analytics. Specifically, we will examine one element within the text analytics field – that of enterprise categorization, and its place within the enterprise. We'll also describe the benefits of enterprise content categorization, and end with some speculation on future directions this field may take. For more on the actual development process for adding enterprise content categorization, see the forthcoming companion white paper *Enterprise Content Categorization – How to Successfully Choose, Develop and Implement a Semantic Strategy*.

However, before we begin, we first need to answer the question of what text analytics means in a broad sense. For some people and organizations, text analytics has largely to do with text mining, the discovery of previously unknown patterns visible when collections of unstructured, electronic text are examined. For others, it is about ontologies – relationships and associations among key words, phrases and entities. For some, it is about sentiment analysis, a newer capability that focuses on characterizing the positive or negative sentiment expressed - about the content of a document, such as products and their features. For still others, text analytics is about content categorization, which is the focus of this paper. When applied at the enterprise level, content categorization is designed to create, maintain and apply the semantic infrastructure of the enterprise.

These capabilities are not mutually exclusive. In fact, the most effective organizations combine these technical advances, such as when they use content categorization capabilities and/or noun phrase extraction capabilities to identify the topic of an expressed sentiment. Text mining can also identify topics, and inform ontology definitions.

Regardless of theoretical subtleties, the best way to define and understand enterprise content categorization in more depth is to lay out a set of core capabilities and features. And while not all applications include all these features, any application with at least some of the features can be considered a categorization activity.

This paper will focus primarily on content categorization as it applies to the enterprise environment, although we will also mention other environments like e-commerce and social media. No paper would be complete without a new acronym, so we will call it Enterprise content categorization (ECC). ECC includes the capabilities and features described in the following section.

Elements of Enterprise Content Categorization: Core Capabilities

Categorization

The most fundamental capability, and often the most difficult to do, is the categorization of the subject matter of documents. To accomplish this, the software must attempt to determine what the document is about. Understanding what the document is about involves evaluating the language the material is written in; how the nouns, adjectives and verbs combine; and what the punctuation means. All these are elements of natural language processing (NLP). Literally, the software breaks apart the words, sentences and paragraphs, and deciphers the content – coming to conclusions regarding the topics, key messages, elements and overall meaning of the text.

Categorization can be accomplished using four basic techniques:

- Statistical – where the software creates a statistical signature of all the words in a set of documents and compares the signature of a document with others in that category.
- Semantic networks – which refers to a set of relationships between concepts and words, including parts of speech and real-world relationships. This can include rules of various types, not just Boolean.
- Linguistic terms – which are basically sets of keywords that are considered (by humans and/or software) to be representative of documents within each category.

- Boolean rules – which are essentially sophisticated search queries that can be used to determine membership in a category.

In all cases, the basic technique is to analyze the text of a document by comparing it to similar sets of documents, or by applying sets of rules to the text to determine if it fits a particular category or subject matter.

While there are pluses and minuses for each of the four techniques outlined above, the relative tradeoffs associated with each are outlined next.

In general, statistical methods are the easiest to develop, but rarely achieve high accuracy because of the difficulty of matching concepts on the basis of statistical measures like frequency of words on documents. The reason is that meaning and scope identification of terminologies from an organization's perspective is usually more complex than that. Linguistic terms are the middle ground – not as easy to develop as statistical methods, because they require more human input – but not as accurate as Boolean. Clearly, Boolean rules are the hardest to develop because they require the most sophisticated human effort; however, they are also the most accurate.

Semantic networks are a special case. They require virtually no effort if you want to apply them to general concepts, because they start with an existing semantic network of general linguistic relationships. But to apply them to any specialized subject area, which is a more typical usage, requires a great deal of effort – at least as much as that of Boolean rules.

This is, of course, an oversimplification on a number of grounds. First, statistical approaches require good training sets, which is not easy for anything more specific than high-level categories like "news," "politics" or "sports." Second, while Boolean rules are more difficult to develop, they tend to be easier to maintain and refine. All in all, Boolean rules are usually the best choice. See the companion paper, E*nterprise Content Categorization – How to Successfully Choose, Develop and Implement a Semantic Strategy*, for a more complete description.

