Dent Clin N Am 46 (2002) 435–462

THE DENTAL
CLINICS OF
NORTH
AMERICA

Information retrieval on the Internet: a guide for dental practitioners

Heiko Spallek, DMD, PhD

Center for Dental Informatics, University of Pittsburgh School of Dental Medicine,
3501 Terrace Street, Pittsburgh, PA 15261 USA

The structure of the Internet

It may be fun to surf the Internet and explore this information space randomly, but there is value in structure when searching for professional information. The Internet has a high degree of entropy—the lack of structure that occurs when energy dissipates from a system. When individuals apply their energy, for instance by selecting and filtering information, the Internet fosters self-organization [1]. The main reason why the Internet has so little structure and is so hard to decipher for many people, including dental professionals, is related to its origin and underlying design concept. Vannevar Bush, Director of the Office of Scientific Research and Development during the World War II, developed the conceptual idea for the modern World Wide Web in 1945. Bush discussed in his seminal Atlantic Monthly article [2] a system that he called "memex." Memex stores information for easy retrieval by humans. This system organizes information not by storing it by alphabetical order, location, or a hierarchy of topics, as most libraries, encyclopedias, or dictionaries do, but by "associative indexing" based on the content of the work. Bush argues that the memex storage system is much more closely related to the way we think than any artificial hierarchical organization system. His article is appropriately entitled "As we may think."

In 1989, Berners-Lee used Bush's ideas for developing a system that he called "mesh" and, in a revised version in 1990, "World Wide Web" [3]. The Web's structure is created by links, which are usually cross-references rather than a hierarchy (what we currently know as "links," Vannevar Bush called "associative trails"). The Internet is well suited to showing that things are related but not how they are related. This is in contrast to the traditional

E-mail address: hspallek@pitt.edu (H. Spallek).

PII: S0011-8532(02)00002-2

information retrieval world, in which providers, such as the National Library of Medicine, supply highly organized access to their information, such as on MEDLINE [4].

Information overload has become one of the most significant problems that users face on the Web; a plethora of information is created by persons who want to act as publishers. Two University of California–Berkley professors say that the Web contains 2.5 billion pages and grows at 7.3 million pages per day. If intranets and database-generated pages are included, the Web contains 550 billion documents (with 95% publicly accessible) [5].

Considering this immense amount of information, why are searchers in the Internet not always happy to have access to so much information? To analyze the discrepancy between available information and satisfaction, let us turn to the information needs of users.

Information needs

Ordinary users

Ordinary users look up all kinds of information, including topics related to dentistry. Although no data are available specifically about searching for dental topics, 40.9 million adults, or 54% of Internet users, use the Web for health care, according to a 2000 study by Cyber Dialog [6]. Users of Web sites depend highly on search engines to find the information for which they are looking.

Health care professionals

A contemporary, comprehensive store of medical knowledge is not a luxury for practicing doctors; it is as vital as an efficient pathology laboratory [7]. Wyatt reports about information needs of physicians in clinics and finds that one third of questions raised during their work remain unresolved because of lack of time or the inconvenience or costs of securing the answer [7]. The Web is a powerful new way to deliver online clinical information, but several problems limit its value to health care professionals. Content is highly distributed and difficult to find, clinical information is not separated from nonclinical information, and the current Web technology is unable to support advanced retrieval capabilities [4]. Studies have shown that health care providers rely heavily on fast and accurate information retrieval. One study showed that 65% of surveyed family practitioners use the Web in their practice; however, despite the plethora of medical information available, only 55% of their searches on the Web resulted in useful information [8]. Another study found that only 10% of searches retrieved relevant information [9]. Yet another study discovered that most searches retrieve only one fourth to one half of the relevant articles on a given topic [10]. Although the preponderance of high-quality clinical Web sites grows daily, there is still a high proportion of nonuseful information on the Web [11].

