Chapter 10

World Wide Web Search Engines

Wen-Chen Hu
Auburn University

Jyh-Haw Yeh
Boise State University

ABSTRACT

The World Wide Web now holds more than 800 million pages covering almost all daily issues. The Web's fast growing size and lack of structural style present a new challenge for information retrieval. Numerous search technologies have been applied to Web search engines; however, the dominant search method has yet to be identified. This chapter provides an overview of the existing technologies for Web search engines and classifies them into six categories: i) hyperlink exploration, ii) information retrieval, iii) metasearches, iv) SQL approaches, v) content-based multimedia searches, and vi) others. At the end of this chapter, a comparative study of major commercial and experimental search engines is presented, and some future research directions for Web search engines are suggested.

INTRODUCTION

One of the most common tasks performed on the Web is to search Web pages, which is also one of the most frustrating and problematic. The situation is getting worse because of the Web's fast growing size and lack of structural style, as well as the inadequacy of existing Web search engine technologies (Lawrence & Giles, 1999a). Traditional search techniques are based on users typing in search keywords which the search services can then use to locate the desired Web pages. However, this approach normally retrieves too many documents, of which only a small fraction are relevant to the users' needs. Furthermore, the most relevant documents do not necessarily appear at the top of the query output list. A number of corporations and research organizations are taking a variety of approaches to try to solve these problems. These approaches are diverse and none of them dominate the field. This chapter provides a survey and classification of the available World Wide Web search engine techniques, with an emphasis on non-traditional approaches. Related Web search technol-

ogy reviews can also be found in (Gudivada et al., 1997; Lawrence & Giles, 1998b; Lawrence & Giles, 1999b; Lu & Feng, 1998).

Requirements of Web Search Engines

It is first necessary to examine what kind of features a Web search engine is expected to have in order to conduct effective and efficient Web searches and what kind of challenges may be faced in the process of developing new Web search techniques. The requirements for a Web search engine are listed below in order of importance:

1. Effective and efficient location and ranking of Web documents.
2. Thorough Web coverage.
3. Up-to-date Web information.
4. Unbiased access to Web pages.
5. An easy-to-use user interface which also allows users to compose any reasonable query.
6. Expressive and useful search results.
7. A system that adapts well to user queries.

Web Search Engine Technologies

Numerous Web search engine technologies have been proposed and each technology employs a very different approach. This survey classifies the technologies into six categories: i) hyperlink exploration, ii) information retrieval, iii) metasearches, iv) SQL approaches, v) content-based multimedia searches, and vi) others. The chapter is organized as follows: Section 2 introduces the general structure of a search engine and Sections 3 to 8 introduce each of the six Web search engine technologies in turn. A comparative study of major commercial and experimental search engines is shown in Section 9 and the final section gives a summary and suggests future research directions.

SEARCH ENGINE STRUCTURE

Two different approaches are applied to Web search services: genuine search engines and directories. The difference lies in how listings are compiled.

- Search engines, such as Google, create their listings automatically.
- A directory, such as Yahoo!, depends on humans for its listings.

Some search engines, known as hybrid search engines, maintain an associated directory. Search engines traditionally consist of three components: the crawler, the indexing software, and the search and ranking software (Greenberg & Garber, 1999; Yuwono & Lee, 1996). Figure 1 shows the system structure of a typical search engine.

Crawler

A crawler is a program that automatically scans various Web sites and collects Web documents from them. Crawlers follow the links on a site to find other relevant pages. Two search algorithms, breadth-first searches and depth-first searches, are widely used by crawlers to traverse the Web. The crawler views the Web as a graph, with the nodes being the objects located at Uniform Resource Locators (URLs). The objects could be HTTPs (Hypertext Transfer Protocols), FTPs (File Transfer Protocols), mailto (e-mail), news, telnet, etc. They also return to sites periodically to look for changes. To speed up the collection of Web documents, several crawlers are usually sent out to traverse the Web at the same time.

Three simple tools can be used to implement an experimental crawler:

- lynx: Lynx is a text browser for Unix systems. For example, the command "lynx -dump -source http://www.w3c.org/" downloads the Web page source code at http://www.w3c.org/.
- java.net: The java.net package of Java language provides plenty of networking utilities. Two classes in the package, java.net.URL and java.net.URLConnection, can be used to download Web pages.
- CPAN (Comprehensive Perl Archive Network): Perl has been used intensively for Web related applications. Some scripts provided by CPAN at http://www.cpan.org/ are useful for crawler construction.