Figure 1: Sample rules from SAS® Enterprise Content Categorization Studio

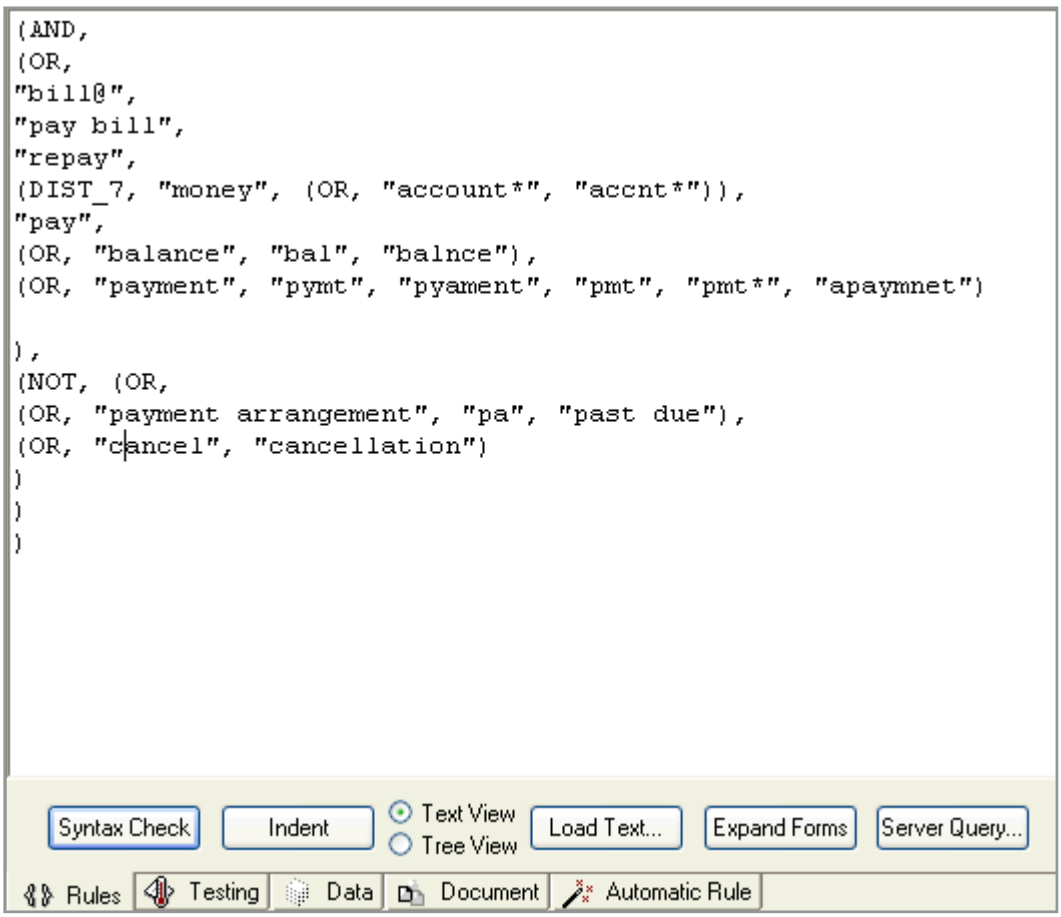


Figure 1, taken from the SAS® Enterprise Content Categorization Studio, exemplifies a number of advantages of Boolean rules, such as being able to incorporate variations in spelling even when there is a lot of variation; the ability to distinguish similar documents with the "NOT" statement; and finally, the power of operators like DIST_7 which says that the word "money" should be counted only if it is within seven words of some variant of the word "account." It is essentially impossible to duplicate that precision with training sets.

Entity Extraction

Entity extraction is the ability of software to recognize specific types of nouns and other phrases, and pull them out of the text of a document. Although this is sometimes claimed to be equivalent to text mining, it is not. With phrase extraction, some of the common types of noun phrases

that have been used are entities such as people's names, company names, and product names; events such as publishing a document; or some other business activity, like selling a product. These entities can then become metadata values used in conjunction with categorization, or – more often -- to feed that metadata[1] into a faceted navigation search interface[2].

Fact and Event Extraction

Facts are sets of higher-order entities; that is, entities with arguments usually expressed in an ontology – a simple model of real world relationships between entity classes – or, in the form of subject-predicate-object (in ontology-speak, a triple), and entity relationships. Fact extraction is similar to and is based upon noun phrase extraction, in that it extracts sets of related entities based on specified triples such as *Person A – works for – Company B*.

Event extraction is a special case of entity and fact extraction. Instead of pulling out all the noun phrases or facts in a document, event extraction identifies words that are associated with events. An event can be anything from a company merging with or buying another company to a company unveiling a new product to a buyer.

Summary

Another basic capability of enterprise content categorization is the ability to generate a summary of a document's contents. This summary is typically somewhere between a snippet (the first 100 or so words in a document), a sophisticated selection of key sentences from the document, and a human-generated abstract of a document. Summaries are often presented with search query results to provide an overview of a document's content.

Clustering

Clustering is software's ability to analyze a document or set of documents and extract sets of frequently co-occurring terms. This includes not just the most frequent terms, but also terms that appear together most frequently to create clusters of terms, or topics. These topics are new, previously unknown groups of terms – and they are more often associated with text mining than with categorization per se. However, this clustering capability

1 Metadata is defined on page 5. [**NOTE**: Be sure to check page number after layout is complete.]

2 A faceted navigation search interface refers to those types of search interfaces that provide a set of facets or dimensions that can be used to filter search results, such as people, organizations, or document source. The facets are designed to supplement relevance ranked search results lists.

is very important to use in conjunction with traditional categorization definitions, because it informs the taxonomy development process, and can be used for elements that may not be classified by the taxonomy itself. Consequently, it provides a vehicle for ongoing development and enhancement.