People who search the Internet are looking for answers to their questions. We can call this process "searching for information" or, more academically, "information retrieval." The following sections cover the basic concepts of information retrieval to help people understand the usage of these concepts by Internet search engines.

Information versus knowledge

Do we know more when we use the Internet because our access to information is improved? To answer this basic question we must define the differences between information and knowledge. Following Nonaka [12] and Huber [13], knowledge is the justified belief that increases an entity's capacity for effective action. Information is messages or meanings that add, restructure, or change knowledge. Information is only the raw material for the production of knowledge. Commonly, knowledge and information are used interchangeably, but knowledge entails a knower and information is self-sufficient. We also could say that knowledge lies more in people than in databases. In conclusion, the Internet in its current appearance only can provide information, not knowledge. This access to information can enhance vastly our ability to accumulate knowledge, however, especially considering that dentistry relies on an aggregation of different sciences, each of which has a huge store of continually changing meanings, beliefs, and advances. Accessing these meanings, beliefs, and advances usually requires an information retrieval system.

What is information retrieval?

The ultimate goal of using an information retrieval system is to retrieve documents that contain information. For users to retrieve documents, they must enter queries that request documents. Most information retrieval system queries consist of typing text at a keyboard, but this might change in the future to voice input. For queries to be matched to documents, there must be an indexing language, which is a set of descriptors that describe the content of the documents and can be entered by users to retrieve them. A search engine is a computer program that uses the indexing language to match queries and documents for users [14]. A search can take place only when documents were first indexed in a process performed either by a human or by an automated process.

What is knowledge management?

Knowledge management is designed to capture documentary, personal, and other types of knowledge and make it available in ways that can help an organization accomplish its goals. Organization and management guru Peter Drucker first used the term "knowledge management" in the mid-1980s. The concept was not translated into commercial computer technology for almost a decade, however [15].

Knowledge management is not a single technology but instead is a collection of indexing, classifying, and information-retrieval technologies coupled with methodologies designed to achieve results desired by the user [15]. Because the Internet mainly provides information instead of knowledge, this article focuses on the information retrieval process.

Effectiveness of information retrieval systems

Although searchers easily can find documents that contain a specific set of keywords, searchers have a notoriously difficult time in formulating search queries that find documents on specific topics effectively. The difficulty lies in finding all relevant documents (avoiding false negatives) without including documents that are not in the topic (avoiding false positives) [16]. To judge the effectiveness of an information retrieval system, recall and precision can be evaluated. Recall is the proportion of relevant documents in a collection that are retrieved (sensitivity); precision is the proportion of retrieved documents in a search that are relevant (positive predictive value) [14]. In other words, precision is a measure of how well a search eliminates unwanted results. The balance between precision and recall is what makes search engines efficient.

Basic concepts for information retrieval on the Internet

Traditional Internet search engines

Search engines, such as AltaVista and HotBot, traditionally consist of three components: the crawler, the index, and the search software. Crawlers, also called spiders, are programs that automatically scan Web sites and create indices of the Uniform Resource Locators (URL; the Web address that usually starts with http://www), the keywords if provided on these scanned Web pages, the links to other Web pages, and the text of the Web pages. Indexing in this context is the process of developing a document representation by assigning content descriptors or terms to the document to identify it as exactly as possible [17]. Crawlers follow all links to find other relevant pages and roam—in theory—the entire Internet. They return to sites periodically to look for changes. When a user submits a search query by providing a keyword, the search engine's software goes through the index to find Web pages with keyword matches and ranks the pages in terms of relevance [18]. In general, search engines retrieve too many Web pages, of which only a small fraction are relevant to the searcher's query. The most relevant documents do not necessarily appear at the top of the result page. The number of documents retrieved should not be used as the sole criterion for assessing the usefulness of a search engine. The retrieval of fewer but more relevant Web pages is often more helpful than listings of hundreds of general documents [19].

Example for a simple search

Problem

Which symptoms characterize periodontitis?

