# Relevance and the Future of Search

Stephen E. Arnold
ArnoldIT, Harrods Creek, Kentucky
sa@arnoldit.com

*Drafted August 14, 2005. This essay forms the basis of Stephen E. Arnold's keynote at Information Today's search conference in London, England, on October 11, 2005.* *Postal Box 320 Harrods Creek, Kentucky 40027. Some of this information appeared in an article written for* Online Magazine *earlier in 2005 and some in my monograph* The Google Legacy: How Google's Internet Search Is Transforming Application Software. Information *about the book is located at* http://www.infonortics.com/publications/google/google-legacy.html).

A popular phrase in marketing circles is *old school*. For example, there are old school motorcycles, and there are old school engineering designs or retro-technology like kitchen toasters that look as if they popped from the Victoria & Albert collection of everyday items.

Old school connotes products or technology that is solid and dependable. Old school is intended to make the customer smile. The overriding sentiment is trust.

Is search old school. Does search make a user smile. Is search dependable? Does search make a user smile? Does search engender trust?

In my article for *Online Magazine*, I wrote:

> Taking a broad view, one must accept that relevance of Web search results can be compromised....The biggest threat, however, is not unscrupulous Web page authors. The risk is that the thirst for revenue may become so great that the only relevant results for a query may be results from paid listings. The future, then, may be the Web as a giant Yellow Pages with the content-rich days left behind like piles of old back up floppies.

The old school approach to relevance was people based. The craft of indexing allowed researchers to have a reasonable amount of confidence that an article or book matched the index terms assigned to it.

The old school process required individuals with a fondness for information and specific training to read documents, consider the content of the document and, in some cases, the author's intent, and then apply index terms from a controlled list. Examples of indices that made use of this approach ranged from ABI/INFORM (a business and management database) to Compendex (an index of engineering information) to PROMT (an index of trade and industry data).

The approach today is new school. The cost of skilled humans has become an issue at many information companies. The solution to expensive people and their time-consuming manual processes is automation.

Indexing automation is the only way to cope with the amount of information available today. The statistics from commercial companies such as FileNet and universities such as the University of California - Berkeley differ very little. Information in digital form is growing by leaps and

bounds. An organization generates twice the data it did the previous year. The Berkeley data suggest that the rate is closer to triple the data in the previous year. The reality is that no one knows what the rate is. Everyone knows that the amount of information available is large. It grows quickly. Nothing short of pulling the plug on every computer in the world will slow data growth.

In an article for CMSWatch.com, I identified the last three major trends in enterprise search. These were the use of Extensible Markup Language, automatic indexing, and automatic classification of textual information. To these we can add text mining. Each of these technologies generate a lot of hype but little clarity about how they actually work.

XML is a buzzword for structured information. XML is also the pot of gold at the end of the semantic Web's rainbow. The idea is that a software script will match specific chunks of information from an XML document - in other words, retrieve pieces of documents rather than entire documents. For the user, the idea of getting the most relevant portion of a document suggests that search systems will answer questions, not just generate a long list of results to wade through.

Automatic indexing suggests that software can "read" a document and identify the key terms, phrases, and concepts in that document. These systems have demonstrated that about two-thirds of their output is usable. To tackle the remaining one-third, humans are needed. Quietly, Autonomy Ltd., Stratify Inc., and dozens of other organizations have created administrative screens so human editors can "tune" the indexing. Automatic really means semi-automatic or sort of automatic. The goal of eliminating human indexing has, somewhat happily to my way of thinking, been slowed by complaints about bad indexing.

Automatic classification is a close cousin of automatic indexing. Companies such as ClearForest and Inxight (a Xerox PARC spin out) offer solutions whose price can hit six figures. For the budget minded, there are open source systems such as and quasi-commercial software such as Rubryx. Kofax, the vendor of high-end scanning systems, purchased an automatic classification vendor (Mohomine) and now includes that feature in its company's scanning systems. A fertile area for Ph.Ds in computer and information science, automatic classification seeks to "read" documents and assign each to a conceptual pigeon-hole. Google practices the black art of on-the-fly classification using 500,000 categories that are not hierarchical.

Text mining is a close cousin of data mining. The idea behind mining is that software can pore through documents or entire collections of content in a variety of file formats. As the software processes the documents, the software notes occurrences, variations, and other types of data anomalies. The user looks at a report that provides a bird's-eye view of the information in the corpus. What's brilliant about the idea of text mining is that it combines automatic indexing and automatic classification technologies, tosses in the notion of clustering, and adds a dollop of catnip—no one has to read the information to know what's in the collection of information.