Clusters are often added as a component of faceted navigation applications, where the cluster of terms is one facet along with more traditional facets like people, organizations, dates and source of content. Clusters are useful as a way to explore new, related content that wouldn't normally be found. However, they are not designed to find specific content, but to discover new, previously unknown content topics and patterns within the text.

Enterprise Content Categorization: Related Semantic Elements

Clearly, enterprise content categorization does not exist in a vacuum. Let's take a look at other semantic elements that are the related products of enterprise content categorization.

Taxonomy and Vocabularies

Thesauri and controlled vocabularies have been a useful, but rather limited, resource for organizations. One important function they play is to establish an official, standard nomenclature to facilitate communication between communities within an organization.

They can be even more useful as a resource for developing a categorization capability. The terms in vocabularies can be used to categorize, but even more importantly, the taxonomies themselves are typically used as the high-level structure, or the foundation, for categorization.

A taxonomy is simply a classification scheme that is normally, but not always, hierarchical. It is a way to provide structure to a set of concepts, ideas, and/or things. The levels of taxonomy are often based on an "Is-A-Child" or "Is-A-Type-Of" relationship. For example, a mammal is a type of animal and a monkey is a type of animal. Very large, multi-level taxonomies have been developed from the standard "life" taxonomy to such specialized taxonomies as the biotechnology taxonomy, MeSH. These taxonomies have tens of thousands of nodes, and ten or more levels.

In the information access world, these large taxonomies are used for indexing each mention of an idea or thing that is in a particular document or set of documents. Their power comes from the hierarchical structure that allows the aggregation of lower levels into higher levels that provide a simple grouping categorization capability. It is also the hierarchical structure that makes taxonomies such a good foundation for developing categorization rules. Categorization can use the hierarchical structure for simple grouping, and then add additional capabilities.

Metadata

Another related semantic element is metadata. There are many different kinds of metadata, and the different types are used in very different ways. Records management, for example, uses mostly system metadata like creation date, document type, and author or publisher. However, for applications that involve finding information, the most important type of metadata is keyword or subject metadata. In the realm of text analytics, the key aspect of metadata is the so-called "automatic" generation of keyword or subject metadata by categorizing applications. It is important to remember that this is not really automatic, because it always requires a great deal of human effort to set up the taxonomies that drive the categorization capabilities. So, beware of vendors who claim that their product will automatically generate all the metadata you need. However, it is true that once the categorization and noun phrase extraction has been developed, it can then be used to generate large amounts of metadata for whatever application you need.

Like enterprise search, content management and taxonomies, metadata has not had a very successful history. However, new technologies that deal with enterprise content categorization have the potential to change that, by reducing the effort required to generate metadata and improving the quality of the metadata. We will take a closer look at how that is done in the section The Hybrid Model of Metadata Generation.

Ontology

The last related semantic element is ontology. In an information or knowledge management environment, ontology is a formal representation of concepts or things, and their relationships. As an ex-philosophy student, I have to admit I found it a bit appalling that the term ontology (the study of the nature of reality) was used to describe such simplistic models.

In ontology, the representation of concepts or things typically consists of a set of types, properties and relationship types. The most basic and common form is called a triple, because it involves a subject, a predicate and an object. For example, you might have the types People and Companies with a set of properties of both types and a set of relationships. From this simple form, you can build large numbers of triples that express relationships such as *People – Work For – Companies and Companies – Pay – People*. From these triples, you can build a detailed model of the entire domain and develop inferences about aspects of the domain that are not formally represented. For example, you can use this model to infer that if a person who works for a company is fired, they are no longer being paid.

From the perspective of enterprise text data management, the relationship is complementary –as taxonomies can be used to add a subject matter component to an ontology-based application. Categorization can also feed the development of identifying instances of triples of subject-predicate-object in unstructured text, through noun phrase extraction and other linguistic capabilities. For instance, it can identify and classify verbs and adjectives for properties. The most exciting area, however, is the ability to integrate structured database content, and semi-structured and unstructured text, to build ever-more powerful and intelligent applications.

Content Categorization in an Enterprise Context

Now that we have an overview of what content categorization and its core capabilities encompass, the next question is: what is it good for? The short answer is: just about anything you can think of that has to do with information. The longer answer will take the rest of this section.