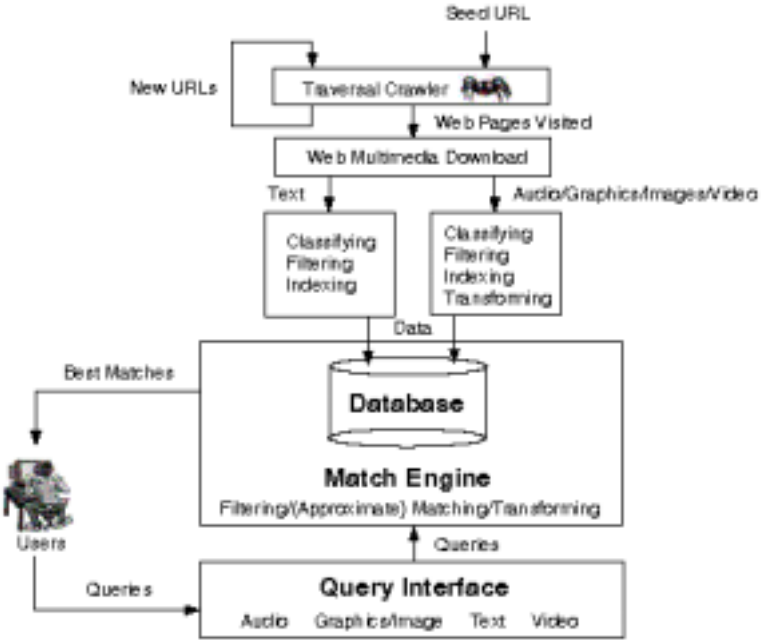
To construct an efficient and practical crawler, some other networking tools have to be used.

Indexing Software

Automatic indexing is the process of algorithmically examining information items to build a data structure that can be quickly searched. Filtering (Baeza-Yates, 1992) is one of the most important pre-processes for indexing. Filtering is a typical transformation in information retrieval and is often used to reduce the size of a document and/or standardize it to simplify searching. Traditional search engines utilize the following information, provided by HTML scripts, to locate the desired Web pages:

- Content: Page content provides the most accurate, full-text information. However, it is also the least-used type of information, since context extraction is still far less practical.
- Descriptions: Page descriptions can either be constructed from the metatags or submitted by Web masters or reviewers.

Figure 1: System structure of a Web search engine.



- Hyperlink: Hyperlinks contain high-quality semantic clues to a page's topic. A hyperlink to a page represents an implicit endorsement of the page being pointed to. (Chakrabarti et al., 1999)
- Hyperlink text: Hyperlink text is normally a title or brief summary of the target page.
- Keywords: Keywords can be extracted from full-text documents or metatags.
- Page title: The title tag, which is only valid in a head section, defines the title of an HTML document.
- Text with a different font: Emphasized text is usually given a different font to highlight its importance.
- The first sentence: The first sentence of a document is also likely to give crucial information related to the document.

Search and Ranking Software

Query processing is the activity of analyzing a query and comparing it to indexes to find relevant items. A user enters a keyword or keywords, along with Boolean modifiers such as "and," "or," or "not," into a search engine, which then scans indexed Web pages for the keywords. To determine in which order to display pages to the user, the engine uses an algorithm to rank pages that contain the keywords (Zhang & Dong, 2000). For example, the engine may count the number of times the keyword appears on a page. To save time and space, the engine may only look for keywords in metatags, which are HTML tags that provide information about a Web page. Unlike most HTML tags, metatags do not affect a document's appearance. Instead, they include such information as a Web page's contents and some relevant keywords. The following six sections give various methods of indexing, searching, and ranking the Web pages.

HYPERLINK EXPLORATION

Hypermedia documents contain cross references to other related documents by using hyperlinks, which allow the user to move easily from one to the other. Links can be tremendously important sources of information for indexers; the creation of a hyperlink by the author of a Web page represents an implicit endorsement of the page being pointed to. This approach is based on identifying two important types of Web pages for a given topic:

- Authorities, which provide the best source of information on the topic, and
- Hubs, which provide collections of links to authorities.

For the example of professional basketball information, the official National Basketball Association site http://www.nba.com/ is considered to be an authority, while the ESPN site http://www.espn.com/ is a hub. Authorities and hubs are either given top ranking in the search results or used to find related Web pages (Dean & Henzinger, 1999).

Analyzing the interconnections of a series of related pages can identify the authorities and hubs for a particular topic. A simple method to update a non-negative authority with a weight x_p and a non-negative hub with a weight y_p is given by (Chakrabarti et al., 1999). If a page is pointed to by many good hubs, its authority weight is updated by using the following formula:

$$x_p = \sum_{q \text{ such that } q \to p} y_q ,$$

where the notation q®$\eth p$ indicates that q links to p. Similarly, if a page points to many good authorities, its hub weight is updated via

