Category: Concepts, Trends, and Challenges

Internet Search Engines

Vijay Kasi
Georgia State University, USA

Radhika Jain
Georgia State University, USA

INTRODUCTION

In the context of the Internet, a search engine can be defined as a software program designed to help one access information, documents, and other content on the World Wide Web. The adoption and growth of the Internet in the last decade has been unprecedented. The World Wide Web has always been applauded for its simplicity and ease of use. This is evident looking at the extent of the knowledge one requires to build a Web page. The flexible nature of the Internet has enabled the rapid growth and adoption of it, making it hard to search for relevant information on the Web. The number of Web pages has been increasing at an astronomical pace, from around 2 million registered domains in 1995 to 233 million registered domains in 2004 (Consortium, 2004). The Internet, considered a distributed database of information, has the CRUD (create, retrieve, update, and delete) rule applied to it. While the Internet has been effective at creating, updating, and deleting content, it has considerably lacked in enabling the retrieval of relevant information. After all, there is no point in having a Web page that has little or no visibility on the Web.

Since the 1990s when the first search program was released, we have come a long way in terms of searching for information. Although we are currently witnessing a tremendous growth in search engine technology, the growth of the Internet has overtaken it, leading to a state in which the existing search engine technology is falling short.

When we apply the metrics of relevance, rigor, efficiency, and effectiveness to the search domain, it becomes very clear that we have progressed on the rigor and efficiency metrics by utilizing abundant computing power to produce faster searches with a lot of information. Rigor and efficiency are evident in the large number of indexed pages by the leading search engines (Barroso, Dean, & Holzle, 2003). However, more research needs to be done to address the relevance and effectiveness metrics. Users typically type in two to three keywords when searching, only to end up with a search result having thousands of Web pages! This has made it increasingly hard to effectively find any useful, relevant information.

Search engines face a number of challenges today requiring them to perform rigorous searches with relevant results efficiently so that they are effective. These challenges include the following ("Search Engines," 2004).

1. The Web is growing at a much faster rate than any present search engine technology can index.
2. Web pages are updated frequently, forcing search engines to revisit them periodically.
3. Dynamically generated Web sites may be slow or difficult to index, or may result in excessive results from a single Web site.
4. Many dynamically generated Web sites are not able to be indexed by search engines.
5. The commercial interests of a search engine can interfere with the order of relevant results the search engine shows.
6. Content that is behind a firewall or that is password protected is not accessible to search engines (such as those found in several digital libraries).[1]
7. Some Web sites have started using tricks such as spamdexing and cloaking to manipulate search en-

Table 1. Search engine issues and challenges

Major Challenges and Concerns	Detail
Spamdexing and cloaking	These tricks are used by Web sites to manipulate search engines to display them as the top results for a set of keywords.
Privacy and security	Search engines index all the content available on the Web without any bounds on the sensitivity of the information.
Information explosion	The Web is growing at a fast rate, with Web pages being updated very frequently, presenting challenges to search engines to index frequently and intensely.
Categorization and representation	Search engines are facing the challenge to categorize and represent the information to the user within the top few matches. The traditional sequential listing of results is posing some challenges.

gines to display them as the top results for a set of keywords. This can make the search results polluted, with more relevant links being pushed down in the result list. This is a result of the popularity of Web searches and the business potential search engines can generate today.

8. Search engines index all the content of the Web without any bounds on the sensitivity of information. This has raised a few security and privacy flags.

With the above background and challenges in mind, we lay out the article as follows. In the next section, we begin with a discussion of search engine evolution. To facilitate the examination and discussion of the search engine development's progress, we break down this discussion into the three generations of search engines. Figure 1 depicts this evolution pictorially and highlights the need for better search engine technologies. Next, we present a brief discussion on the contemporary state of search engine technology and various types of content searches available today. With this background, the next section documents various concerns about existing search engines setting the stage for better search engine technology. These concerns include information overload, relevance, representation, and categorization. Finally, we briefly address the research efforts under way to alleviate these concerns and then present our conclusion.

BACKGROUND OF SEARCH ENGINES

Search engines have developed from just being research projects in the early 1990s (e.g., Archie, Gopher, Veronica, and Jughead) to some of the most visited Web sites such as Google, Yahoo!, and AskJeeves (Marckini, 2001). We classify search engines as first, second, and third generation (Nobles, 2003) as a guide for discussion and not to deduce much from the classification.

First-Generation Search Engines

This generation includes the early entrants in the search domain. Search engines existed even before the invention of the World Wide Web. Search tools such as Archie and Veronica searched using FTP (file transfer protocol) and gopher protocols long before HTTP (hypertext transfer protocol) came into play. Search technology was very primitive. With Archie one could search for file names, while with Veronica, one could search for text files and file names. The World Wide Web and its standards enabled the development of robots and crawler or spider software that "wandered" around the Web to index all the links it could find to perform a search on them later. MIT's World Wide Web wanderer's launch in 1993 was the first of many spiders such as RSBE, WWWworm, and Jumpstation. ALIWEB (Archie-Like Indexing of the Web) debuted in 1994; instead of relying on a Web spider to create an index, it relied on webmasters of participating Web sites to post their own index information for each Web page that they wanted to be listed. Thus, ALIWEB gave webmasters their first opportunity to use META tags to position their sites on a search tool. This gave birth to the META tag within the HTML (hypertext markup language) header of a Web page. The popularity of the Web and introduction of Web browsers such as Netscape and Internet Explorer gave rise to the development of search engines with Web interfaces such as Yahoo!, WebCrawler, Lycos, and Galaxy. This in turn gave rise to the second generation of search engines, discussed next.

Figure 1. Search-engine evolution