Of course, people who work with categorization to extract meaning from documents have a tendency to get caught up in the beauty and excitement of the process, so they think of it as something intrinsically valuable. The people who pay them, on the other hand, would like there to be a bit more practical value. We will discuss the business benefits in a later section, but first let's look at some of the different ways to approach the enterprise context of categorization.

The first context I'd like to review is content categorization in an enterprise environment, which refers to any organization – commercial, non-profit or government – that has significant information access needs. One reason to focus here is that this is the area in which technologies can deliver the most value – if approached correctly.

If an organization has significant information access needs, then there is a right way and a wrong way to approach enterprise adoption. The wrong way is to view content categorization as simply another IT project that involves doing a standard requirements gathering; doing a standard software evaluation based on features and technical specs; buying and installing the software; implementing an application for whatever group is leading the project; and then declaring the project done.

There are two main things wrong with this "standard" process. First, it does not take into account the need to understand the rich role of language or semantics, and how that requires a different approach than the norm. Second, it lacks a strategic vision that sees language as part of the infrastructure of the enterprise; specifically, the semantic infrastructure.

While it is possible to get value out of a point content categorization project, the way to get maximum value is to develop a strategic vision of how it can fit within – and ultimately transform – your organization. This doesn't mean that your first text analytics project has to be a giant, multi-year effort. But having this strategic vision will ensure that your first and your last project – and all in between – will deliver maximum value, avoid dead ends and duplicate effort, and create the means to solve many information access problems.

Introduction to the Semantic Infrastructure

A good way to think about a semantic infrastructure approach is to compare it with your IT infrastructure. An IT infrastructure creates a platform for applications to run on and enables those applications to communicate with each other. In the same way, a semantic infrastructure is a platform for building a broad variety of applications and enabling those applications to communicate – at a higher level than simply exchanging data – with each other. It has more to do with the network than specific uses of the network.

There are three dimensions to a semantic infrastructure:

- Content and content structure, such as metadata standards, taxonomies and other structures.
- Technology, which can include applications such as search and content management.
- A team of people who are dedicated to maintaining, refining and facilitating the application of the infrastructure's various elements.

It is beyond the scope of this paper to fully discuss the nature of a semantic infrastructure and how to best approach it. But basically the key is to map and understand each of the three elements and how they interact. I'm always amazed at how many organizations do not have a clear idea of what content and/or content structure they have.

None of this is particularly new or radical. The point is that your organization has a semantic infrastructure, whether you consciously know it or not. The difference among organizations relates to how well structured and recognized their semantic infrastructure is.

A tragedy of modern organizations in this age of information and sophisticated knowledge workers is how little effort typically goes into the semantic infrastructure. For example, most organizations have less than one full-time person for search functions, and librarians are seen as fluff. Employees at these organizations then complain about how they are swamped with information and have to spend so much wasted time looking for things. The reason is that they haven't developed an understanding of their semantic infrastructure.

However, the situation may be about to change. The contribution of enterprise content categorization could be the key factor in bringing about that change. However, for that change to happen – and for search and content management and even categorization to really deliver value – organizations need to understand their own semantic infrastructure.

Regardless of the form it takes, and the methods you use to develop it, a deep and accurate understanding of your organization's semantic infrastructure is the foundation that will support all your information and knowledge management initiatives. What you use your semantic infrastructure for is a foundation upon which you can build and support a range of applications. This can range from enterprise-wide applications, like search or content management, to specific analytical projects, like a text mining or customer support application. We will discuss the benefits of this approach in the next section.

Enterprise Content Categorization: Stopping the Bleeding

When considering the potential benefits of text analytics, the basic issue is the cost of poor information access. Numerous studies conducted by IDC, Forrester and others have examined the amount of time lost in

searching for information. Unfortunately, the amount of time and money organizations lose every year is so high that it is hard to believe those studies. But believe it!

I've seen studies that simply add the amount of time employees spend searching (and therefore not working) to the cost of not finding and having to recreate content along with other associated costs. These studies generate figures in the range of US$12 million annually per 1,000 employees. [See the IDC Study, "***The Hidden Costs of Information Work***" – Susan Feldman et al, May 2009.]

Of course, in a number of contexts the cost of not finding information can be much, much higher. While it is harder to quantify, a few stories can help. The first is the cost of an unsuccessful search into the question of whether a competing company has any patents around a particular drug that you're thinking about developing. In this case (which was an actual case), the cost of not finding the right information was five years of wasted effort by a large pharmaceutical company. And what about the cost of missing key documents in an e-discovery environment?

Regardless of how you add it up, the costs of poor information access are huge. And that means that even a small percentage gain can lead to large cost savings. Just using the $12 million a year per 1,000 employees figure, you can see that a 10% increase in time savings is $1.2 million a year per 1,000 employees. If your organization has 10,000 employees, that's $12 million a year. The actual costs, and therefore the actual savings, are much larger when you count all the related costs of poor information access.