$$y_p = \sum_{q \text{ such that } p \to q} x_q \ .$$

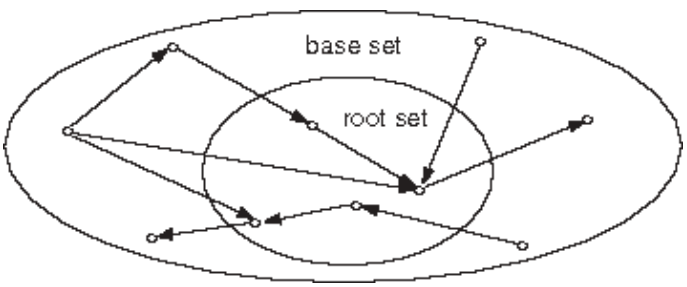
Unfortunately, applying the above formulas to the entire Web to find authorities and hubs is impracticable. Ideally, the formulas are applied to a small collection $S_{sð}$ of pages which contain plenty of relevant documents. The concepts of a root set and a base set have been proposed by (Kleinberg, 1999) to find $S_{sð}$. The root set is usually constructed by collecting the *t* highest-ranked pages for the query sð from a search engine such as Google or Yahoo!. However, the root set may not contain most of the strongest authorities. A base set is therefore built by including any page pointed to by a page in the root set and any page that points to a page in the root set. Figure 2 shows an example of a root set and a base set. The above formulas can then be applied to a much smaller set, the base set, instead of the entire Web.

In addition to the methods used to find authorities and hubs, a number of search methods based on connectivity have been proposed. A comparative study of various hypertext link analysis algorithms is given in (Borodin et al., 2001). The most widely used method is a Page Rank model (Brin & Page, 1998), which suggests the reputation of a page on a topic is proportional to the sum of the reputation weights of pages pointing to it on the same topic. That is, links emanating from pages with high reputations are weighted more heavily. The concepts of authorities and hubs, together with the Page Rank model, can also be used to compute the reputation rank of a page; those topics for which the page has a good reputation are then identified (Rafiei & Mendelzon, 2000). Some other ad hoc methods include an HVV (Hyperlink Vector Voting) method (Li, 1998) and a system known as WebQuery (Carriere & Kazman, 1997). The former method uses the content of hyperlinks to a document to rank its relevance to the query terms, while the latter system studies the structural relationships among the nodes returned in a content-based query and gives the highest ranking to the most highly connected nodes. An improved algorithm obtained by augmenting with content analysis is introduced in (Bharat & Henzinger, 1998).

INFORMATION RETRIEVAL (IR)

IR techniques are widely used in Web document searches (Gudivada, 1997). Among them, relevance feedback and data clustering are two of the most popular techniques used by search engines. The former method has not so far been applied to any commercial products because it requires some interaction with users, who normally prefer to use a keyword-only interface. The latter method has achieved more success since it does not require any interaction with users to achieve acceptable results.

Figure 2: Expanding the root set into a base set.



Relevance Feedback

An initial query is usually a wild guess. Retrieved query results are then used to help construct a more precise query or modify the database indexes (Chang & Hsu, 1999). For example, if the following query is submitted to a search engine:

Which TOYOTA dealer in Atlanta has the lowest price for a Corolla 2002? the engine may produce the following list of ranked results:

1. Get the BEST price on a new Toyota, Lexus car or truck. http://www.toyotaforless.com/
2. Toyota of Glendale¾ðYour #1 Toyota dealer. http://www.toyota-of-glendale.com/
3. Leith Toyota¾ðRaleigh, North Carolina. http://www.leithtoyota.com/f_more_about_us.html
4. Atlanta rental cars & auto rentals. http://www.bnm.com/atl2.htm

This list includes three relevant results: 1, 2, and 3; and one irrelevant result: 4. The following two relevance feedback methods can be used to improve the search results:

- Query modification: Adjusts the initial query in an attempt to avoid unrelated or less related query results. For example, the above query could be modified by adding a condition excluding rental cars.
- Indexing modification: Through feedback from the users, system administrators can modify an unrelated document's terms to render it unrelated or less related to such a query. For example, the information concerning rental cars could be removed from the database indexes of car sales and prices.

For the above example, the search results after modification should not include result #4.

Data Clustering

Data clustering is used to improve the search results by dividing the whole data set into data clusters. Each data cluster contains objects of high similarity, and clusters are produced that group documents relevant to the user's query separately from irrelevant ones. For example, the formula below gives a similarity measure:

$$S_{D_i,D_j} = \frac{2\sum_{k=1}^{L}\left(weight_{ik}\cdot weight_{jk}\right)}{\sum_{k=1}^{L} weight_{ik}^2 + \sum_{k=1}^{L} weight_{jk}^2},$$

where $weight_{ik}$ is the weight assigned to $term_k$ in a document D_i (Baeza-Yates, 1992). Clustering should not be based on the whole Web resource, but on smaller separate query results. In (Zamir & Etzioni, 1998), a Suffix Tree Clustering (STC) algorithm based on phrases shared between documents is used to create clusters. Beside clustering the search results, a proposed similarity function has been used to cluster similar queries according to their contents as well as user logs (Wen, Nie, & Zhang, 2001). The resulting clusters can provide useful information for Frequently Asked Queries (FAQ) identification. Another Web document clustering algorithm is suggested in (Chang & Hsu, 1999).

