JOURNAL OF OBJECT TECHNOLOGY

Online at http://www.jot.fm. Published by ETH Zurich, Chair of Software Engineering ©**JOT, 2008**

Vol. 7, No. 3, March-April 2008

On Web Search: Some Activities and Challenges

Won Kim, Ok-Ran Jeong, Hyungsuk Ji, Sangwon Lee

Abstract

Internet search engines have become an indispensable part of everyday living and business today. Although the capabilities of Internet search engines are incrementally improving steadily, it may be time for us to explore a few new directions that can take the search engines to the next level. In this article, we will summarize the current activities in advancing the state of Internet search engines, and explore a few directions of research and development.

1 INTRODUCTION

Thanks to Internet search engine companies such as Google, Yahoo, Baidu, MSN, Naver, etc., and mass participation and collaboration by hundreds of millions of people around the world, people today are able to find all sorts of knowledge they seek instantly, inexpensively, and from almost anywhere. A quiet but tectonic revolution has been unfolding. Inspired by Internet search engines, search engines are being included within large Web sites, such as e-commerce sites, corporate sites, and social networking sites. However, search engines have very difficult technical challenges to wrestle with, for the problems they have set out to address are the longstanding natural language understanding problems. As it is exceedingly difficult to have software understand free-form text in arbitrary domains, search engines have taken to analyzing user-supplied keywords, search histories that give clues, demographic profiles, the number of links into Web pages, etc. to match the user's intent with the contents of the Web pages. Although this approach has been tuned and optimized to achieve remarkable success and delivered tremendous value to people, there are some serious long-term challenges. In the remainder of this article, we will briefly summarize the current technical activities led by Internet search engine companies to advance the state of Internet search engines, and examine a few new directions of research and development to meet the long-term challenges.

2 TECHNICAL ACTIVITIES

In China, Baidu is the market leader, and in Korea, Naver is the dominant market leader in the search and online advertising market. However, worldwide, and in the

Cite this writing as follows: Won Kim, Ok-Ran Jeong, Hyungsuk Ji, Sangwon Lee: "On Web Search: Some Activities and Challenges", in Journal of Object Technology, vol. 7, no. 3, March - April 08, pp. 49 - 54 http://www.jot.fm/issues/issue_2008_03/column5/

US, Google is the dominant leader, with Yahoo, MSN, and Ask trailing rather badly. These leading search engine companies, an array of lesser-known search engine companies, and a host of startups are working hard to make Internet search engines more accurate and more useful. Broadly, the technical activities fall into four categories.

First is to improve the quality of the search results by tuning search algorithms. Hakia, Accoona, and Powerset are some of the startups trying to incorporate more sophisticated algorithms into their search engines [Stross 2007]. We will examine this aspect further in the next section.

Second is to radically improve the presentation of the search results by organizing them into sub-themes and including multimedia data. This has become possible because of the wide adoption of broadband in the homes, and the increasing computing power and memory capacity in the desktops and laptops. Such sites as Naver, in Korea, already present the search results in a number of categories. For example, the results of a search for "Seo TaeJi" (the lead singer of a now-disbanded but still enormously famous rock band in Korea) are shown in 15 different categories: search result, sponsored links, jisik-iN (knowledgeable person), blogs, dictionary, café, sites, books, news, music, videos, images, in-depth materials, Web documents, and related search terms. The number of sub-themes may be a bit too much; however, the general approach appears to be the way to go. Mahalo's hand-built search-results page, similarly, groups the search results into sub-themes. For example, the search results on "Paul Potts" (the silky and powerful voiced winner of 'Britain's Got Talent' show in 2007) are grouped into 11 sub-themes: search result, (Paul Potts) photos, videos, gossip and blogs, news and articles, biographies and profiles, timeline, tour schedule, DVDs and merchandise, related searches, and user recommended links (for Paul Potts).

It is not just the Internet upstarts that attempt to improve the presentation of the search results [Helft 2007]. Microsoft's LiveSearch service, for example, includes, as search results for "digital camera," photos and links to reviews, and shopping information, besides the standard URLs and snippets, for the most popular digital cameras. Ask.com has introduced a service called Ask3D which displays the search results in three panels that combine standard search results with suggestions for related queries, blog items, videos, photos, news articles, and shopping information. Even Google has a service called universal search which mixes videos, photos, news articles and other items with standard search results.

Third is to bring humans into the search. Search engines are emerging that mix the automated search results with submissions and votes by users, and reviews and classification by paid editors. This approach is called a hybrid search or social search. Startups offering this human-powered search include Squidoo, Sproose, NosyJoe, Bessed, ChaCha, and Mahalo [Stross 2007]. ChaCha offers its users the opportunity to chat online with a human search assistant. Bessed has users nominate best Web pages for a topic, and has its editors review and refine them. Mahalo has hired over 30 editors to create search results for 10,000 terms related to popular topical areas, including entertainment, travel, health, and technology.