That sounds good, but keep in mind that enterprise content management and enterprise search vendors have promised those kinds of savings through improved access to information for over a decade. Unfortunately, the outcome has been continued frustration and wasted time. Let's take a look at enterprise content categorization, enterprise content management (ECM) and enterprise search, to see just how we can change the game.

Benefits of Enterprise Content Categorization

The real problem with ECM and enterprise search has not been the investment cost, unless you got stuck paying $5 million a year. The real problem is that they did not deliver significant improvements in information access. This is where enterprise content categorization rides to the rescue.

Enterprise content categorization improves the quality of search by reducing the cost of adding metadata, by generating higher quality metadata, and by providing more consistent metadata. We've known for a long time that the way to improve search is by adding more and better metadata, but people have resisted the idea because of concerns over cost –misplaced concerns if you look at the actual numbers, not the accounting version of the costs. The idea has also been resisted because of unfamiliarity with issues of meaning, semantics, library science and the like.

With enterprise content categorization, those concerns don't go away, but the cost comes way down to a more manageable level – while the quality improves tremendously.

Improving Enterprise Content Management

Content management is an area organizations have struggled with for years. Many keep trying to get real value from their considerable investment – purchasing new systems over and over again – but few realize any significant improvement. Many of the problems involve complex software, and workflow and publishing policy and procedures. But a great deal of the problem with under-performing ECMs has to do with metadata. Without good metadata, ECM doesn't help with the fundamental problems of finding information, regardless of how it might help with record keeping. And getting good metadata has been plagued with issues, such as getting authors to add any metadata at all; getting authors to add consistent metadata; and, most important of all, getting real value from the most important piece of metadata – keywords.

Having author and publisher metadata is helpful, but it only goes so far. Having good titles and descriptions also helps. But what people primarily know about what they are searching for is the subject of the document – and that is where keywords come into play.

In earlier attempts to develop good keywords, many organizations tried imposing taxonomy, or a controlled vocabulary, on authors who were forced to select keywords from the approved terms. This works better than freeform keywords, but many authors balked at the extra effort required to scan a complex taxonomy for the right word(s), and so the tendency was to select as few keywords as possible and to do a cursory job. Authors, of course, feel their real job is to write the document, not categorize it. Categorization is a very different skill than understanding the subject matter of a document well enough to write it.

The quality of keywords without a controlled vocabulary or taxonomy was even more abysmal. But simply adding a taxonomy turned out not to improve the quality enough to justify the extra cost.

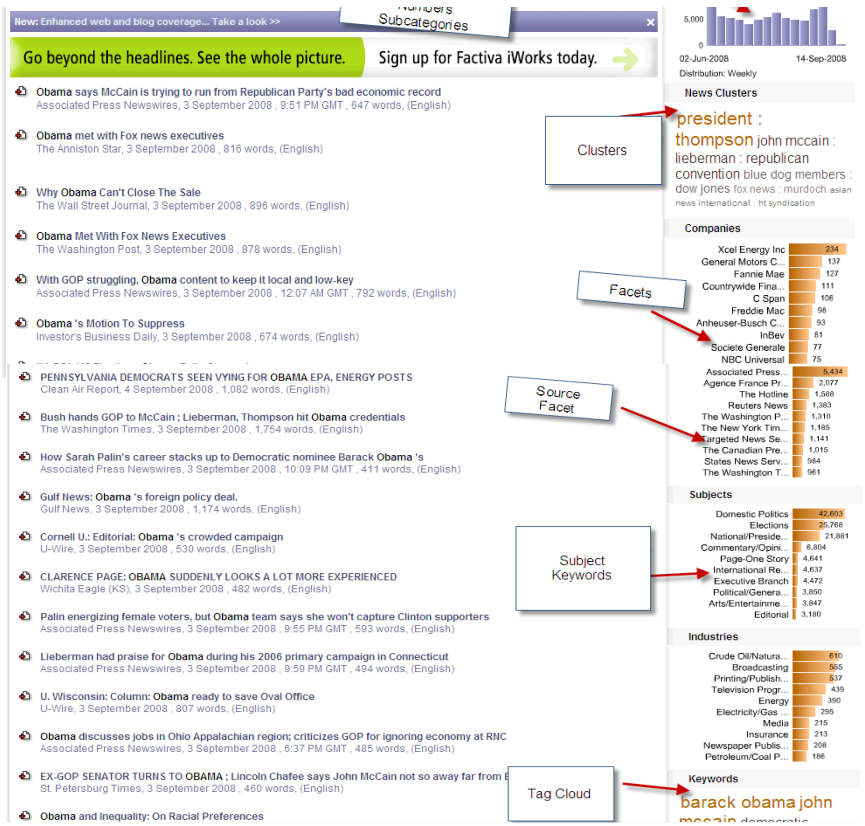
Another solution was to have a team of librarians categorize documents, because they are experts at categorization. This didn't work for a variety of reasons. First, librarians tend to create categorizations that are more suitable for experts, not for novices (both subject matter novices and categorization novices). So, the results were once again poor. Also, the cost of having librarians tag everything can be prohibitively high, or at least it is usually perceived as being too high.