METASEARCHES

None of the current search engines is able to cover the Web comprehensively. Using an individual search engine may miss some critical information that is provided by other

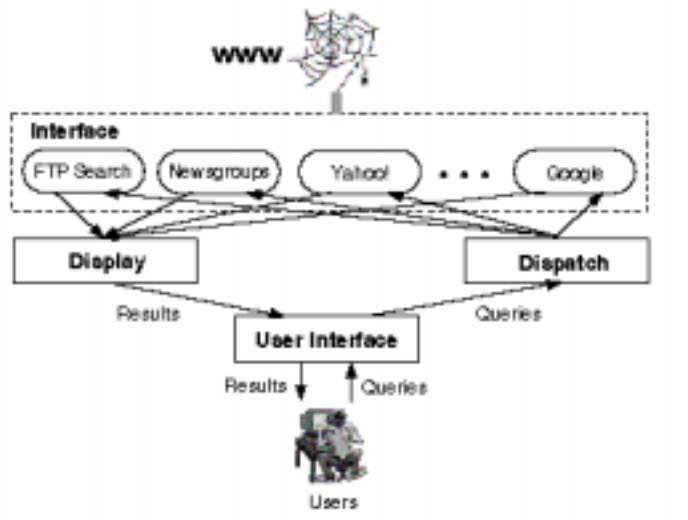
engines. Metasearch engines (Dreilinger & Howe, 1997; Howe & Dreilinger, 1997; Selberg & Etzioni, 1997) conduct a search using several other search engines simultaneously and then present the results in some sort of integrated format. This lets users see at a glance which particular search engine returned the best results for a query without having to search each one individually. They typically do not use their own Web indexes. Figure 3 shows the system structure of a metasearch engine, which consists of three major components:

- Dispatch: Determines to which search engines a specific query is sent. The selection is usually based on network and local computational resources, as well as the long-term performance of search engines on specific query terms.
- Interface: Adapts the user's query format to match the format of a particular search engine, which varies from engine to engine.
- Display: Raw results from the selected search engines are integrated for display to the user. Each search engine also produces different raw results from other search engines and these must be combined to give a uniform format for ease-of-use.

Current search engines provide a multiplicity of interfaces and results which make the construction of metasearch engines a very difficult task. The STARTS protocol (Gravano et al., 1997) has been proposed to standardize internet retrievals and searches. The goals are to choose the best sources (search engines) to evaluate a query, submit the query to the sources selected, and finally merge the query results obtained from the different sources. However, this protocol has received little recognition since none of the most-often-used search engines apply it. Another approach (Huang, Hemmje, & Neuhold, 2000) to solving this problem is to use an adaptive model which employs a "mediator-wrapper" architecture. The mediator provides users with integrated access to multiple heterogeneous data sources, while each wrapper represents access to a specific data source. It maps a query from a general mediator format into the specific wrapper format required by each search engine.

Metasearch engines rely on the summaries and ranks of URLs returned by standard search engines. However, not all standard search engines give unbiased results and this will distort the metasearch results. The NEC Research Institute (NECI) metasearch engine

Figure 3: System structure of a metasearch engine.



(Lawrence & Giles, 1998a) solved this problem by downloading and analyzing each document and then displaying results in a format that shows the query terms in context. This helps users more readily determine if the document is relevant without having to download each page. The authors of Q-pilot (Sugiura & Etzioni, 2000) noticed that thousands of specialized, topic-specific search engines are accessible on the Web and these topic-specific engines return far better results for "on topic" queries than standard search engines. Q-pilot dynamically routes each user query to the most appropriate specialized search engines by using two methods: neighborhood-based topic identification, and query expansion.

SQL APPROACHES

Learning how to use a new language is normally an arduous task for users. However, a new system which uses a familiar language is usually adopted relatively smoothly by the users. SQL (Structured Query Language) is a well-known and widely-used database language. SQL approaches (Florescu, Levy, & Mendelzon, 1998; Mendelzon & Milo, 1998) view the World Wide Web as a huge database where each record matches a Web page, and use SQL-like languages to support effective and flexible query processing. A typical SQL-like language syntax (Konopnicki & Shmueli, 1998; Mendelzon, Mihaila, & Milo, 1997; Spertus & Stein, 2000) is

Query := **select** Attribute_List **from** Domain_Specifications
[**where** Search_Conditions];

Three query examples are given below to show the use of the language.

SQL Example 1 *Find pages in the World Wide Web Consortium (W3C) site where the pages have fewer than 2000 bytes.*

select url **from** http://www.w3c.org/ **where** bytes < 2000;

url is a page's URL and each page has attributes such as bytes, keywords, and text.

SQL Example 2 *Find educational pages containing the keyword "database."*

select url **from** http://%.edu/ **where** 'database' **in** keywords;

Regular expressions are widely used in the query language, e.g., the symbol '%' is a wild card matching any string. The **in** predicate checks whether the string "database" is one of the keywords.

SQL Example 3 *Find documents about "XML" in the W3C Web site where the documents have paths of length two or less from the root page.*

select d.url, d.title
from Document d **such that** 'http://www.w3c.org/' =|®ð|® ® d
where d.text **like** '%XML%';

The symbol '|ð' is an alternation and the symbol '®ð' is a link. The string " =|®|®®" is a regular expression that represents the set of paths of length of one or two. The **like** predicate is used for string matching in this example.

