Auto-Categorization: Coming to a library or intranet near you.

Automatic categorization – it's coming, it could be big, and it could hurt. The only trouble is that it's not automatic. I have yet to see a product that did not need or was not improved by human intervention. But, other than that, almost everything they are saying about it is true.

In this article I will offer answers to a few simple questions that need to be asked when exploring this new type of software: What is it? Why is it suddenly everywhere? What can it do? What can't it do? How should information professionals approach it? And is it dangerous?

What is it?

Simply, automatic categorization is a new type of software that assigns documents into subject matter categories based on a wide variety of techniques. These techniques include statistical Bayesian analysis of the patterns of words in the document, clustering of sets of documents based on similarities, advanced vector machines that represent every word and its frequency with a vector, neural networks, sophisticated linguistic inferences, the use of pre-existing sets of categories, and seeding categories with keywords.

From this list of techniques, it's pretty clear that none of this software categorizes the way humans do. And that is both its strength and its weakness.

There are also a very large number of companies offering their version of this new software and, of course, most claim that their approach is the best, the fastest, and the smartest. And they are all wrong. And they are all right. In other words, the whole product space is wide open with no clear leader and no clear correct or best approach. And this is what makes it hard for the information professional to evaluate.

Why is it suddenly everywhere?

An informal survey puts the number of categorization companies at nearly 50. And more and more search and content management companies are scrambling to incorporate categorization into their products. Why has there been such an explosion of companies in the last two years? It seems to me that the answer is twofold. First, the development of new techniques in recent years has allowed companies without major resources to create versions of categorization software that work as well or better than existing software from the early leaders like Autonomy and they are offering them at ¼ of the price.

However, the real reason that auto categorization has taken off in the last year or so is much simpler – it's that search stinks and users can't find anything and categorizing content enables a browse or search/browse functionality and users prefer browsing and browsing is more successful than simple keyword search and facilitates knowledge discovery and even after years of trying to teach users how to do advanced searching, they won't.

Oh, and one more factor in the above list of "ands", company's don't want to pay librarians to categorize their content because they think it's too expensive. They are wrong, at least when you factor in the time employees waste trying in vain to find that document that they just have to have to answer that customer's question and the customer just left and went with their competitor. Despite that, they still won't pay for humans to categorize their content, but they are more likely to pay anywhere from 250K to 750K for software that does a worse job.

What can it do?

The first and best thing it can do is to very quickly scan every word in a document and analyze the frequencies of patterns of words and based on a comparison with an existing taxonomy, assign the document to a particular category in the taxonomy.

Some other things that are being done with this software are clustering or taxonomy building in which the software is simply pointed at a collection of documents, say 10,000 to 100,000, and search through all the combinations of words to find clumps or clusters of documents that appear to below together. It's not as successful as the first use, but it can be an interesting way of aiding a human in creating or refining a taxonomy.

Another capability that can be found in some of the software is the ability to create an automatic summary of a document. Of course, it's not really a summary, certainly not in the way a human creates a summary. Rather, the software scans through the document and tries to find sentences that are important. Mostly, importance is measured through general rules that have little to do with the meaning of the sentence, but rather, are based on such known rules as the first sentence of the first paragraph is often important, the last sentence of the first paragraph is often important.

This type of summarization has not shown itself to be particularly valuable as a component of a search result list, but by tying the summary to a categorization rather than a keyword search term, the result can be more valuable. This is particularly true if instead of using the summary as a means for users to determine if a document is the one the user wants, the summary is used at categorization time by an editor to determine if a provisional categorization is a good one.

Another feature is metadata generation. The idea is that the software categorizes the document and then searches for keywords that are related to the category. This can be useful even if the suggested metadata isn't simply taken, since authors or editors work better selecting from an existing set of keywords than when starting fresh with a blank field.

A closely related feature that some companies offer is noun phrase extraction or as one company, Inxight, calls it, a thing finder. This list of noun phrases can be used to characterize the document or can be combined with the noun phrases of a collection of documents to generate a catalog of entities covered by the collection. One example

might be to generate a list of company names and then use that list to scan other or new documents to determine which documents deal with which particular companies.

One company used a thing finder to uncover all the job postings and resumes on sites all over the Internet and aggregated them into a service bringing together job and employee seekers. They could do this because job postings and resumes have certain terms that can be used to identify them and then it was possible to match people and companies, industry and skills, and so on.