The Hybrid Model of Metadata Generation

Enterprise content categorization can help solve all three problem areas for ECM – if a hybrid model is used. A hybrid model consists of authors, editors-librarians, and enterprise content categorization software. In this model, a document is first submitted to the ECM system, which is integrated with enterprise content categorization so that the categorization software can analyze the document and generate a wealth of metadata. Some of this metadata can be generated through the ECM software – because things like author and publisher can be automatically generated. What enterprise content categorization adds is a way to generate metadata for the problematic areas – descriptions and keywords.

The key is to get all three pieces working together. The software can generate or suggest a description (summarization capability), and it can use the enterprise content categorization noun phrase extraction to generate a variety of facet values, such as by finding all the product names, company names, people, etc. that are mentioned in the document. It then lists those in a faceted display. Finally, it applies categorization to suggest the main topics of the document. For an example of a faceted search interface, see Figure 2.

Figure 2: Metadata from SAS Enterprise Content Categorization extends the rudimentary concepts from content management systems to deliver more meaningful, searchable facets.



At that point in the process, authors face a much simpler task than trying to create all the metadata themselves. And let's face it, it's hard enough to get authors to think up a few keywords, much less create all the metadata that faceted navigation needs. In this new model, the author is faced with the simpler task of agreeing with the suggestions made by the software, or rejecting them. Most of the time, I imagine authors will accept most, if not all, of the software suggestions. In cases where the software makes one of those silly suggestions that plague all "automatic" systems, it is very easy to spot that remark and delete it.

The role of the librarian or taxonomist is now very different. Instead of asking them to tag the documents themselves, they develop the categorization taxonomy and then monitor how it is used by the authors.

This is the way to get metadata into your ECM repository and thereby finally get full value from all the money, time and effort that have gone into your ECM initiatives.

Improving Enterprise Search with Faceted Navigation

The hybrid model works well for conventional ECM and for a conventional enterprise search application built on top of it. But it works even better for the newer, more effective approach to search – the one that uses faceted navigation.

Faceted navigation is one of the most important new approaches to enterprise search, as well as to e-commerce applications. However, faceted navigation requires a lot of metadata. And that is where enterprise content categorization rescues all the money, time and effort that have historically been poured into enterprise search.

Faceted navigation took off first in the e-commerce arena for two reasons. First, there was an immediate and obvious economic payoff by making it easier for customers to find what they wanted to buy. Second, e-commerce had product catalogs filled with all the metadata that was needed, like type, model, price, feature lists, and more.

For search in general, and for enterprise search in particular, the situation was much different. There aren't any predefined and pre-populated facets with all that metadata, so it all has to be added to an organization's unstructured content. Also, with enterprise search the variety of information needs is much more complex than with e-commerce, where the main activity involves buying something. While the basic solution for generating metadata for enterprise search is to integrate it with enterprise content management, many issues make the situation more complex. This starts with design issues, like determining what is the best balance between ease of finding information (the more facets the better), and the effort required to generate that metadata.

Enterprise content categorization changes the entire debate by making metadata generation much easier and more effective. However, it doesn't eliminate the tradeoff entirely. While simply moving the best tradeoff point might not sound like a lot, it is actually *essential* for one very important reason.

I have worked on a number of enterprise faceted search projects, and the one thing that can kill the value of a faceted navigation application is having

too few facets. This situation is often driven by a desire to minimize the cost of adding metadata. The right number of facets is a design question that depends on a variety of factors, but the only designs I've seen that simply did not work at all were the ones that offered only two or three browse structures for searches.

In general, facets work so well because they offer multiple filters designed to work together and to support a variety of users with different interests and different types of knowledge. With too few facets, the entire user experience changes from a set of simple selections from well-defined facets to a browse through a complex structure – which is much more cognitively difficult. What enterprise content categorization adds is the most essential facet of all, subject matter expertise.

But what about a situation where you can't add metadata through a content management system? This situation arises for Internet search and for enterprise search systems that incorporate external content into the index. In this case, enterprise content categorization can be directly integrated into the search engine and used to "automatically" generate metadata. The term "automatically" requires clarification for two reasons. First, it often works best in that environment to have editors work with the output to refine and fine tune the application of categorization capabilities to incoming content. Second, metadata creation is only automatic after the development of good categorization and noun phrase extraction capabilities.

