Editorial Preface

Search Engine Technology: A Closer Look at Its Future

Mahesh S. Raisinghani, Associate Editor, IRMJ

ABSTRACT

This editorial preface will describe the need for new technology for search engines, the two generations of search engine technologies used, and the need for a third generation search engine. We will discuss the limitations of today’s search engines and difficulties in determining relevancy of the search and the need for a context-sensitive ranking system, and take a closer look at the future of search engines based on the open source model.

INTRODUCTION

The Internet, particularly the World Wide Web, is a vast source of information that is growing at an explosive rate. More than 7 million publicly available pages are added to the World Wide Web everyday, according to a study conducted by Cyveillance, a Washington DC-based market research firm (WiseNut Search Engine White Paper, 2001). This growth rate means that the 2.1 billion unique pages available on the Web will double in the span of about one year. In the early days, search results were very basic and largely depended on what was on the Web page. Important factors included keyword density, title, and where in the document keywords appeared.

The first generation added relevancy for META tags, keywords in the domain name, and a few bonus points for having keywords in the URL. Basic spam filters emerged that eliminated keyword stuffing and same color text. The portals also made their appearance, and engines started looking like giant billboards and overstuffed yellow pages.

The second generation added much in the way of off page criteria and link analysis. Some of the major components employed are tracking clicks, page reputation, link popularity, temporal tracking, and link quality. Other advances include term vectors, statistics analysis, cache data, and context where two-word keyword pairs are extracted from a page to better categorize it.

The third generation of search engines is already underway. It adds word stemming and a thesaurus on top of the term vector database to assist in keeping a search in context. Auto extraction of keyword pairs also helps automatically categorize a page, where searches like ‘shop for’ or ‘find’ trigger totally different search results based on the context or intent of the person doing the searching (Nobles, 2003).

HOW SEARCH ENGINES WORK

The term “search engine” is often used generically to describe both crawler-based search engines and human-powered directories. These two types of search engines gather their listings in radically differ-
ent ways. Crawler-based search engines, such as Google, create their listings automatically. They “crawl” or “spider” the web, then people search through what they have found. If we change our web pages, crawler-based search engines eventually find these changes, and that can affect how we are listed. Page titles, body copy and other elements all play a role.

Crawler-based search engines have three major elements. First is the spider, also called the crawler. The spider visits a web page, reads it, and then follows links to other pages within the site. This is what it means when someone refers to a site being “spidered” or “crawled”. The spider returns to the site on a regular basis, such as every month or two, to look for changes.

The second part of the Crawler-based search engine, is the index. The index, sometimes called the catalog, is like a giant book containing a copy of every web page that the spider finds. If a web page changes, the book is updated with the new information.

Search engine software is the third part of a Crawler-based search engine. This is the program that sifts through the millions of pages recorded in the index to find matches to a search and rank them in order of what it believes is most relevant (How search engines work, 2002).

HUMAN-POWERED DIRECTORIES

A human-powered directory, such as the Open Directory, depends on humans for its listings. When directories have enough websites for a category and it is “full” more categories and subcategories are created. There are thousands of directories on the web today and many of them specialize in particular subject areas. When you submit a web site to a directory it is usually accompanied with a short description of the web site and some keywords that best describe the website’s content. The directory editors then review the site by visiting it on the web and then add it to their category. In some cases they may even create a new category or subcategory. Usually listings appear in alphabetical order but you may also have the possibility to pay a small subscription fee to be listed at the top (How search engines work, 2002).

Changing the web page has no effect on its listing. Things that are useful for improving a listing with a search engine have nothing to do with improving a listing in a directory. The only exception is that a good site, with good content, might be more likely to get reviewed for free than a poor site. Usually, a hybrid search engine will favor one type of listings over another. For example, MSN Search is more likely to present human-powered listings from LookSmart. However, it does also present crawler-based results (as provided by Inktomi), especially for more obscure queries.

Major search engines generally provide listings from a variety of sources, which they may get from third-party search providers or through their own efforts. Table 1 illustrates where each search engine gets the main results it displays.