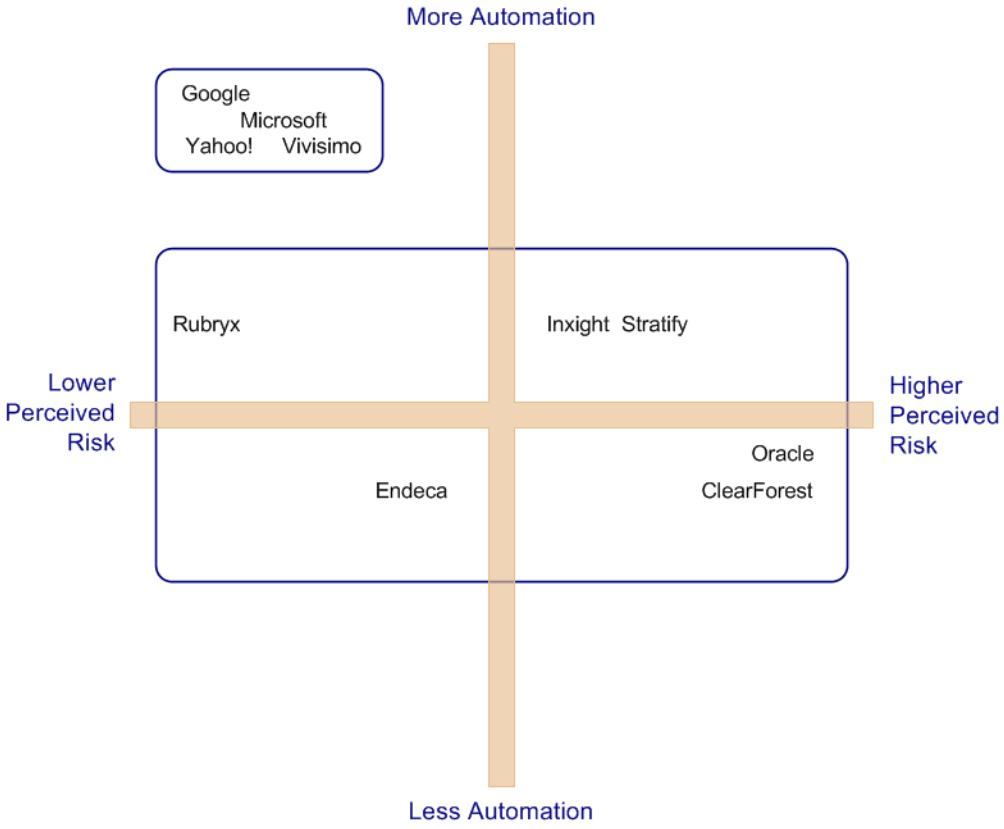
To summarize, the new school of indexing focuses on shifting as much work as possible to automated systems. The volume of information available for indexing makes automation a necessity. The newer approaches give the user some benefits:

- Access to content by topics or categories
- Visual displays of topics or content coverage in the form of numbers or graphic presentations
- Freedom from having to read documents to find out what's in them.

There are some exceptions to these benefits. A good example comes from law. When a large company is involved in high stakes litigation, highly-skilled, presumably well-paid professionals read, and annotate documents. Litigation support systems provide some of the benefits of the type of automatic text processing mentioned above; however, lawyers and expert witnesses manually process text. Machine-processes are not yet good enough when the client may lose large sums of money or face jail time.

The graphic below shows the unusual situation in which information sensitive people find themselves. The market for automatic processing of content is fairly concentrated. In fact, most of the activity is focused on the low-risk, higher-automation sector and what might be characterized as a commercial market including organizations experimenting with automated processes and organizations investing in more specialized processes. The diagram includes only organizations active in text mining. The diagram excludes more specialized types of automated data mining such as processing gene sequences or monitoring credit transaction for anomalous behaviors.

### Relevance and Risk – Automation Diagram




This diagram shows two “spaces” where relevance has importance. The smaller rectangle represents the pursuit of relevance to balance the needs of users of organic search (unpaid listings) versus the needs of advertisers and their inorganic listings (various types of paid listings). The larger rectangle shows the broader spectrum of relevance in organizations. In this sector, when search results pose little risk to the organization lower cost, more labor intensive systems are desirable. When the stakes are higher, higher cost systems are selected. These systems may also be labor intensive; however, the licensee wants to invest in software, systems, and people in order to reduce risk.

This type of market distribution presents several challenges for those concerned with relevance. Let's look at several of the potentially more interesting hurdles from two points of view: the user of the service and the company deploying the automated process.

Let's look at the application of automated text processing for the advertiser-supported indexing services, where the user gets a useful service for free. The problem that becomes immediately evident is that automated processes can be tuned to meet the needs of the advertising portion of the business. Clearly, relevancy is important in the list of Web pages matching a user's query. Matching the appropriate ad to the user's query is important.