Beware of companies that claim their solution is fully automatic. What they really mean is that with a lot of effort developing and refining series of training sets of documents, the software can do a reasonable job of generating metadata. And even if it does a reasonably decent job of generating metadata, "fully automatic" systems will not continue to learn and develop over time. For that characteristic, you need the subject matter experts to step in and refine the results.

Obviously, a fair amount of human effort goes into developing taxonomies. But once they are created, it is possible to simply point the software at incoming content and have it automatically generate metadata using categorization, summarization, and noun phrase and/or fact extraction to populate a search interface. For a good example from early adopters of this approach, check news website searches. These software tools can populate facets, such as people, organizations or companies, and even subject terms. This capability is often combined with clusters for exploring

related topics and tag clouds. Enterprise search can use the same techniques to incorporate external content or other federated content that might not have gone through a content management metadata generation cycle.

Foundation for Informed Search-Based Applications

According to a number of analysts speaking at the Enterprise Search Summit in New York 2010, search-based applications are where most of the action in enterprise search is going to be for the next few years. Search-based applications include a huge range of technology capabilities, including business intelligence, customer intelligence, social media, content aggregation and/or cleanup, call center applications, and much more. What they all have in common is the need for better metadata to make them really work. This is what text analytics, particularly enterprise content categorization, can provide.

This area also includes one of the other core capabilities of text analytics – sentiment analysis – which has taken off in the area of customer intelligence. Knowing that people are talking about your products and services is important, but knowing that they are saying positive or negative things is even more important. Developing sentiment analysis capabilities is very much like categorization, so the best vendors of text analytics are combining them. In fact, having categorization, noun phrase extraction, and sentiment integrated into a single application enables development of much more sophisticated applications.

By adding structure to unstructured content – through categorization, noun phrase extraction, ontologies and the like –both structured and unstructured information can be integrated to create even more powerful and useful applications. What those applications are will vary, but can include:

- A variety of e-discovery applications.
- Smarter customer call center applications that can utilize unstructured content.
- CRM applications that add semantics and intelligence.
- Expertise location to supplement any Enterprise 2.0 initiatives.
- An intelligent work space – where search is integrated into applications so employees can find information without interrupting what they are doing.

Efficiency Gains: Saving Time

There is one final, important point to make about the benefits of enterprise content categorization. Too often, it is easy to dismiss all the time savings that form the basis of so much ROI literature, because it's hard to see how it really leads to true bottom-line value. The feeling is that the time employees save is soft money that doesn't show up on the accounting reports.

To counter this feeling (and it is mostly a feeling rather than any hard numbers), some analysts have started trying to come up with better ways to think about time savings. One very good way is to express time savings in terms of the number of full time employees (FTE) that could be hired without raising the operating costs of the company. Companies can generate a spreadsheet of the additional projects they could undertake with those extra employees, and then estimate the increased revenue and profit those extra projects could produce.

Another way to think about time savings is to take the money saved and pass it on to your clients rather than hire additional staff. The two are not mutually exclusive, of course.

A third way to think about time savings is to generate a spreadsheet of what your existing employees could do with extra time. With less time spent searching, they will have more time for doing their jobs and producing value. Or, they might spend some of that time thinking and working smarter. Proposals could be produced with less effort and time, reports could be written on schedule, and customer support people could spend more time helping customers and less time looking for answers. Employee productivity could be improved too, because less time would be spent looking for – or worse, reinventing – existing documents that couldn't be found.

Finally, improved access to information enables organizations to build a more integrated enterprise with improved communication and collaboration. It has even been shown to improve worker morale and satisfaction. Nothing is more frustrating than the feeling that you are wasting time instead of doing a good job.

Future Directions

It is nearly impossible to predict where a field as new and unsettled as this will go in the next five years. Rather than try to make any full prediction of all possible directions, what follows are some ideas and possibilities that do not try for completeness. The future is very bright and very open.

New Features: Building Blocks

Currently, every enterprise content categorization project basically has to start from scratch, and this is costly. I don't see this completely changing, because conventional taxonomy projects almost always involve at least some customization, and categorization taxonomies are complex and dependent on the individual corpus of documents. However, as with conventional taxonomies, there are ways to speed up the development process.

Just as many existing taxonomy projects can find "starter" taxonomies that can be adapted to a new corporate environment, as more categorization taxonomies are built, it should become possible to find and adapt existing "starter" categorization taxonomies. "Starter" taxonomies are high-level categorization taxonomies for specific industries or subject areas that cover the broad, general categories and structures of that area. As more easily adaptable "starter" taxonomies become available, the development process will speed up considerably. For example, if you had a high-level federal government categorization taxonomy based on public information, you could use that to develop a deeper taxonomy that could be applied to internal content. This starter taxonomy could also be used, at a much lower cost, to develop agency-specific categorization taxonomies.