LIMITATIONS OF TODAY’S SEARCH ENGINES

The Web has experienced continuous growth since its creation. As of March 2002, the largest search engine contained approximately 968 million indexed pages in its database. As the number of available Web pages grows, users experience increasing difficulty in finding documents relevant to their interests. One of the underlying reasons for this is that most search engines find matches based on keywords, regardless of their meanings. Other main reasons for obtaining poor search results are that many words have multiple meanings. For instance, two people
searching for “wildcats” may be looking for two completely different things (wild animals and sports teams), yet they will get exactly the same results. To provide the user with more useful information, we need a system that includes information about the conceptual frame of the queries as well as its keywords. This is the goal of KeyConcept, a search engine that retrieves documents based on a combination of keyword and conceptual matching. The documents are automatically classified to determine the concepts to which they belong. The Query concepts are determined automatically from a small description of the query or explicitly entered by the user. KeyConcept is shown to significantly improve search result precision through its use of conceptual retrieval.

### DIFFICULTIES IN DETERMINING RELEVANCY

The Collection frequency weighting does not, however, rank pages with the greatest relevancy at the top of the list. Nor does it assess the importance of a given page relative to other pages. Moreover, basing the rank of a Web page solely on the content of the page itself and in particular the content of the Meta tag, which does not even appear as part of the text of the page can cause problems for search engines. This is because savvy Web page authors can use a technique known as spamming, repeating a “hot” keyword many times in the title or the Meta tag to raise the rank of the page.

### Table 1. Search engine results chart

<table>
<thead>
<tr>
<th>Search Engine</th>
<th>Type Of Main Results</th>
<th>Provider Of Main Results</th>
<th>Paid Results</th>
<th>Directory &amp;/or Backup Results</th>
</tr>
</thead>
<tbody>
<tr>
<td>AllTheWeb</td>
<td>Crawler</td>
<td>AllTheWeb (Overture-owned)</td>
<td>Overture</td>
<td>n/a</td>
</tr>
<tr>
<td>AltaVista</td>
<td>Crawler</td>
<td>AltaVista (Overture-owned)</td>
<td>Overture</td>
<td>LookSmart</td>
</tr>
<tr>
<td>AOL Search</td>
<td>Crawler</td>
<td>Google</td>
<td>Google</td>
<td>Open Directory</td>
</tr>
<tr>
<td>Ask Jeeves</td>
<td>Crawler</td>
<td>Teoma (Ask-owned)</td>
<td>Google</td>
<td>Open Directory</td>
</tr>
<tr>
<td>Google</td>
<td>Crawler</td>
<td>Google</td>
<td>Google</td>
<td>Open Directory</td>
</tr>
<tr>
<td>HotBot</td>
<td>Crawler</td>
<td>Inktomi (Yahoo-owned)</td>
<td>Overture</td>
<td>Also available: AllTheWeb, Google, Teoma</td>
</tr>
<tr>
<td>LookSmart</td>
<td>Human</td>
<td>LookSmart/Zeal</td>
<td>LookSmart</td>
<td>Backup from Inktomi</td>
</tr>
<tr>
<td>Lycos</td>
<td>Crawler</td>
<td>AllTheWeb (Overture-owned)</td>
<td>Overture</td>
<td>Open Directory</td>
</tr>
<tr>
<td>MSN Search</td>
<td>Human</td>
<td>LookSmart/Zeal</td>
<td>Overture</td>
<td>Backup from Inktomi</td>
</tr>
<tr>
<td>Netscape</td>
<td>Crawler</td>
<td>Google</td>
<td>Google</td>
<td>Open Directory</td>
</tr>
<tr>
<td>Overture</td>
<td>Paid</td>
<td>Overture</td>
<td>Overture</td>
<td>Backup from Inktomi</td>
</tr>
<tr>
<td>Open Directory</td>
<td>Human</td>
<td>Open Directory</td>
<td>n/a</td>
<td>n/a</td>
</tr>
<tr>
<td>Teoma</td>
<td>Crawler</td>
<td>Teoma</td>
<td>Google</td>
<td>n/a</td>
</tr>
<tr>
<td>Yahoo</td>
<td>Crawler</td>
<td>Google</td>
<td>Overture</td>
<td>Yahoo</td>
</tr>
</tbody>
</table>

Source: searchenginewatch.com
without adding any value to the content of that page.