However, if the list of "hits"—sometimes called organic search results—is highly relevant so that the "hits" answer the user's question with a high degree of accuracy, the desire of the user to click on advertisements may be lowered. Based on the research conducted for The Google Legacy, users of Google click on advertisements when the search results don't provide the specific information required or when the prospect of looking through a laundry list is too daunting. The ad provides a short cut to the needed information. What this suggested is that Google, Microsoft, and Yahoo among others want to do a good job with organic search results but not the absolute best possible job of matching the user's query. Similarly, these types of companies want to do the best possible job of matching advertisements to the user's query. It doesn't take much experience with text mining systems to see that the organization using the text mining system can adjust the results of the algorithms to strike a balance between organic search results' relevancy and the relevancy of the ads in order to keep users and advertisers happy—that is, coming back to the search system and buying ads to reach those users. Tip the balance too far for the user and the ad revenue may tend downward. Tip the balance too far to the advertiser, then the user may become annoyed that a non-advertiser "hit" in the result list did not provide useful information.

Another issue is that if a Web developer can spoof the organic results in some way, the user may see an irrelevant citation. A sufficiently annoyed user may click to another search engine, thus triggering a downturn in the number of clicks for the advertisers. Because Web indexing systems can be fooled by clever developers, relevancy is a challenge to the Googles, Microsofts, and Yahoos.

The issue of relevancy is equally important in the band of vendors running parallel to the x axis of the diagram. These companies provide automation tools to organizations wanting to bring some order to their burgeoning content. These vendors address what Donald Rumsfeld described as "not knowing what we know."

The issue that surfaces is that relevance comes at a price. For those applications where the risks are high, there is minimal automation. The best example occurs in litigation. The sophisticated litigation and document management support systems in law firms are never replacements for attorneys who sift and read documents. Similarly in such life or death applications as those posed by extremists, automatic systems cannot replace "feet on the street" or specialists analyzing documents and message traffic.

Automation in support of relevancy has its limits. Unlike the "tuning" that seems to be part of the ad-supported indexing systems' balancing act, commercial applications of technology are gated by cost. The technology comes at different price points from the less than $100 product from Rubryx to the seven figure installations from ClearForest, Stratify, and other specialist vendors.

To summarize, relevance is likely to be compromised (intentionally or unintentionally) by "free" Internet search systems. These companies have to walk a fine line between driving traffic to Web sites in the results list or to the landing pages of the advertisers. Get out of kilter either way and the free indexing service has a multitude of problems.

The commercial issue of relevance boils down to what the organization is willing to pay. Lower-cost systems may not do a very good job of indexing or adding value to the tags included in the index. Without the resources to add humans to the mix, the likelihood of getting irrelevant "hits" seems a foregone conclusion. For organizations with the resources to buy sophisticated systems and hire professionals to refine and tune the systems, relevance may be higher. However, when the stakes are high, organizations are not likely to put their trust in automated systems.

## Two Problems, Many Solutions

Relevance, like beauty, is in the eye of the beholder. Users of online search-and-retrieval systems have high expectations. Part of the reason is that more people have more experience using the Internet's various finding tools. For the most part, these tools work well. When Google applied its voting approach to Web search results, users helped make Google one of the premier search systems. Even enterprise search systems were influenced. Corporate procurement teams added a new question to their repertoire, "How is your system like Google's?" reminded vendors of enterprise search systems of Google's growing influence.

### Web Search

With Google results, users found that most queries were met with a list of highly relevant results. For Google-anado, there was the "feeling lucky" button. Click the link, and Google displayed the result most relevant to the user's query. For queries for hotel accommodations or cheese vendors, Google's "feeling lucky" function was spot on most of the time.

Google's approach to relevance is far richer than voting. Since 1998, Google has filed hundreds of patents and been awarded more than 100 patents for search-related activities. A significant number of Google patents focus on improving the relevance of organic and inorganic results. Inorganic results are ads or other types of for-fee listings.

Among the processes Google uses are:

- Query expansion. Google looks at a query and determines what other sites are likely to have relevance to that particular query.
- Clustering. Google groups sites, concepts, and many other factors. By looking at other items in a cluster, Google can provide suggestions as evidenced by its Google Suggest function.
- Semantic and linguistic processes. Since 1996, Google has not pounded its semantic and linguistic drums. However, patents filed as early as 2001 point to a growing reliance on routines that combine statistical and linguistic processes. These processes are applied to organic and inorganic results. For participants in AdSense, these processes identify other Web pages where a particular ad is likely to be germane. For AdWords, these processes help expand the Web pages on which content and ads have an appropriate relationship.

Google is a work in progress. These techniques, while advanced, are becoming part of other Web search engines' processes. There is good news and bad news with this type of increasing sophistication from the blending of behavioral data (voting or popularity determined by clicks or links) and layers of statistical, semantic, and linguistic processes. The good news is that for many queries, these processes allow systems to recommend other places to look for information. The term facets is now applied to showing what are "More like this" and "See also" references in old school systems.