A related area could be the development of standard building blocks of categorization rules for specific areas. For example, there are certain actions that a customer support group engages in that are somewhat independent of the specific products supported. There are even more actions in common when you look at specific industries, such as computers, phone or wireless devices, and so on. If a standard categorization taxonomy is designed with this in mind, it should be possible to develop components, or building blocks, that can mixed and matched, and then combined with more product-specific categorization rules. This should further speed up the process of development. For example, if you had a categorization taxonomy that covered basic customer support actions for

a phone company, you should be able to combine it with a catalog and/or categorization taxonomy for each phone company's products.

New Features: Full Platforms and Embedded Applications

Within the text analytics vendor community there are two seemingly conflicting trends. One trend is toward full-featured platforms with all text analytics capabilities, and the other is toward embedding specific functionality within applications. There are one or two rare vendors who do both. Some text analytics companies have virtually disappeared within business intelligence or customer intelligence application companies, or inside content management applications. And, with the current emphasis – some would say hype – on social media, these kinds of embedded applications will certainly continue, despite concerns about how difficult it is to get real value out of social media applications.

Faceted navigation applications took off first in the area of e-commerce, and are now being applied more slowly to the larger enterprise market. I see a similar trend for text analytics. The enterprise market will continue to call for the full platform of text analytics offerings. As with other information technologies, organizations may initially purchase text analytics to solve a specific problem, but over time they find more and more applications for their new "toy." I refer those companies to the section in this paper: Content Categorization: in an Enterprise Context.

Overall, I see the enterprise market continuing to grow just as enterprise search did, but with better results for the customer. It probably won't be a rapid bubble of growth until a basic level of awareness is achieved in the market. That can take time, because many organizations are not aware of the full potential for text analytics. The other factor that might slow the growth of text analytics is customer fatigue. After all, enterprises have heard it all before from search and ECM vendors, and they will likely need to be convinced that enterprise content categorization really is different, and that it really does have the potential for dramatic improvements in finding information. Hopefully this paper will contribute to that effort.

New Kinds of Applications

While text analytics has the capacity to enhance existing applications like search and content management, I expect to see new products and new product areas developed as more people become aware of all the capabilities text analytics offers their field. One such new product is expertise analysis, which builds on sentiment analysis.

Sentiment analysis has been a primary focus in text analytics for the last year or so. However, the same technology that can do sentiment analysis can be adapted to other areas – one of which is expertise analysis. Instead of setting up rules and applying statistical models to extract positive and negative sentiment from text, the same approach can characterize how expert the writer of the document is in various subjects mentioned in the document. Consequently, it can rank the writer as expert, general or novice.

When expertise analysis is combined with the idea that experts and novices operate at different levels of specificity of categorization and language, there are a whole range of possible applications and/or enhancements to existing applications, such as:

- Taxonomy/ontology development – will impact everything from evaluating user contributions, to having expert-general-novice versions or sections of the taxonomy, to how to present the taxonomy.
- Search – will include query analysis, information presentation and different clustering algorithms.
- ECM – the software will suggest different levels of tags.
- Social media analytics – will identify community expertise and build on that in a number of ways.
- Knowledge management – expertise location – will not be dependent on self- or community-ranking.

Finally, as more people develop more sophisticated applications and do more research into our understanding of categorization, I expect that new capabilities based on that deeper theoretical understanding will be developed and added to the text analytics mix. Where that will lead to is anyone's guess. It should be quite exciting to see the results!

Strategic Vision

To build on that last thought, let's close with a strategic look at the future. The first part of that vision is to realize that when you add intelligence to enterprise applications, the possibilities are staggering. Imagine all your business processes being streamlined because they are now at least partially automated. Remember all the excitement about how we were all going to create new and more intelligent enterprises with the dawning of the information age? That was before we got swamped with information overload and the growth of information silos. Well, enterprise content categorization has the potential to help fulfill some of that promise, if not fulfill those grandiose visions.

To achieve these dramatic results, you must first create the foundation for building an integrated enterprise – not technically integrated into a single network, but semantically integrated. We thought that federated search might have been able to do this, until we realized that simply having more access to more information doesn't solve anything, but getting smarter access ***does***. Consider one small example. Instead of trying to rid your organization of information silos, enterprise content categorization and ontologies allow you to take advantage of why people build silos. Silos are built because they can speed up access to information by limiting it, and by creating shorthand to refer to it (otherwise known as jargon). Silos are based on shared interests and common information needs, and it's unlikely that we could ever get rid of them. In fact, we don't want to get rid of silos.

As it turns out, with enterprise content categorization we won't have to get rid of silos. Instead, we can build bridges between silos that retain all their advantages, while enabling a sophisticated, inter-silo communication to take place. And that will enable us to semi-automatically translate jargon and special concepts and local categorization preferences between all the tribes that exist within a modern organization.

So – what is possible with a semantically integrated enterprise?

Use your imagination!