Search engine used

Excite

URL of the search engine

http://www.excite.com/

Query syntax

"periodontitis"

Result description

Excite returns 810 results—too many to review them carefully. The first one is already disappointing because it does not seem to come from a professional resource. It is a "Cockatoo Press Web site" that talks about periodontitis but "is maintained by a non-native speaker of the English language" whose name and credentials are unknown to the reader. Cockatoo is actually a travel guide Web site. The second-highest ranked result links to a bad-breath page. Although bad breath might be one symptom of a periodontitis, this resource does little more than pitch a new kind of mouth rinse that can be ordered via the Web site.

URL
http://www.cockatoo.com/periodontitis/index.htm;
http://www.badbreath-infocenter.com/

Tips to improve the search effectiveness

- Be as specific as possible
- Check spelling
- Use proper capitalization
- Try using synonyms or related words
- Avoid computer-specific terms such as file or disk
- Try to add more keywords if the search returns too many results

Catalogs

Catalogs, such as Yahoo!, work with descriptions of Web pages submitted by Web site owners or by Yahoo!'s own editors, who have reviewed the pages manually. Because catalogs do not use crawlers, they do not automatically find Web pages. Human-generated descriptions usually produce a more relevant response to searches, however, because the editorial process of creating these descriptions is more accurate than the computer-generated index of traditional search engines.

Catalogs can be browsed following hierarchically organized categories, such as Health->Medicine->Dentistry->Periodontics, or they can be searched by keyword in the same way as traditional search engines. The catalog engine tries to match the search term with the resource description in the database.

More and more search engines, such as AltaVista and Excite, take a hybrid approach by also using browsable catalogs. Yahoo!'s search results, for instance, show the positions of matches within the Yahoo! hierarchy. This positioning allows searchers to find related categories and higher-level concepts that might even fit the information need better [18].

In general, manually produced catalogs are more outdated than traditional search engines because humans work slower than the search engines' crawlers. New Web pages might not be found in a catalog, or changes to existing pages might not have been updated yet.

Keyword search

There are two basic types of search methods: one tries to match the searcher's term with one or more words in the text of the Web pages to be searched, and the other method tries to match the same term with information about the Web pages to be searched. The information about the documents is called metadata and must be chosen carefully by the creator of the Web pages—usually the Web designer—or by the entity creating a document collection, such as an Internet search engine operator. For instance, a patient education Web page about tooth brushing might not include the term "prevention"; thus, a dentist searching for "oral prevention measures" might never find it. A careful Web designer would include the phrase "oral prevention measures" in the metadata of this document. These metadata additions are called metatags; they are part of HTML, the hypertext markup language, which is used to design Web pages. Unlike most HTML tags, however, metatags do not affect a document's appearance. Instead, they include such information as a Web page's contents and some relevant keywords behind the scenes. Most engines look for keywords in metatags first for their indexing algorithms. In the past, some Web designers have subverted keyword-based techniques by loading their metatags with keywords that do not relate to their site's content. Search engine operators have taken steps to counteract this, however. For example, some search engines lower the rankings of sites that use keywords unrelated to their content [18].

Example for a catalog search

Problem

How does localized juvenile periodontitis fit into the classification of periodontal diseases?

Search engine used

Yahoo! Catalog

URL of the search engine

http://www.yahoo.com/

Query syntax

Yahoo->Medicine->Dentistry->Periodontics

Result description

After reaching the appropriate category "Periodontics" from Yahoo!'s home page, the first listing is the American Academy of Periodontology, which is described as "offers information on gum disease and treatment and a periodontist referral service." Following the link, a second search is required on the American Academy of Periodontology's homepage for "Localized Juvenile Periodontitis." Ultimately, a full-text article entitled "Development of a Classification System for Periodontal Diseases and Conditions" can be found, which summarizes the classification for periodontal diseases and conditions.