Various SQL-like languages have been proposed for Web search engines. The methods introduced previously treat the Web as a graph of discrete objects; another object-oriented approach (Arocena & Mendelzon, 1998) considers the Web as a graph of structured objects. However, neither approach has achieved much success because of its complicated syntax, especially for the latter method.

CONTENT-BASED MULTIMEDIA SEARCHES

In order to allow for the wide range of new types of data which are now available on the World Wide Web, including audio, video, graphics, and images, the use of hypermedia was introduced to extend the capabilities of hypertext. The first internet search engine, Archie, was created in 1990; however, it was not until the introduction of multimedia to the browser Mosaic that the number of Internet documents began to increase explosively. Only a few multimedia search engines are available currently, most of which use name or keyword matching where the keywords are entered by Web reviewers rather than using automatic indexing. The low number of content-based multimedia search engines is mainly due to the difficulty of automated multimedia indexing. Numerous multimedia indexing methods have been suggested in the literature (Chang & Hsu, 1992; Yoshitaka & Ichikawa, 1999), yet most do not meet the efficiency requirements of Web multimedia searches, where users expect both a prompt response and the search of a huge volume of Web multimedia data. A few content-based image and video search engines are available on-line (Benitez, Beigi, & Chang, 1998; Gevers & Smeulders, 1999; Lew, 2000; Smith & Chang, 1997; Taycher, Cascia, & Sclaroff, 1997). Various indexing methods are applied to locate the desired images or video. The major technologies include using camera/object motion, colors, examples, locations, positional color/texture, shapes, sketches, text, and texture as well as relevance feedback (Flickner et al., 1995). However, a de facto Web image or video search engine is still out of reach because the system's key component¾ðimage or video collection and indexing¾ðis either not yet fully automated or not practicable. Similarly, effective Web audio search engines have yet to be constructed since audio information retrieval (Foote, 1999) is considered to be one of the most difficult challenges for multimedia retrieval.

OTHERS

Apart from the above major search techniques, some ad hoc methods worth mentioning include:

- Work aimed at making the components needed for Web searches more efficient and effective, such as better ranking algorithms and more efficient crawlers. In (Zhang and Dong, 2000), a ranking algorithm based on a Markov model is proposed. It synthesizes the relevance, authority, integrativity, and novelty of each Web resource, and can be computed efficiently through solving a group of linear equations. A variety of other improved ranking algorithms can be found in (Dwork et al., 2001; Singhal & Kaszkiel, 2001).
- Various enhanced crawlers can be found in the literature (Aggarwal, Al-Garawi, & Yu, 2001; Edwards, McCurley, & Tomlin, 2001; Najork & Wiener, 2001). Some crawlers are extensible, personally customized, relocatable, scalable, and Web-site-specific (Heydon & Najork, 1999; Miller & Bharat, 1998). Web viewers usually consider certain Web pages more important. A crawler which collects those "important" pages first is advantageous for users (Cho, Garcia-Molina, & Page, 1998).
- Artificial Intelligence (AI) can also be used to collect and recommend Web pages. The Webnaut system (Nick & Themis, 2001) learns the user's interests and can adapt as his or her interests change over time. The learning process is driven by user feedback to an intelligent agent's filtered selections.
- To make the system easier to use, an interface has been designed to accept and understand a natural language query (Ask Jeeves, 2002).

MAJOR SEARCH ENGINES

Some of the currently available major commercial search engines are listed in Table 1, although many table entries are incomplete as some of the information is classified as confidential due to business considerations (Search Engine Watch, 2002). Most search services are backed up by or are cooperating with several other services. This is because an independent or stand-alone service contains less information and thus tends to lose its users. In the table, the column Backup gives the major backup information provider, and most unfilled methods use keyword matching to locate the desired documents. Most search engines on the list not only provide Web search services but also act as portals, which are Web home bases from which users can access a variety of services, including searches, e-commerce, chat rooms, news, etc. Table 2 lists some major experimental search engines, which use advanced search technologies not yet implemented by the commercial search engines. The list in Table 2 is a snapshot of the current situation; the list is highly volatile either because a successful experimental search engine is usually commercialized in a short time or because a prototype system is normally removed after its founders leave the organization. The two tables list major general-purpose search engines; special-purpose search engines including specialty searches, regional searches, kid searches, etc. are not considered in this chapter. They use much smaller databases and therefore give more precise and limited search results.

SUMMARY

In less than a decade, the World Wide Web has become one of the three major media, with the other two being print and television. Searching for Web pages is both one of the

Table 1: Major commercial Web search engines. SE: Search Engine and AS: Answering Service.