Fourth is to bring search technology to the enterprise. Internet search engines have been developed for people to look for Web pages on the open Internet. Search technology that can perform unified searches of all types of documents has become very important in enterprises. Enterprises store and manage their data using a wide variety of software systems, including file systems, database systems, data warehouses, content management systems, ERPs, CRMs, SCMs, email servers, Web servers, etc. Security and authorization are very important to enterprises. To meet the needs of the enterprises of different scales, various enterprise search products have come to the market [Hoover 2007]. Autonomy, Endeca, Fast Search and Transfer offer scalable and sophisticated products for large enterprises. Microsoft plans an upgrade of its SharePoint Server for large enterprises. For small enterprises, Google, a joint venture of IBM and Yahoo, and Microsoft offer cheap or free, but limited-capabilties, products.

3 CHALLENGES

There are several challenges to Internet search engines. One of them is the obvious need for continued improvement in the quality of the search results by better matching the needs, preferences and intents of the user with the contents of the Web pages, or, equivalently, by reducing irrelevant search results. Today many irrelevant search results often occupy the first few pages of the search results, forcing people to wade through the search results, and iterate the search process by typing in different keywords or expressions, hoping to find what they look for.

There are various types of irrelevant search results. A first type is due to the search being outside the user's needs and preferences. An example is a search for "Korean consulate general's office in Houston." The user wants the homepage of the Korean consulate general's office in Houston, Texas, United States. The search results do not include the homepage of the consulate general's office of Korea in Houston. The top search result happens to include a correct link, among several other links. Other top results include the Web pages of a scholarship fund, a training center for overseas study, etc., that happen to contain the term "Korean consulate general's office in Houston." Another example is a search for a brief definition and some examples of the term "labeled data" (in supervised machine learning). The top search result is irrelevant "Briefs-Label Data Standards," and the next one is an announcement of a workshop on machine learning, and the next one is an esoteric technical paper on supervised learning, and so forth.

A second type of irrelevant search results is spam or commercials. High ranking search results often contain advertisements, or short abstracts of books or articles that one needs to purchase. These are posted to the Web by design. An example is a search for "photos of antique brass bottles." Five of the top search results are advertisements. The hybrid search engines mentioned in the previous section can have their editors sift spam. However, spam poses a serious problem for automated search engines.

A third type of irrelevant search results arises from ambiguous words or concept level mismatches. Examples of the former include "apple" the fruit vs. "Apple" the company; "Pascal" the person vs. "Pascal" the programming language, "match" in sports vs. "match" for lighting fire, "conductor" of an orchestra vs. "conductor" of a trolley, etc. Examples of the latter include "China" and "Asia," "action movies" and "movies," etc.; that is, China "is a kind of" Asian country, action movies are "a kind of" movie, etc. If one is interested in learning about "China slavery ring," one should not type in "Asian slavery ring."

In natural languages, people can express the same meaning in different ways, use different words to express the same meaning with different nuances, convey cynicism with exactly the opposite words, use slang expressions, use coded or abbreviated words and expressions, etc. This in general is too much for any software to understand. For example, when one types in "leave Britney alone" on YouTube, it instantly delivers a video of a young man bitterly telling the world not to drag the singer Britney Spears through mud. However, if one types in "do not bother Britney," YouTube cannot find the same video. In restricted domains, by building a large database of related expressions, synonyms, different nuances, different emotions they convey, etc., this type of challenge can be met to some extent.

The recently emerging efforts to improve the presentation of the search results by organizing them in sub-themes may be further generalized into "multi-faceted" search. Much of the world's knowledge has more than one facet. For example, "Tae Kwon Do" (the Korean martial art) has history, techniques, governing body, rules in tournaments, luminaries, schools, books, movies, etc. Although all facets of Tae Kwon Do may be in a single Web site, different people are likely to be interested in different facets. The practice of organizing the search results, for example, as shown earlier, on "Paul Potts" in terms of photos, videos, gossip and blogs, news and articles, biographies and profile, etc. may be made general by applying it to most of the world's knowledge, rather than just 10,000 common terms. Grouping the search results in terms of various facets (or sub-themes) would often be helpful to the users, since the facets can correspond more closely to the needs, interests, and preferences of the users. The users may select the right facets from the search results, rather than having to wade through pages of a heterogeneous mixture of the URLs and snippets of relevant results, irrelevant results, and even spam.