URL
http://www.perio.org/

Tips to improve the search effectiveness

- Use catalog for acquiring a general overview about a topic
- Follow the category tree to find the subcategory of interest
- Follow promising links and continue with a site search to find specific information

Usually searchers enter a keyword (or keywords along with Boolean modifiers, such as "and," "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 searcher, the engine uses an algorithm to rank the pages that contain the keyword. For example, the engine may count the number of times the keywords appears on a page.

Advanced concepts for information retrieval on the Internet

Special-purpose search engines and indices

As opposed to general search engines, domain-specific search engines allow searching of subsets of the Internet related to one or more topics [19]. Various special purpose search engines and knowledge stores for health

Example for Boolean keyword search

Problem

Can I use Amoxicillin for treating a case of localized juvenile periodontitis?

Search engine used

AltaVista: advanced search

URL of the search engine

http://www.altavista.com/

Query syntax

“amoxicillin AND ‘localized juvenile periodontitis’ ”

Result description

A Boolean operator was used to combine the word “amoxicillin” logically with the phrase “localized juvenile periodontitis.” This resulted in 11 links to pages that all contained the word “amoxicillin” and the exact phrase “localized juvenile periodontitis.” Already the first result, an article by Ramin Mollazadegan, Faculty of Odontology, Karolinska Institute, Sweden, fits the information need.

URL
http://www.dsg.ki.se/odonfak/kov/exarb/Ramin_Mollazad.html

Tips to improve the search effectiveness

- Structure carefully the complex Boolean query
- Review help of the search engine to see supported Boolean operators and capitalization rules
- Use an asterix (*) to indicate missing letters if one is uncertain about spelling or one wants to cover more meanings, such as “dent*” for “dental” and “dentistry” (some search engines use a % or a ? for one missing character)
- Try using synonyms or related words

Example for search in a special-purpose search engine

Problem

Where do I find treatment guidelines for localized juvenile periodontitis provided by periodontal societies?

Search engine used

HealthWeb→Dentistry

URL of the search engine

http://www.healthweb.org/

Query syntax

"periodontology"

Result description

Although HealthWeb was unable to find any result when searching with a specific query "treatment and guidelines and periodontitis," it provided eight links to national and international periodontal societies when searching for the broader term "periodontology."

URL
URLs of eight periodontal societies

Tips to improve the search effectiveness

- Use broad terms
- Try using synonyms or related words
- Use for concept/topic search rather than for specific terms

care professionals have been developed in the past and have either disappeared or survived [4,20,21]. If compiled with effort and accuracy, they have the potential to shorten the search for relevant information. These subsets are usually human-made, however, and they lag behind the rapidly changing Internet. Searchers also must trust the creator of the particular subset that important Web sites have not been omitted.

Special-purpose search engines differ widely in their goals, technical sophistication, completeness, and usability. The MegaSite project compares the largest health-related Web resources thoroughly [22]. Some examples of special purpose search engines are listed in Table 1.

Table 2 shows a list of commonly used search engines ordered by category. Considering that more than 3500 search engines, link list, and business

Table 1
Special purpose search engines and indices for dentistry

Dental image databases	
Oral pathology image database http://www.uiowa.edu/~oprm/AtlasHome.html	Online oral pathology atlas: histologic and clinical images and definitions eg, four pictures of a periodontal abscess (two clinical, two histologic)
DERWeb→image library http://www.derweb.ac.uk/	Registration (free) required Total of 2600 clinical dental images with clinical description Boolean search for images thumbnail pictures are watermarked; purchase necessary before use eg, "periodontitis" results in 22 images
Consumer health information	
HealthFinder http://www.healthfinder.gov/	Provides consumer health information for the US population in many languages, but mostly in English and Spanish Includes more than 1500 consumer health topics
MEDLINEplus health information→Dental Health http://www.nlm.nih.gov/medlineplus/dentalhealth.html	Consumer-oriented directory of dental Web sites Prior editorial review limits listings to appropriate Web sites Large number of Frequently Asked Questions documents listed for various topics
General dental resources	
Internet dental resources http://amia.dental.pitt.edu/resources/	Categorized listing of dental Web sites, eg, case studies, dental colleges Comprehensive listing of dental mailing lists including short description
Dental-related internet resources http://www.dental-resources.com/	Directory of Internet resources related to dentistry, eg, continuing education, dental laboratories, education sites Well structured for browsing, but no search available Slightly outdated
HealthWeb→Dentistry http://healthweb.org/dentistry/	Collaborative project of health sciences libraries and National Library of Medicine all listings are reviewed by editorial board Supports wildcards and Boolean operators Uses medical subject headings keywords (see MEDLINE)
Dental meta search engine	
MedNet's dental databases http://www.mednets.com/sdental.htm	Access to 25 searchable Web sites, eg, dental implant glossary, ADA Web site search engine Various Web site listings by category, eg, clinical information, diseases, dental hygenists

