Web Search Tools: Recent Developments

Heting Chu
Associate Professor
Palmer School of Library & Information Science
Long Island University
720 Northern Blvd.
Brookville, NY 11548, U.S.A.
E-mail: hchu@liu.edu

【ABSTRACT】
Recent developments of Web search tools are illustrated and discussed with particular emphases on precise search, natural language search, multilingual search and translation, new ranking and selection techniques, multimedia searching, filtered or customized search, intelligent agents as well as browsing and searching integration. In addition, the author explores the future direction of Web search tool developments, which includes efforts in increasing search precision, and enhancing concept search capability of those services.

Introduction
Web search tools, in the form of search engines and directories, have been flourishing in both quality and quantity since 1994. Numerous studies have been conducted to describe and explore their capabilities in recent years with the intention of improving them (Koch, 1999; Schwartz, 1998). Indeed, existing features of Web search tools have been enhanced and new features are being added as the designer and the end user alike learn more about this latest system for information retrieval. The current paper attempts to summarize and discuss recent developments of Web search tools. A brief outlook into the future direction of Web search development will also be provided.

Recent Developments
Precise Search
Web search tools, in general, are notorious for its low precision search performance. The
problem becomes so prominent that people tend to accept the following expression: “You will be guaranteed to retrieve what you are not looking for when searching on the Web.” Most Web search tools present thousands, if not millions, of items to the end users for any single query. Yet, the end user neither has the time to sift through the results nor the skills to narrow down the search.

Controlled vocabulary has traditionally been used to solve the problem of low precision in other search environments, such as online systems and print indexes. It is natural for people to resort to controlled vocabulary for increasing precision of Web searches. OCLC’s NetFirst (http://www.oclc.org/oclc/netfirst/index.htm), using Dewey Decimal Classification and Library of Congress Subject Headings, is one of the pioneer projects in endeavors of this kind. Similar efforts can also be found at BUBL LINK (http://link.bubl.ac.uk/), Napier University Library’s Internet Resource (http://www.napier.ac.uk/depts/library/intres/ir000999.html) and SOSIG (The SOcial Sciences Information Gateway, http://www.esrc.bris.ac.uk/roads/cgi/search.pl). McKiernan’s “Beyond Bookmarks: Schemes for Organizing the Web” (http://www.public.iastate.edu/~CYBERSTACKS/CTW.htm) strives to bring together various projects applying controlled vocabularies in organizing and representing Web documents. However, searching the Web with controlled vocabulary remains a big challenge in many ways given the huge size and ever changing nature of the Web as well as the complexities involved in asking the public to use controlled vocabulary for searching.

XML (eXtensible Markup Language), a standard completed in 1998 by the World Wide Web Consortium (W3C), is designed, among other purposes, to remedy the problem of low precision in Web searching by marking up what a particular piece of information is about in a Web document (Bosak and Bray, 1999). In contrast, the famed HTML can only specify what the information looks like. In other words, XML can be used to index Web documents semantically with the help of such tags as <price> ... </price> and <date> ... </date>. Once a Web document is indexed semantically rather than by its appearance, accuracy in Web resource retrieval can be expected to improve. Network users, especially those in the business sector, embrace XML with great enthusiasm and start implementing it for e-commerce and similar tasks. XML, by its very nature, will have a significant impact on improving the precision of Web searches.

**Natural Language Search**

In the previous section, we discussed how controlled vocabulary could increase the precision of network searching. On the other hand, end users would always desire to conduct searches the same way as they ask a question, thus eliminating the need for doing concept analysis, selecting search term(s) and constructing search strategies. Ask Jeeves (http://www.aj.com) appears to be just the right search tool that meets the aspiration of natural language search.

Ask Jeeves made its debut in the Web searching arena in 1998. It allows people to enter a search query in question format (e.g. Why is the sky blue?). It then decomposes the question automatically behind the scene following the
procedure required for any common searches, and present search results to the user. In doing so, the user is relieved from the most difficult part of searches, which includes such tasks as concept analysis and the use of Boolean operators.

InQuizzit (http://www.inquizzit.com) also offers natural language search products and services by implementing natural language understanding techniques researched by many talented minds for years. However, it now only sells its products for profits, and its free service has been discontinued shortly after it was introduced to the public via New York Times (Riordan, 1998).

In sum, natural language searching has a long way to go even in other computerized search environments such as online systems or CD-ROM databases. Nevertheless, we know that the Web has been a showcase for many new searching technologies. Therefore, we would expect to see more advancement in natural language searching on the Web.