No.	Name	URL	Type	Backup	Method
1	AOL Search	http://search.aol.com/	Hybrid SE	Open Directory	
2	AltaVista	http://www.altavista.com/	SE	LookSmart	
3	Ask Jeeves language	http://www.askjeeves.com/	AS		natural
4	Direct Hit	http://www.directhit.com/	SE	HotBot	hyperlink
5	Excite	http://www.excite.com/	SE	LookSmart	
6	FAST Search	http://www.alltheweb.com/			scalability
7	Google	http://www.google.com/	SE		hyperlink
8	HotBot	http://www.hotbot.com/	Hybrid SE	Direct Hit	
9	IWon	http://www.iwon.com/	Hybrid SE	Inktomi	
10	Inktomi	http://www.inktomi.com/	SE		
11	LookSmart	http://www.looksmart.com/	Directory	Inktomi	reviewers
12	Lycos	http://www.lycos.com/	Directory	Open Directory	
13	MSN Search	http://search.msn.com/	Directory	LookSmart	
14	Netscape Search	http://search.netscape.com/	SE	Open Directory	
15	Northern Light	http://www.northernlight.com/	SE		filtering
16	Open Directory	http://dmoz.org/	Directory		volunteers
17	RealNames	http://www.realnames.com/			keywords
18	Yahoo!	http://www.yahoo.com/	Directory	Google	reviewers

most common tasks performed on the Web and one of the most frustrating and problematic. This chapter gave an overview of the current technologies for Web search engines with an emphasis on non-traditional approaches and classified the technologies into six categories. However, apart from the traditional keyword matching techniques, no one method dominates Web search engine technologies. The major reason for this is that the amount of information posted on the World Wide Web is huge and the page formats vary widely.

Future Directions

Users of search engines often submit ambiguous queries. Ambiguous queries can be categorized into four types: i) disorderly, ii) incomplete, iii) incorrect, and iv) superfluous queries. Below are examples of perfect and ambiguous queries and the ranked search results from Infoseek at http://www.infoseek.com/ for the book "Intelligent multimedia information retrieval," edited by Mark T. Maybury (Maybury, 1997).

- Perfect query: *Intelligent multimedia information retrieval*
  1. Intelligent multimedia information retrieval

- Disorderly query: *Multimedia information intelligent retrieval*
  1. Artificial intelligence, fuzzy logic and neural networks
  2. Intelligent access to information: research in natural language, information retrieval, computer vision, multimedia and database
  3. Multimedia color PC notebooks
  4. Intelligent multimedia information retrieval

- Incomplete query: *Multimedia information retrieval*
  1. Abstract Stein Mulleller Thiel 95
  2. Corpora Oct 1998 to -: Corpora: TWLT 14: language technology in multimedia information
  3. 3 2.1 Introduction to the workplan
     ...
  6. Intelligent multimedia information retrieval

Table 2: Major experimental Web search engines.

No.	Name	URL	Method
1	Clever	http://www.almaden.ibm.com/cs/k53/clever.html	hyperlink
2	Grouper	http://longinus.cs.washington.edu/grouper2.html	clustering
3	HuskySearch	http://huskysearch.cs.washington.edu/	metasearch
4	ImageRover	http://www.cs.bu.edu/groups/ivc/ImageRover/Home.html	image
5	ImageScape	http://skynet.liacs.nl/	image
6	Inquirus	http://www.neci.nj.nec.com/homepages/lawrence /inquirus.html	metasearch
7	Mercator	http://www.ctr.columbia.edu/metaseek/	image
8	MetaSEEk	http://www.research.compaq.com/SRC/mercator/	crawler
9	PicToSeek	http://zomax.wins.uva.nl:5345/ret_user/	image
10	W3QS	http://www.cs.technion.ac.il/~konop/w3qs.html	SQL
11	WebOQL	http://www.cs.tornoto.edu/~gus/weboql/	Object SQL
12	WebSQL	http://www.cs.toronto.edu/~websql/	SQL

- Incorrect query: *Intelligent multi-media information retrieval*
  1. Artificial intelligence research laboratory at Iowa State University
  2. Vasant Honavar's home in cyberspace
  3. CIIR multi-media indexing
  ...
  31. Intelligent multimedia information retrieval

- Superfluous query: *Intelligent multimedia information retrieval systems*
  1. Research in multimedia and multimodal parsing and generation
  2. Intelligent multimedia information retrieval

This example shows that even a slight variation in the query produces significant differences among the search results. Users tend to submit ambiguous queries to search engines, most of which use the technology of keyword matching to look for the desired pages. The ambiguity creates undesired search results if keyword matching is used.

Since the introduction of eXtensible Markup Language (XML) (XML, 2002), more and more Web documents are published in XML. An XML document not only provides the same information such as keywords, hyperlinks, descriptions, etc., that a function-like HTML document supplies, but also structural information. The structural information is one of the most crucial features of an XML document, and is not supplied by an HTML document. XML document searches (Hu et al., 2001) are expected to be the next major research direction for Web search engines as using this structural information is likely to give much better search results.