There is one oddity in today's search paradigm. The Web pages are hierarchically structured, and the hierarchical structure itself embodies the relationships among the Web pages. However, as the search paradigm is strictly "input = keywords, output = URLs", the relationships among the Web pages are left to the users to determine by navigating through them. It would be helpful if the search engines can reduce the need for the users to navigate through the hierarchical structures of the Web pages they receive as search results. For example, suppose that a student is interested in learning about the research and publication activities on "Web technology and u-commerce" of the computer science departments of several graduate schools in the US. He will need to visit the homepages of all the universities he is curious about, and navigate through them to learn about the professors in the "Web technology and u-commerce" or related research group, their publications, Ph.D. theses, etc. Since there is a reasonable degree of commonality in the hierarchical structures of the homepages, for example, of universities in the US, and the labels on the Web pages, it may be useful if the user can receive the hierarchical structure of a part of the homepage (i.e., a partial Web rooted at the computer science department's research areas, rather than the entire university) of each of the universities he is interested in. There are some technical issues to overcome before this can be realized. For example, although the structure used for each research group within a particular computer science department (e.g., the University of Illinois at Urbana-Champaign, USA) is rather uniform, there are some small but non-negligible differences: the faculty for some research groups is divided into subgroups, some research groups have “seminars”, some research group has “affiliated faculty” and “related researchers,” while other research groups do not. Further, on the homepage of the UIUC computer science department, there are groups named “Algorithms and Theory” and “Parallel Computing and System,” while on the homepage of the Cornell University computer science department, apparently the same groups are named “Theory of Computing,” and “Scientific and Parallel Computing.” Such intra and inter-heterogeneity needs to be dealt with.

4 CONCLUDING REMARKS

The World Wide Web and the search engines are in the center of a revolution in the history of mankind that we are witnessing today, and, despite the exceedingly difficult nature of the problem of matching elementary expressions by hundreds of millions of people with the world’s accumulated knowledge stored on the World Wide Web, the services that the search engines have been able to deliver are truly amazing and wondrous. In this article, we first summarized the current technical activities by search engine companies aimed at improving the quality of the search results. Then we outlined a few research and development directions to meet the long-term challenges facing the search engines, hoping that these will lead to the next step up in the evolution of the search engines, and consequently, much better services to the people of the world.

ACKNOWLEDGMENTS

This research was supported by the MIC (Ministry of Information and Communication), Korea, under the ITRC (Information Technology Research Center) support program supervised by the IITA (Institute of Information Technology Advancement) (IITA-2008-(C1090-0801-0046)).

REFERENCES

[Helft 2007] Miguel Helft. “New-Look Search Sites Aim to Close Google Gap,” The New York Times, September 27, 2007.

[Hoover 2007] J. Nicholas Hoover. “Microsoft’s Ready To Be Your Enterprise Search Vendor,” Information Week, Nov. 10, 2007.

[Stross 2007] Randall Stross. “The Human Touch That May Loosen Google’s Grip,” The New York Times, June 24, 2007.

About the authors

Won Kim is a Professor and Univeristy Fellow with the School of Information and Communication Engineering at Sungkyunkwan University, Suwon, S. Korea. He is Editor-in-Chief of ACM Transactions on Internet Technology (www.acm.org/toit). He is Global General Chair of the Human.Society@Internet International Conference. He is the recipient of the ACM 2001 Distinguished Services Award, and is an ACM Fellow. He can be reached at wonkim@skku.edu

Ok-Ran Jeong is a research professor with the School of Information and Communication Engineering at Sungkyunkwan University, Korea. She was a visiting scholar in the computer science department of the University of Illinois at Urbana-Champaign, and, before that, a post doctoral researcher in the Center for e-Business Technology in Seoul National University. She received a Ph.D. in computer science from Ewha Womans University. Her research interests include Web technology (Web architecture, Web mining, intelligent techniques) and u-commerce applications. She can be reached at orjeong@ece.skku.ac.kr.

Hyungsuk Ji is a research professor with the School of Information and Communication Engineering at Sungkyunkwan University, Korea. He received a Ph.D. in cognitive science from the Institut National Polytechnique de Grenoble, France. His research interests include corpus linguistics, cognitive linguistics, semantic representation with a computational model and other areas in natural language processing (computational linguistics). He participated in developing the Atlas Project http://dico.isc.cnrs.fr.

Sang-Won Lee is an Assistant Professor with the School of Information and Communication Engineering at Sungkyunkwan University, Suwon, S. Korea. Before that, he was a research professor at Ewha Womans University and a technical staff at Oracle, Korea. He received a Ph.D degree from the Computer Science Department of Seoul National University in 1999. His research interest is in flash-based database technology. He can be reached at swlee@skku.edu