Actual Uses

Auto-categorization software had its start in the news and content provider arena and it is there that it still finds it's most successful and developed application. The reason is clear, it is an environment in which you have 1,000's or tens of thousands of documents a day to categorize and one very clear advantage that this software has is speed. There are many other reasons this area is successful: The material is written by professionals who know how to write good titles and opening paragraphs that the software can use to categorize the material. In addition, the level of categorization tends to be relatively shallow and broad which makes it easier for the software to find a good match.

Another feature of this market is that there are already a number of editors who not only know the subject matter and vocabulary, but also have had experience in categorizing similar content.

A related market is in sites that categorize web sites on the Internet into a browse taxonomy. Yes, Yahoo started with all human editors and no you can't eliminate editors, but you can create a system that makes editors more productive by supporting, not replacing them.

A new and intriguing market is in the intelligence industry. They, like the publishing industry, have huge volumes of content to categorize. However there are two features of the intelligence industry that are different; they need a finer granularity of categorization and not coincidentally, the material is not designated for a community of readers but is routed to one or a few experts.

Not only does the intelligence industry need more specific categories, they also need to categorize content, not just at the document level, but at the paragraph level. This requires a level of sophistication beyond the early simple Bayesian statistics.

Basically, I would say that whatever you use categorization for now can be improved, enhanced, and made more economical by the addition of auto-categorization software, although the highest value areas are still where there is a large influx of documents, preferably well written by professionals, that need to be categorized into a fairly shallow or general taxonomy, or else have very highly developed and specialized vocabularies like the pharmaceutical industry.

I'd like to end this section by looking at a special and difficult area, but potentially very rewarding, the corporate intranet. It is difficult because all the things that make news feeds work very well when pushed through the auto-categorizer are missing on almost all corporate intranets.

The content is written by a really wild mix of writers, some good, some bad, some so bad they're scary. Some of the content is pure literature which is unfortunately sitting next to an accounting document which is next to a scientific research paper. Some of the content has good titles and some has very bad titles and some has every page on the site with the same title. Some of the documents are a single page of HTML and some are book length PDF documents and some have about a paragraph of content but links to all sorts of other pages or sites.

In addition, the economics are wrong. Most corporate intranets, no matter how big, don't have thousands of new pages being published every day, with one exception, which is when news feeds are being posted, but then they have their own categorization that need to be integrated with the categorization schema or taxonomy of the rest of the intranet.

Nevertheless, I think that ultimately, it is the corporate intranet that will see the most lucrative employment of auto-categorization software. Certainly the need is great although different and there are a very large number of corporate intranets which makes the challenge worthwhile. On Intranets one challenge is to create a taxonomy rather than to categorize 1,000's of documents. Another challenge is to create either a very broad taxonomy or else integrate a number of taxonomies. Finally, there is a need for both general categorization to support browsing and a very deep, specific taxonomy that supports quickly finding a particular document or even a paragraph.

One reason for auto-categorization on intranet is the sheer amount of unstructured but very valuable content that resides on corporate intranets. However, because of the factors noted above, it will need a different balance of automatic and manual categorization and it will call for better auto-categorization than has been adequate for news feeds.

Another reason is the current drive to "portalize" intranets. Portal software is itself an attempt to solve the infoglut problem but without a good taxonomic foundation, portals too often end up as very expensive replacements for bookmarks.

<u>What Can't it do?</u>

First and foremost, it cannot completely replace a librarian or information architect although it can make them more productive, save them time, and produce a better end product. The software itself without some human rules based categorization can't currently achieve more than about 90% accuracy – which sounds like a lot until you realize that 1 our of every ten documents listed in a search or browse interface will be wrong. And more importantly, it will be wrong in inexplicable ways, ways which will cause users to lose confidence in the system.

While it is much faster than a human categorizer and doesn't require vacation days and medical benefits, it is not as good as a human categorizer. It can't understand meaning like a human can, and it can't summarize like a human, because it doesn't understand the meaning in the document and because it doesn't bring the meaningful contexts that humans bring to the task of categorization.

One thing that early AI efforts taught us is that while speed is significant, speed alone cannot make up for a lack of understanding of meaning.

How should information professionals approach it?

It is still very difficult to accurately evaluate this type of software or even know what to look for, what is important, and what is hype. The answers will vary from situation to situation, but there are a few things to keep it mind.