REFERENCES

Aggarwal, C. C., F. Al-Garawi, and P. S. Yu (2001). Intelligent crawling on the World Wide Web with arbitrary predicates. In *Proceedings of the 10th International World Wide Web Conference*, Hong Kong.

Arocena, G. O., and A. O. Mendelzon (1998). WebOQL: Restructuring documents, databases and Webs. In *Proceedings of the 14th International Conference on Data Engineering*, Orlando, Florida.

Ask Jeeves. *http://www.askjeeves.com/*

Baeza-Yates, R. A. (1992). Introduction to data structures and algorithms related to information retrieval. In W. B. Frakes and R. A. Baeza-Yates, editors, *Information Retrieval Data Structures & Algorithms*, pages 13-27, Prentice-Hall.

Benitez, A. B., M. Beigi, and S.-F. Chang (1998). Using relevance feedback in content-based image metasearch. *IEEE Internet Computing*, 2(4):59-69.

Bharat, K., and M. Henzinger (1998). Improved algorithms for topic distillation in a hyperlinked environment. In *Proceedings of the 21st International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 104-111.

Borodin, A., G. O. Roberts, J. S. Rosenthal, and P. Tsaparas (2001). Finding authorities and hubs from link structures on the World Wide Web. In *Proceedings of the 10th International World Wide Web Conference*, Hong Kong.

Brin, S., and L. Page (1998). The anatomy of a large-scale hypertextual Web search engine. *Computer Networks and ISDN Systems*, 30:107-117.

Carriere, J., and R. Kazman (1997). WebQuery: Searching and visualizing the Web through connectivity. *Computer Networks and ISDN Systems*, 29(11):1257-1267.

Chakrabarti, S., B. E. Dom, S. R. Kumar, P. Raghavan, S. Rajagopalan, A. Tomkins, D. Gibson, and J. Kleinberg (1999). Mining the Web's link structure. *IEEE Computer*, 32(8):60-67.

Chang, S.-K., and A. Hsu (1992). Image information systems: Where do we go from here? *IEEE Transactions on Knowledge and Data Engineering*, *Special Issue Celebrating the 40th Anniversary of the Computer Society*, 4(5):431-442.

Chang, C.-H., and C.-C. Hsu (1999). Enabling concept-based relevance feedback for information retrieval on the WWW. *IEEE Transactions on Knowledge and Data Engineering*, 11(4):595-609.

Cho, J., H. Garcia-Molina, and L. Page (1998). Efficient crawling through URL ordering. In *Proceedings of the 7th World Wide Web Conference*, Brisbane, Australia.

Dean, J., and M. R. Henzinger (1999). Finding Related Web Pages in the World Wide Web. In *Proceedings of the 8th International World Wide Web Conference*, pages 389-401, Toronto, Canada.

Dreilinger, D., and A. E. Howe (1997). Experiences with selecting search engines using metasearch. *ACM Transactions on Information Systems*, 15(3):195-222.

Dwork, C., R. Kumar, M. Naor, and D. Sivakumar (2001). Rank aggregation methods for the Web. In *Proceedings of the 10th International World Wide Web Conference*, Hong Kong.

Edwards, J., K. McCurley, and J. Tomlin (2001). An adaptive model for optimizing performance of an incremental Web crawler. In *Proceedings of the 10th International World Wide Web Conference*, Hong Kong.

Flickner, M., *et al.* (1995). Query by image and video content: The QBIC system. *IEEE Computer*, 28(9):23-32.

Florescu, D., A. Levy, and A. Mendelzon (1998). Database techniques for the World Wide Web: A survey. *ACM SIGMOD Record*, 27(3):59-74.

Foote, J. (1999). An overview of audio information retrieval. *Multimedia Systems*, 7(1):2-10.

Garofalakis, J., P. Kappos, and D. Mourloukos (1999). Web site optimization using page popularity. *IEEE Internet Computing*, 3(4):22-29.

Gevers, T., and A. Smeulders (1999). The PicToSeek WWW image search system. In *Proceedings of the IEEE International Conference on Multimedia Computing and Systems*, Florence, Italy, pages 264-269.

Gravano, L., K. Chang, H. Garcia-Molina, C. Lagoze, and A. Paepcke (1997). STARTS: Stanford protocol proposal for internet retrieval and search. In *Proceedings of the ACM SIGMOD International Conference on Management of Data*.

Greenberg, I., and L. Garber (1999). Searching for new search technologies. *IEEE Computer*, 32(8):4-11.

Gudivada, K. N., V. V. Raghavan, W. I. Grosky, and R. Kasanagottu (1997). Information retrieval on the World Wide Web. *IEEE Internet Computing*, 1(5):58-68.

Heydon, A., and M. Najork (1999). Mercator: A scalable, extensible Web crawler. *World Wide Web*, 2(4):219-229.

Howe, A. E., and D. Dreilinger (1997). SavvySearch: A meta-search engine that learns which search engines to query. *AI Magazine*, 18(2).