In an attempt to avoid falling victim to spamming, many search engines severely penalize pages that appear to be using this technique. Inevitably, legitimate pages are often unduly penalized. Recently, some search engines have begun to rely on the valuable information buried in the structure of the Web itself to rank pages in a more objective way. Unlike standard paper documents, the Web includes hypertext, which links one page to another. This link structure can supply valuable information about the importance of a given Web page. The inbound links — all the links that point to a particular page — provide a relatively accurate assessment of the value of that destination page. Because many of these inbound links originate from authors other than the one who created the page being considered, they tend to give a more objective measure of the quality or importance of that page. By making a link to another page, the author of the originating page sanctions the destination page in some way. So theoretically, the more inbound links a page has, the more likely it is to be important. Simple counting of inbound links, however, will not tell the whole story. If a page has only one inbound link, but that link comes from a highly valued page such as the Yahoo! homepage, that page might reasonably be ranked higher than a page that has several inbound links coming from less significant pages. Under this link analysis system, the rank of a given page will improve if it contains lots of links from other pages, especially if they come from highly valued pages. Many experts attribute the early success of Google to its PageRank™ method, which uses this type of link analysis ranking system.

NEED FOR A CONTEXT-SENSITIVE RANKING SYSTEM

By design, the PageRank™ method only looks at the connectivity graph of the link structures, ignoring the context of these links. PageRank objectively tells us the relative importance of a page but fails to determine its relevance. Consider a searcher looking for the fundamental theory of quantum physics. The Yahoo! homepage, while one of the most important pages on the Web according to the PageRank™ system, is certainly not the ultimate authority on quantum physics. Further, the PageRank™ system still relies on traditional methods of text analysis to determine relevancy, which as discussed above, may eliminate relevant pages by erroneously penalizing them as a defense against spamming. So it is better to rely on a link analysis-based ranking system that is context-sensitive. The WiseNut’s patent-pending WiseRank™ does just that.

The comScore Media Metrix qSearch service measures search-specific traffic on the Internet. Search data is gathered by monitoring the web activities of 1.5 million English-speakers worldwide (1 million in the United States) via proxy metering. Proxy metering allows comScore to see exactly how those within its panel have surfed the web. From this data, the company then extracts activity that’s considered to be specifically search-related.

FUTURE OF SEARCH ENGINE TECHNOLOGY

There are several reasons why the current system is not working or does not work effectively. On one hand, the Internet is growing at an exponential rate. On the other hand, spammers are growing at an exponential rate as well. In essence, the engines need a way to store more pages,
combat spam, and still provide pertinent results.

The third generation of search engine technology adds Web maps that although not searchable, are a useful filtering tool to get rid of duplicate sites and many stand alone pages that drive traffic to only a few destinations. This means pages like doorways, gateways, entry, and/or splash will soon get filtered out. They will also be extracting as much data as possible about our individual searching habits. All the major engines plan on building personal profiles, little robots that ‘come to know you’ over a period of time, based on past searching habits.

Search is a game of intellectual property, innovation, and market position. Indeed, search is such a complicated and expensive undertaking that analysts have pegged the cost of market entry at well over $100 million. Commercial search-engine providers soon could face a serious competitor if the vision of some open-source developers materializes. A team of open-source programmers recently launched a project called Nutch to provide search-engine software for free (Battelle, 2003).

If and when the Nutch project begins to build momentum, perhaps a handful of commercial niche engines will be created using Nutch technology as their foundation. For example, commercial search engines could tweak Nutch’s algorithm enough to prevent spammers from polluting the results and might have access to enough capital to become formidable players in the commercial search-engine business (Hook, 2003). Interestingly, early Nutch development was supported in part by Overture’s R&D division, and an Overture official serves on the Nutch board (Battelle, 2003).

**LATEST DEVELOPMENTS IN SEARCH ENGINES**

Google has released Channels an enhanced reporting featured for its Adsense program that allows publishers to track multiple sites, or groups of pages to gauge their performance individually. Channels provide us with a way to view detailed reporting about the performance of our pages. By pasting channel-specific ad code into any combination of pages, company can track a variety of metrics across their sites. Companies can Track their desktop performance versus the performance of their laptops, or compare our motorcycle pages to our automobile pages, by assigning each group of pages to a specific channel and comparing results in our custom channel reports. Companies can even assign a channel to each of their separate domains, to see where clicks are coming from. Other new features include downloadable CSV reports, support for ads within frames, and alternate ad colors for Public Service Announcement (PSA) (Lozano, 2004).