**Multilingual Search & Translation**

As Internet is a global network, it carries multilingual information even though English appears to be the dominating language. Therefore, the tools created for searching such information have to address this multilingual issue by either supporting multilingual search or providing translation facilities. AltaVista has taken the lead in both areas while many other search services (e.g., Excite, HotBot and Yahoo!) followed suit especially in the domain of multilingual searching (He, 1999).

When Web search tools were initially developed, English was the only language that people could use for searching. However, AltaVista swiftly introduced multilingual searching not long after its inception, and has ever since kept enhancing this feature in collaboration with other companies specialized in language processing. At present, people can search the Web using more than two dozen different languages (e.g., English, Chinese, French, German, Spanish and Russian). In addition to multilingual searching, AltaVista also offers translation capability using the Systran software (http://www.systransoft.com) so that any search results can be translated from English to French, German, Spanish, Portuguese, or Italian, and visa versa (Balkin, 1999). Although the translation quality is not quite satisfactory due to the inherent limitation of automatic translation, AltaVista did its best in this respect.

Compared with AltaVista, other search tools such as Excite and Yahoo! can only support multilingual searching. The number of languages they cover is also smaller than AltaVista. Even so, these search tools together with AltaVista make it possible for people to enjoy the multilingual and international nature of information available on the Internet.

**New Ranking/Selection Techniques**

Ranking mechanism has been introduced to Web searching as an effort to help the user in locating a “needle” in a huge “hay stack.” Algorithms commonly employed for ranking search results include term frequency, term proximity, term location, document date, term completeness, and term weight (Huwe, 1999). While these algorithms rank search results according to the intrinsic attributes of Web docu-
ment, the following three ranking methods focus on the extrinsic features generated by people when they access the target Web documents or when they create their own sites.

Google! (http://www.google.com) assumes that the greater number of times a site is pointed to by other sites, the more important it becomes. This practice appears very similar to citation analysis - an established method used for, among other things, evaluating the quality of paper publications. Borrowing some terminology from citation analysis, we can call the site being pointed to as the "cited site" while the site referring to the cited site becomes the "citing site." In other words, Google! has developed a mechanism for analyzing and ranking the links pointing to a cited site using PageRank, an algorithm named after one of its creators - Larry Page. The importance of citing sites is also considered in the rank computation. Vidmar (1999) refers to such practice as "relevance ranking based on backlinks."

Direct Hit (http://www.directhit.com), also known as "the popularity engine," ranks a site by how many users actually visit (thus the word "hit") it. If we regard AltaVista and similar search engines as author-controlled services (i.e. search results are determined by how well keywords match the content of a Web page), if we call Yahoo and similar directory-based search tools as editor controlled services (i.e. editors locate and catalog sites by examining them one by one), Direct Hit represents a third kind of search mechanism: user-controlled service in which search rankings depend upon the choices made by other users (Frauenfelder, 1998). The higher the number of times a site is visited by others, the more popular that site becomes. Clever(http://www.almaden.ibm.com/cs/k53/clever.html) adopts an algorithm developed by Jon Kleinberg for analyzing the myriad of hyperlinks and then identifying two kinds of useful Web sites: 1) authorities - the best sources of information on a particular topic, and 2) hubs - collections of links to those locations. This process is called "hypersearching" by its developers. (Chakrabarti et al, 1999) Two steps are involved in ranking Web sites. First, sites are ranked by using the traditional ranking methods such as frequency and proximity for analyzing hyperlinks contained within. Second, these hyperlinks are further scrutinized using the Web citation analysis method that Google! employs. The reiteration of these two steps locates and fine-tunes search results which are eventually presented as pages of relevant links, separated into hubs and authorities. (Frauenfelder, 1998) Clever is officially an IBM's research project. So far it offers only demo searches.

While traditional ranking methods may cause problems such as word stuffing (i.e. repeating some keywords deliberately in order to get a better ranking) or make little sense in situations like "pay for placement" (i.e. ranking search results based on the amount of fees paid), the three techniques described above open up a new dimension for ranking and judging Web sites. Obviously, all three practices are still at a developing stage. They nevertheless look to be quite promising in improving network search performances.

**Multimedia Searching**

One distinctive feature of the Web is its ability of storing, transferring and presenting
multimedia information. The need for locating and obtaining multimedia is thus steadily increasing as the Web grows rapidly and becomes popular among network users.