Hu, W.-C., Y. Zhong, W.-C. Lin, and J.-F. Chen (2001). An XML World Wide Web search engine using approximate structural matching. In *Proceedings of the 5th World Multi-Conference on Systemics, Cybernetics and Informatics*, Orlando, Florida, July 22-25.

Huang, L., M. Hemmje, and E. J. Neuhold (2000). ADMIRE: An adaptive data model for meta search engines. *Computer Networks* (*The International Journal of Computer and*

Telecommunications Networking), 33(1-6):431-448.
Kleinberg, J. M (1999). Authoritative sources in a hyperlinked environment. *JACM*, 46(5):604-632.
Konopnicki, D., and O. Shmueli (1998). Information gathering in the World Wide Web: The W3QL query language and the W3QS system. *ACM Transactions on Database Systems*, 23(4):369-410.
Lawrence, S., and C. L. Giles (1998a). Context and page analysis for improved Web search. *IEEE Internet Computing*, 2(4):38-46.
Lawrence, S., and C. L. Giles (1998b). Searching the World Wide Web. *Science*, 280:98-100.
Lawrence, S., and C. L. Giles (1999a). Accessibility of information on the Web. *Nature*, 400:107-109.
Lawrence, S., and C. L. Giles (1999b). Searching the Web: General and scientific information access. *IEEE Communications*, 37(1):116-122.
Lew, M. S. (2000). Next generation Web searches for visual content. *IEEE Computer*, 33(11):46-53.
Li, Y. (1998). Toward a qualitative search engine. *IEEE Internet Computing*, 2(4):24-29.
Lu, H., and L. Feng (1998). Integrating database and World Wide Web technologies. *World Wide Web*, 1(2):73-86.
Maybury, M. T. (1997). *Intelligent multimedia information retrieval*. MIT Press.
Mendelzon, A. O., and T. Milo (1998). Formal models of Web queries. *Information Systems*, 23(8):615-637.
Mendelzon, A. O., G. Mihaila, and T. Milo (1997). Querying the World Wide Web. *International Journal on Digital Libraries*, 1(1):54-67.
Miller, R. C., and K. Bharat (1998). SPHINX: A framework for creating personal, site-specific Web crawlers. In *Proceedings of the 7th International World Wide Web Conference*, Brisbane, Australia.
Najork, M. A., and J. Wiener (2001). Breadth-first search crawling yields high-quality pages. In *Proceedings of the 10th International World Wide Web Conference*, pages 114-118, Hong Kong.
Nick, Z. Z., and P. Themis (2001). Web search using a genetic algorithm. *IEEE Internet Computing*, 5(2):18-26.
Rafiei, D., and A. O. Mendelzon (2000). What is this page known for? Computing Web page reputations. In *Proceedings of the 9th International World Wide Web Conference* Amsterdam, Netherlands.
Search Engine Watch. *http://www.searchenginewatch.com/*
Selberg, E., and O. Etzioni (1997). The MetaCrawler architecture for resource aggregation on the Web. *IEEE Expert*, 12(1):8-14.
Singhal, A., and M. Kaszkiel (2001). A case study in Web search using TREC algorithms. In *Proceedings of the 10th International World Wide Web Conference*, pages 708-716, Hong Kong.
Smith, J. R., and S.-F. Chang (1997). An image and video search engine for the World-Wide Web. In *Proceedings of the Symposium on Electronic Imaging: Science and Technology¾ðStorage and Retrieval for Image and Video Databases V, IS&T/SPIE*, San Jose, California.
Spertus, E., and L. A. Stein (2000). Squeal: A structured query language for the Web. In *Proceedings of the 9th International World Wide Web Conference*, Amsterdam, Netherlands.
Sugiura, A., and O. Etzioni (2000). Query routing for Web search engines: Architecture and

experiments. In *Proceedings of the 9th International World Wide Web Conference*, Amsterdam, Netherlands.

Taycher, L., M. L. Cascia, and S. Sclaroff (1997). Image digestion and relevance feedback in the ImageRover WWW search engine. In *Proceedings of the International Conference on Visual Information*, San Diego.

XML (eXtensible Markup Language). *http://www.w3c.org/XML/*

Wen, J.-R., J.-Y. Nie, and H.-J. Zhang (2001). Clustering user queries of a search engine. In *Proceedings of the 10th International World Wide Web Conference*, Hong Kong.

Yoshitaka, A., and T. Ichikawa (1999). A survey on content-based retrieval for multimedia databases. *IEEE Transactions on Knowledge and Data Engineering*, 11(1):81-93.

Yuwono, B., and D. L. Lee (1996). WISE: A World Wide Web resource database system. *IEEE Transactions on Knowledge and Data Engineering*, 8(4):548-554.

Zamir, O., and O. Etzioni (1998). Web document clustering: A feasibility demonstration. In *Proceedings of the 19th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 46-54, Melbourne.

Zhang, D., and Y. Dong (2000). An efficient algorithm to rank Web resources. In *Proceedings of the 9th International World Wide Web Conference*, Amsterdam, Netherlands.