At Search Engine Strategies, all the spider-based search engines confirmed their commitment to a comprehensive search index containing as much relevant content as their spiders can find.

Google confirmed its stance against paid inclusion, saying its spiders should find all relevant content. Teoma/Ask Jeeves backed away from XML paid inclusion. Its communities-focused algorithm had problems accurately targeting data structured in an XML format. The company retains a per-URL paid inclusion program to facilitate more rapid inclusion of URLs and content into the index with content collected by the spider.

Yahoo! has announced a new paid inclusion program, Site Match, part of its content acquisition program (CAP). Site Match is a shift from the earlier Inktomi, AltaVisa, and FAST inclusion programs. It introduces CPC to all commercial inclusion (other than the Yahoo directory). Yahoo’s new paid inclusion philosophy eliminates the free ride for clicks of past per-URL inclusion programs. Yahoo created a similar pricing
model for XML and per-URL inclusion programs. A lack of per-click fees for traffic meant low incentive for Webmasters or Web marketers to keep page content relevant for earlier per-URL paid inclusion programs at Yahoo divisions (Inktomi, AltaVista, and FAST) (Lee, 2004).

Kanoodle has joined the competition between Overture and Google for distributed search marketing dollars, launching a content-targeted sponsored links program. The new program, ContextTarget, is currently the only available sponsored links product built solely for content targeting, rather than being driven by keyword matching. Contextual ads appear on web pages based on the context or topic of the page. For example, if we are visiting a web page that discusses clinical depression, the contextual ads displayed on that particular web page would be related to mental illness and depression. Thus, we might see ads from pharmaceutical companies, medical information sites, and treatment centers for mental illness.

What makes Kanoodle’s ContextTarget program unique is that advertisers can choose where they want their ads to appear by selecting specific categories. Currently, Overture and AdSense do not allow that level of control.” Rather than relying on a dynamic page scan of content pages to determine the appropriate ad based on a key word, ContextTarget maps the content page to a taxonomy of broader subject categories within content sites (i.e., publishers), so that ads appear by subject,” said Jillmarie J. Giardina, Kanoodle’s Director of Marketing.” An advertiser chooses one or more categories and enters a bid for placement on pages under each,” she continued. “This ensures only relevant ads appear on content pages.” Additionally, partners can block competitors’ ads from appearing on their site. The partner company can block by URL and by category (Thurow, 2004).

Kanoodle ContextTarget program allows advertisers to:

- Bid distinctly on their contextual ads.
- Pause ContextTarget listings independently of keyword listings.
- Write ads specific to the contextual environment.
- Schedule ContextTarget listings to automatically go live or dead at specific times, with no need for manual adjustment.

eBay Stores URLs will include dashes between the words so the individual words will be considered by search engines when determining the relevance of Stores for a given search. eBay is changing the default structure of its Stores URLs to make it easier for search engines to recognize keywords in Store names more easily and provide better indexing. The auction site has worked to make it easier for search engines to read the content of Stores, and all Store URLs that members previously used will continue to work and are being redirected to the new page URLs (AuctionBytes & Search Engine Lowdown, 2004).

CONCLUSION

Researchers are poised to revolutionize search technology over the next few years. The most common thrust is to personalize search engines. Adele Howe, a computer science professor at Colorado State University in Fort Collins, and Gabriel Somlo, a CSU graduate student, have built a proof of concept called QueryTracker, a software agent that sits between a user and a conventional search engine and looks for information of recurring interest, such as the latest news about a user’s chronic illness. QueryTracker submits a user’s query to the
search engine once a day and returns results from new Web pages and pages that have changed since the previous search. The magic in QueryTracker comes from its automatic generation of an additional daily query—which Howe says is often superior to the user’s original query—based on what it learns about the user’s interests and priorities over time. It filters the results of both queries for relevance and sends them to the user (Anthes, 2004).

Studies repeatedly show that more than 80 percent of Web users rely on search engines to reach their destination. Currently, most search engines fail to capture the bulk of the “invisible web” - resources in databases inaccessible to the engines indexing crawlers. Clearly we have a need for new improved technology in search engines in order to cope with the increased Internet usage around the world. In order to retrieve the search at faster rate, the search engines have to crawl the Web faster using the latest and most powerful technologies and ensure that the results are the most relevant based on the context and the freshest available on the Web.

REFERENCES