First, be very wary of the results of bake offs proving that one product is more accurate. I've seen results that show that Mohomine beats Inxight, but Inxight beats Quiver, but Quiver beats Mohomine, and then in a last minute come form behind victory, Inxight beats Mohomine and Mohomine beats everyone. In other words, something is fishy. And what it is, is that there is no clear method of comparing results. Too much depends on the specific content that is chosen as the test material, the editors or information architects that administer the test, and perhaps the time of day or phases of the moon.

Also, be very wary of vendors that tell you that their software really is automatic and you can just open the box, install the software, point it at your content and out pops your corporate yahoo ready to solve your information needs. It is true that the software is getting better and that with the addition of features like Applied Semantic's Ontology or H5Technologies Subject Matter Framework, the software can do a pretty good job of assigning documents to a general category without any human intervention, nevertheless, even with these world knowledge starting points, there is still a lot of work for information professionals to do.

In fact, the early success stories show, it is in conjunction with information professionals that you find the most successful implementations. Working well with humans then becomes an important area to investigate in your evaluation. One feature to look for is being able to customize the software, tweaking the algorithms in ways beyond simply selecting training sets.

Another feature is how well the software supports work flow for not only editors but for non-editors also. One model that I believe is by far the best for most if not all corporate intranet environments is a distributed work flow model that supports subject matter experts and authors with provisional categorization and metadata and then routes their work to a central team of editors or information architects who with the aid of features like auto-summarization can quickly say "good job" or "I don't think so" (or perhaps the more politically prudent, "May I respectfully suggest that your document might better

belong in the HR vacation policy category not in the large green things that grow category.").

Is it dangerous?

Oh I hope so! Very dangerous! And getting more dangerous as it develops and we find more interesting uses for it. But not necessarily in the ways you might think.

First, it is not particularly dangerous in terms of job security for information professionals, unless someone makes a big mistake. The danger is that someone with a background more on the computer side of information science deciding that this new software really is automatic and that means they can get rid of those pesky librarians. This is a possibility, but the good news is that it will tend to be a self-correcting mistake. When users start to howl about their automatically categorized content as loudly as they do now about search, the software is all we need crowd should get the message.

However, there is another danger which librarians need to be aware of. It's being aware of the difference between a reference library and an information system designed for normal users. Not that I would imply that librarians aren't normal, but well, you know. Anyway, categorization software can be used to create dense, marvelously complex taxonomies that store every document in its place and only its place. These taxonomies can be a thing of beauty - to a taxonomist or librarian, but they can make it very difficult for users to find anything.

Rather than a danger to information professionals, auto-categorization can not only enhance their ability to solve user's information problems, it may even elevate their status to something close to what it should be. Not only will librarians and information architects produce more and more economically, but they will have expensive software associated with the task and, as we all know, in today's corporations unless there is expensive software involved, no one will think you're valuable.

Well, OK, maybe that's a bit overstated, but speaking of the place of the information professional in today's corporations, auto-categorization software has the potential to highlight what should already be clear, that the information professional is engaged in a fundamental infrastructure activity. Information professionals are or should be involved in the creation and maintenance of the intellectual infrastructure of their organization. While technology and organizational infrastructures have received more attention and resources, some of the imbalance could be righted through the intelligent utilization and integration of new software, new methods of working with both content providers and content consumers, and new ways of presenting information.

One of the most central pieces of this intellectual infrastructure is categorization. And like all infrastructure activities, integration with other components is essential. Categorization needs to be incorporated into content creation through content management tools and at the same time incorporated into content consumption through search and portal and collaboration software. There are, of course, many non-software

components that need to be integrated as well, but that takes us outside the scope of the current article.

So, in conclusion, I think it's pretty clear that auto-categorization will ultimately enhance both the power and the prestige of the information professional. Until it gets so good, so intelligent, so insightful, that it takes over completely. Of course, by then, we will have either merged humanity with machines and created a new cyborg race or all humanity will have retired to a life of the meaningless pursuit of idle pleasure and machines will have inherited the earth.

I wouldn't worry about it, yet.

Names and URL's of companies interviewed and/or mentioned for the article:

Applied Semantics	http://www.appliedsemantics.com
Autonomy	http://www.autonomy.com
H5Technologies	http://www.h5technologies.com
Inxight	http://www.inxight.com
Mohomine	http://www.mohomine.com
Quiver	http://www.quiver.com
TopicalNet	http://www.topicalnet.com
Verity	http://www.verity.com