Introduction

The Internet, particularly the World Wide Web, is a vast source of information that is growing at an explosive rate. More than 7,000,000 publicly available pages are added to the World Wide Web every day, according to a study conducted by Cyveillance, a Washington D.C.-based market research firm.

This growth rate means the 2.1 billion unique pages available on the Web will double in the span of about one year. The challenges for today's search technologies are twofold. Search engines must be fast enough to crawl the exploding volume of new Web pages in order to provide the most up-to-date information. As the number of pages on the Web grows, so will the number of results search engines return. So, it will be increasingly important for search engines to present results in a way that makes it quick and easy for users to find exactly the information they're looking for. Search engines have to overcome both these challenges.

Limitations of Today's Search Engines

The Web has experienced exponential growth since its creation. 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 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 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, 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! home page, 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.

The 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! home page, 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 previously, 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 (e.g., WiseNut’s patent-pending WiseRank™).

Effective Utilization of Search Engines

What we need to know is that search engine optimization, which is actually the effective utilization of search engines to draw traffic to a Web site, is an art. It is an ongoing, continuously evolving, high-maintenance process that includes the customization of a site for better search engine ranking. After developing a Web site and selecting the best hosting company, don’t rush out and submit it to search engines immediately. A Web site manager would be wise to take a little time to do the following (Dahm, 2003; Nobles, 2003):

1. Fine-Tune the Title Tag to Increase Traffic to the Site: Improving the title tag is one technique that applies to just
about all search engines. The appearance of key words within the page title is one of the biggest factors determining a Web site’s score in many engines. Changing page titles to include some of the site’s key words can greatly increase the chance that a page will appear with a strong ranking in a query for those key words.

2. Create Gateway Pages that are Specific to the Focus of Each Site: Keyword selection must be done carefully with great forethought and understanding of the search engine’s selection criteria for key words. The larger the number of keywords that are used, the more the relevance of any one key word is diluted. One way to get around this is to create gateway pages.

3. Ensuring that Site Technology Won’t Confuse the Search Engines: Often, the latest technology being built into a site can confuse the search engine spiders. Frames, CGI scripts, image maps, and dynamically generated pages are all technologies that many spiders don’t know how to read. With frames, for instance, the syntax of the FRAMESET tag fundamentally changes the structure of an HTML document. This can cause problems for search engines and browsers that don’t understand the tag. Some browsers can’t find the body of the page, and viewing a page through these browsers can create a blank page.

4. Setting Up a NOFRAMES Section on the Page: Every page that uses frames should include a NOFRAMES section on the page. This tag will not affect the way a page looks, but it will help a page get listed with the major search engines.

5. Performing a Maintenance Check: All Web sites should be thoroughly tested using a site-maintenance tool in order to catch errors in operation before customers are brought to the site. HTML errors can hinder a search engine spider’s ability to index a site; it also can keep a search engine from reading a page or cause it to be viewed in a manner different from how it was intended. In fact, a recent report by Jupiter Communications suggested 46% of users have left a preferred Web site because of a site-related problem. With NetMechanic’s HTML Toolbox or another site maintenance tool, all Webmasters, from the novice to the expert, can avoid potential visitor disasters due to site errors.

6. Finding the Best Submission Service: Selecting a search engine submission service requires careful thought and important decisions. Using an auto-submission service is a good place to begin. Most search engines, like Alta Vista, HotBot, and InfoSeek, automatically spider a site, index it, and hopefully add it to their search database without any human involvement. Some engines, like Yahoo, are done completely with human review and for many reasons are best submitted individually.

To improve site rankings and increase understanding of the listing process, there are many tasks that can be done on a regular or semi-regular basis. Optimizing rankings within the search engines also helps to ensure that a site attracts the right traffic. Some of the monthly and weekly tasks are:

Crunching and Examining Log Files: Data contained in log files are an excellent resource for identifying which engines are sending the majority of traffic to a site. It also can show which keywords or gateway pages are generating the strongest traffic and what visitors are doing when they enter the site.
Searching the Search Engines: Conduct a search of the search engines to analyze where the highest rankings of the site have materialized and what keywords are generating the best rankings. Different search engines use different rules to rank pages. Individual gateway pages should be created based on the knowledge and interpretation of what each search engine is using to determine top rankings. Several pages can be tested on one or more engines, and the pages that have the most success can be kept, while the unsuccessful pages can be dumped or revised to achieve a higher ranking.