The other bit of good news is that Microsoft and Yahoo have little choice but improve their search systems for users and for advertisers. Without a strong counter to Google, both companies face the prospect of ceding an important source of revenue to Google.

The bad news is that as the competitive heat rises, the temptation to fiddle with the controls increases as well. Egregious changes are not desirable. However, fine adjustments are likely to become normative. Google's algorithm for relevancy is a bouncing Betty chunk of mathematics. This type of algorithm changes as new values become known. When one factor reaches a certain level, a trigger sets off one or more changes in other parts of the algorithm. Figuring out what causes a specific change occupies many 20-somethings in the search engine optimization industry.

The other piece of bad news is that the need for computing horsepower and bandwidth goes up with each additional algorithmic process. Much search and retrieval work requires highly parallelized execution of core processes. More processes demand that the search company invest in faster systems. Unless these costs are repaid in some way, the computations can bring a search system to its knees. Alta Vista and Excite both hit the computational load / cost wall and suffered as a result.

### *In the Organization*

If we shift our attention to relevance within an organization, a somewhat different situation appears to be forming.

Enterprise search is becoming a commodity. IBM, Microsoft, and Oracle have the option of making search part of other enterprise software. For example, Microsoft can include search in the forthcoming Vista operating system. Oracle already includes a serviceable text search system in its database management system. IBM has ginned up another suite of enterprise search technologies and promises to provide them at no or low cost. How can vendors of enterprise search systems compete with companies who are larger and give away search as part of their enterprise software bundle? The answer is, "Do something different?"

The "different" for many vendors of search systems is to focus on providing more functionality and more bells and whistles that the "free" search systems offer. For vendors of enterprise search without the money and resources to develop a better mousetrap, their future is less than cheery.

However, for those with technical insight and access to cash, the future is bright. The outlook for software that adds value to enterprise search systems is bright. The types of functions that are likely to become more widely available in the next 12 to 24 months include:

- Algorithms that interact with content without requiring the customer to invest significantly in infrastructure
- Technology that includes easier to use tools for training systems to have sensitivity to particular words, phrases, and proper nouns used in an organization. Human effort can be better allocated when relevancy-enhancing tags do not require manual intervention.
- Hooks that allow standing queries to be integrated into specific business processes. The manual processes required to map queries to work tasks makes a useful function too expensive for many organizations.
- Pricing that puts these relevancy-enhancing tools within reach of more organizations. The market for value-adding search utilities is skewed to those with deep pockets. Lower price points combined with ease of use and increased relevancy functions will expand the pool of potential buyers.

Entrepreneurs and investors are likely to see this as an opportunity to make money. Users will benefit because search systems will be more likely to return an "answer," not a list of results to be scanned.

The bad news is that these systems have the potential to come back and bite management on the ankle. The notion of learning what information is closely related to a particular deal is useful to a salesperson. However, that same information in the hands of an attorney investigating alleged collusion over pricing may be the evidence to convict.

## Looking Forward

Relevance is not dead. Relevance has been given a pacemaker and a new regimen. The old notions of precision and recall are essentially too "old school" to be of much use in today's information-centric world. The "new school" teaches that relevance is situational. System managers and users need knobs, controls, switches, and settings to tune a search system to deliver what's needed at any given time.

On the Web, the same relevance engines drive the free Web search and the for-fee matching of ad blather to the users' queries. The difference is in how the algorithmic bag of tricks is applied to a particular corpus for a particular purpose. One doesn't want relevance to be left to chance in either situation.

In the enterprise, relevance is subject to gating by the resources of the organization. Organizations with more cash to spend are likely to have an information advantage over organizations who must use less sophisticated systems. However, more sophisticated systems require a careful and considered approach to deciding:

- What to index
- Who can search what content
- When are indexes updated
- How to handle versions of content.

A user may not find a relevant document for the simple reason that access to that document is prohibited. An attorney may not find a document for the simple reason that the documents was not retained to be indexed. A vice president of finance may not find a document because it was on a server excluded from the indexing process. These examples underline the fact that relevance is a slippery fish regardless of technology. Policy comes into play and makes the rules about much of relevance.

Users of free search systems need to be aware of the "new school's" approach to relevance. Precision and recall are less important than balancing various economic and behavioral forces. Users of enterprise search systems need to be aware of limits on relevance imposed by two different systems. One one hand, technology may make enterprise search more useful to workers in task specific jobs. On the other hand, organizational policies may make the issue of relevance irrelevant. Documents germane to a query may not be available for a range of reasons, some benign others less benign.

Finding a document that answers a user's question is easier today than at any other time in the short history of online search. The future problems are no longer just technical. Looking forward, relevance is now faced with challenges that include financial, political, legal, and technical aspects. The quest for relevance shares some features of knights seeking the Holy Grail. Brave people know it's out there. Finding it requires conviction and effort.