Multimedia information comprises any combination of audio, video, images and texts. While the traditional method, being dubbed as “description-based search” (i.e. using description data in text form such as artist, year, or title of work), is still applicable for searching multimedia, a new approach named “content-based search” (i.e. using multimedia attributes such as color, texture, layout or frequency) for locating multimedia is gaining recognition and momentum. Up to this moment, search tools have been most successful for searching still images (e.g. pictures and photos) as opposed to moving images like movies or videos.

Initially, tools specialized in image searching (e.g. QBIC -http://www.qbic.almaden.ibm.com/ and WebSeek -http://www.ctr.columbia.edu/webseek/) were developed. Later, general purpose search tools such as AltaVista (http://image.altavista.com) and Yahoo!(http://ipix.yahoo.com) have incorporated image searching capability into their operations. All these tools employ both the description as well as content-based approaches in their search algorithms because neither method yields perfect search results alone. Aside from what has been discussed above, HotBot and some other tools also explored image (both still and moving) searching by making use of file extension (e.g. gif, jpg, mpg, and mov), HTML tags (e.g., <img...>), and other features (e.g. image related keywords in page titles) imbedded in Web documents.

In comparison, search tools specifically devoted to audio and video information retrieval are small in quantity and scale. While tools such as HotBot can also be used to search for audio/video information by specifying file type (e.g. audio and video) and file extension (e.g. au and wav) in advanced searches, they are not sufficient for locating the large amount of audio data available on the Web particularly as a result of the newly developed music format - MP3. Therefore, search tools (e.g. Snap - http://snap.com) have been tailored or designed for searching audio/video information. For instance, Snap has recently formed partnership with Real Networks, a major multimedia software maker, to imbed Snap search engine in Real Player G2, a free program from Real Networks that enable people to listen to music and watch video over the Internet. Under that plan, users of Real Player program can search the Internet for music, other audio or video files that can be played using the software. In addition, Snap and Real Networks are going to jointly start a multimedia search site (http://speed.snap.com) for rich graphics, and for locating and downloading audio and video at faster speed (Richtel, 1999).

Despite these efforts, it still remains relatively easier for people to find textual information than multimedia on the Internet. However, researchers continue to explore new mechanisms, especially from the content-based perspective, for multimedia searching.

**Filtered or Customized Search**

All the Web search tools are built for people to use. Yet end users always have various backgrounds and different information needs. In order to serve the user effectively, many search services
promote filtered or customized searches. Such services are also known as collaborative filtering or personalized search. "Collaborative filtering" is a method for recommending choices to the user based on the user's answers to selected questions and the answers of other people in the past. It presumes that people who have agreed with each other previously on certain matters will most likely agree again in the future. Currently there are quite a few programs for collaborative filtering (mdigiova@eddy.berkeley.edu). The word "filter" in this context means selecting appropriate information for the end user rather than the traditional implication of separating the good from the inappropriate (e.g. Web filtering programs for blocking offensive and pornographic sites away from children). In other words, the objective for having filtered searches is to provide the right information to the right user.

iAtlas (http://www.iatlas.com) is a search tool specially designed for reducing information overload. In cooperation with Inktomi, infoUSA and many others, iAtlas has developed a filtering technology that includes keyword search, business filters, geographic filters, Web filters, and content filters. Obviously, the user can filter what s/he finds based on keyword search using any combination of the four filters iAtlas provides. With the business filters, one can choose from industry (e.g. cars or doctors), company size (e.g. 1-19 employees), and type (e.g. public or single location). Using the geographical filters, the user can select metro area (e.g. Atlanta or New York) and state (e.g. California or New Jersey). The Web filters allow the user to specify a particular Web site, page length, and number of links (e.g. 5-10 links). The content filters cover features or document type (e.g. images or Java) and last modified (e.g. last modified before January 11, 1999).

Many common-purpose search tools such as Excite and InfoSeek (now part of the Go Network - http://infoseek.go.com) also implement filtering in their general framework. For example, InfoSeek/Go has a link named "Personalized Start Page." The user normally can click the button for personalization at a search site and then create a profile by specifying some personal attributes such as gender and e-mail address. The profiles then become the basis for the search tools to provide filtered or personalized services. As the creation of profiles may affect one's privacy, users should be careful when asked about personal information by Web search services.