Future Trends in Search Engines

A group of international developers is working to produce a Nutch search engine that may rival the commercial search bands. Nutch is a nascent effort to implement an open-source Web search engine. Web search is a basic requirement for Internet navigation, yet the number of Web search engines is decreasing. Today’s oligopoly could soon be a monopoly, with a single company controlling nearly all Web searches for its commercial gain. That would not be good for users of the Internet.

Nutch provides a transparent alternative to commercial Web search engines. Only open source search results can be trusted completely to be without bias (or, at least, their bias is public). All existing major search engines have proprietary ranking formulas and will not explain why a given page ranks as it does. Additionally, some search engines determine which sites to index based on payments rather than on the merits of the sites themselves. Nutch, on the other hand, has nothing to hide and no motive to bias its results or its crawler in any way other than to try to give each user the best results possible.

Nutch aims to enable anyone to easily and cost-effectively deploy a world-class Web search engine. This is a substantial challenge. To succeed, Nutch software must be able to:

- fetch several billion pages per month
- maintain an index of these pages
- search that index up to 1,000 times per second
- provide very high-quality search results
- operate at minimal cost

(Nutch, 2003; Loren, 2003)

Nutch has not yet been tuned for quality. There are 10 or 20 knobs that we can twiddle to adjust the ranking formula. They have started developing software to do this tuning automatically, but the current code just contains guesses. With a little tuning, Nutch should be able to get results that are competitive with major search engines.

Google has released Channels, an enhanced reporting feature for its Adsense program, that allows publishers to track multiple sites or groups of pages in order to gauge their performance individually. Channels provides 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, companies can track a variety of metrics across their sites. Companies can track their desktop performance vs. the performance of their laptops, or compare motorcycle pages to automobile pages by assigning each group of pages to a specific channel and comparing results in the custom channel reports. Companies can even assign a channel to each of their separate domains in order 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 announcements (PSAs) (Lozano, 2004).

At search engine strategies, all the spider-based search engines confirmed their commitment to a comprehensive search in-
dex 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!, meanwhile, announced a completely new paid inclusion program, Site Match, part of its content acquisition program (CAP). Site Match is a shift from the earlier Inktomi, AltaVista, 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. “While they are currently sold through the same channels, contextual and search advertising present advertisers with an inherently different value proposition,” said Josh Stylman, Managing Partner at Reprise Media. “To optimize the effectiveness of contextual campaigns, marketers must possess the ability to control bids, offer separate messaging and have simple turn-key tracking that is partitioned off from their search campaigns.”

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. “Kanoodle.com also allows publishers to determine their own categories,” Giardina further explained. “For example, if AOL wants to block other ISPs or another publisher wants to block Financial News Providers, the company can effectively block these categories. All of Kanoodle.com’s ads are checked for relevancy and for inclusion in a forbidden 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 that 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 result will be that buyers will more easily find our Store when they are searching the Web for items we are selling, eBay’s announcement read. eBay said an important way for Stores to gain visibility and traffic is through appearing in search engines off eBay. The auction site has worked to make it easier for search engines to read the content of Stores, and said the URL changes is part of that effort. eBay said all Store URLs that members previously used will continue to work and are being redirected to the new page URLs (AuctionBytes & Search EngineLowdown, 2004).

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% of Web users rely on search engines to reach their destination. Currently, all search engines fail to capture the bulk of the invisible Web—resources in databases inaccessible to the engines indexing crawlers. The above research clearly indicate the need for new improved technology in the search engines in order to cope with the increased Internet usage around the world and to retrieve the search at a faster rate; using the latest and most powerful technologies, the search engines have to crawl the Web faster and ensure results are the freshest and best available on the Web.

References