**Intelligent Agents**

Another recent development of Web search tools is intelligent agents, which are relevant to the filtering concept we discussed in the previous section. The major difference between an intelligent agent and a filter however lies in that the former employs artificial intelligence (AI) techniques while the latter may not. Intelligent agents, also known as bots, are software programs that are able to perform certain tasks without human intervention.

BotSpot (http://www.bobspot.com) is a Yahoo! type directory for all kinds of bots available on the net. People working at BotSpot classify intelligent agents by subject, with reviews and historical perspectives. There are also intelligent agents designed for specific purposes. For instance, TracerLock (http://peacefire.org/tracerlock) monitors search engines
and notifies the registrant by e-mail when new pages appear on the Web containing specific key-
com) searches for information specified by the
user in lists and Usenet newsgroups, and then e-
mail back search results on a daily basis (Rankin,
1999). It closely resembles to the SDI (Selective
Dissemination of Information) service provided
by librarians and other information professionals
over the years.

In addition, intelligent agents are used in
classifying and clustering materials collected by
Web search tools. For example, Northern Light is
able to organize search results into custom search
folders such as "Higher education" and
"Education administration." Meta-search tools
(e.g. Inference Find - http://www.inference.com/)
have also made use of intelligent agents for simi-
lar purpose. As research in artificial intelligence
advances, more applications using the AI tech-
nique will be developed for Web searching, a fast
growing area in information retrieval.

**Browsing & Searching Integration**

All the Web search tools can be broadly cat-
ergized into two types: directories such as
Yahoo! for browsing and search engines like
Altavista for searching. When one intends to
search for a specific piece of information, search
engines should be the appropriate choice. Other-
wise, directories should be used if the user is
looking for information about a topic or simply
would like to browse to find out what is available.
Such kind of distinction could be easily discerned
when Web search tools were just created in mid
90s. However, as each type of search tools has its
own strengths and weaknesses, it became the
norm rather than the exception that both browsing
and searching are supported at one single search
site. The boundary between browsing and search-
ing is diminishing in the mist of this integration
adopted by most search tools.

Altavista in the past tried to maintain its dis-
tinctive searching feature and refrain from the
inclusion of a resource directory. However, it
began to incorporate the directory developed by
LookSmart at its site as directories become
increasingly popular among search tool users.
Lycos and HotBot, both a search engine by
nature, now list Open directory at their home
pages. Yahoo!, the best known directory in the
history of Web search tools, has partnered with
various search engines (the most recent one is
Inktomi) to allow its user to conduct searches in
addition to browsing. Within this integrated
framework, the user can choose between brows-
ing and searching at the same place or perform
searches within a browsable category. In a sense,
the integration of browsing and searching greatly
enhances the user-friendliness of Web search
tools as the user no longer needs to switch
between different sites in order to get the right
browsing or searching capability.

**Possible Future Directions**

It is apparent that Web search tools, as an
emerging species for information retrieval, have
made significant progress since their inception.
However, the problem of low precision in net-
work retrieval remains very serious despite various efforts and attempts for improvement.
The essence of the problem is rooted in the nature
of data available from the Internet and the way it
has been processed and represented. While Web resources are rich, they meanwhile become obsolete instantly and are of mixed quality. Furthermore, effective algorithms for representing information on the Web are to be found. Purely extracting keywords or phrases from networked documents (e.g. Web pages) will not alleviate or eliminate the low precision problem of Web searching at all.

Among the other features desirable for Web search tools of next generation, concept search appears to need immediate attention. How could search tools in the future not merely match a query with keywords/phrases in a database? For example, when one searches for information about "violence among athletes," documents that happen to mention this phrase should not be retrieved. Excite remains to be one of the few search tools that support concept searching. But their performance still shows the deficiency of their concept search capability. On one hand, the gigantic amount of data on the Web can only be efficiently processed and represented using automatic methods. On the other hand, automatic indexing algorithms developed so far would just permit people to search the Web mechanically (i.e. term matching) instead of intellectually (i.e. concept searching). Further research advancement in artificial intelligence seems to be the only possible approach for successfully implementing concept search in Web search tools using the automatic method.

**Concluding Remarks**

Web search tools are still evolving and developing even though they have been improved and enhanced constantly in the past few years. New features and capabilities will be introduced and get incorporated into existing search tools. To keep up-to-date with the new developments of Web search tools, we should closely monitor those search sites. Meanwhile, the following three sites also encapsulate exactly what we are looking for:


**References**