Table 1 (*continued*)

Medical search engines relevant to dentistry	
Medline (NLM's premium bibliographic database) http://www.ncbi.nlm.nih.gov/entrez/	PubMed allows access to 11 million citations from Medline dated back to 1965 Medical subject headings allow one to search by a controlled vocabulary Tutorial for learning how to use PubMed most efficiently eg, 12,068 results for keywords search "periodontitis"
Medical Matrix http://www.medmatrix.org/	Registration (free) required eg, 14 sites and 42 links for "periodontitis" search: mainly links to guidelines and treatment standard documents

directories are operating on the Internet, the list is far from complete; however, it provides a starting point for the information searching professional.

Ranking and matching criteria

One of the main problems of Internet search engines is their inability to limit the search results to an appropriate number of relevant documents. To include the relevant results, almost all search engines offer infamously too many search results, of which a few may be relevant to the searcher's interest. As long as the most relevant documents appear on top of the result page, this does not represent a problem for the efficiency of the information retrieval. In most cases, however, the relevant search results are interspersed among up to thousands of search results. Although search engines try to improve this situation, users of search engines continuously experience the weaknesses of relevancy ranking. To solve this problem, search engine experts must struggle with using a computer algorithm to approximate the complex human value of something being relevant to a person's interest [23].

Although the development of useful intelligent agents that adapt to the individual user's preferences has just started, significant research efforts were undertaken to measure the relevance of Web pages toward a larger user group, such as all Web site creators. Although intrinsically insufficient for the individual searcher, these group-based "judgments" about relevance are a first step toward a useful ranking of Internet search engine results. The search engine Google, for instance, calculates the popularity score for a particular site by totaling the sites that contain links to that site. High link popularity leads to an improved ranking [23]. The strength of Google's PageRank feature, which is named after Larry Page, a former PhD student at Stanford who created Google with Sergy Brin, another former Stanford graduate student, is the method by which it automatically extracts people's editorial judgment to improve relevancy [24].

Table 2
Internet search engine list (description omits category-specific characteristics, such as eliminating duplicates by meta search engines)

General search engines	
AltaVista http://www.altavista.com/	Language- and date-specific Boolean queries supported Document-translation service embedded People finder and yellow pages
AllTheWeb http://www.alltheweb.com/	Uncluttered user interface Specialized searches for pictures, audio and video Displays for each search the results of available videos and pictures as well eg, finds 13,888 results for "periodontitis" (and 54 video clips)
Google http://www.google.com/	Uncluttered user interfac every fast and easy to use Results well organized by relevancy Useful extras: cached copies of result pages in case actual page cannot be reached, translation service, similar pages feature No distracting banner advertisement, but color-coded sponsored links eg, finds 20,300 results for "periodontitis"
Lycos http://www.go.com/	Portal-style search engine Provides additional features, such as parental control, personalized home page, top 50 of search terms Search results listings interrupted multiple times by "featured listings" which are advertisements eg, finds 13,857 results for "periodontitis"
NorthernLight http://www.northernlight.com/	Several specialized searches, such as business search alert feature for changed search results Results are ordered by ranking and organized into custom search folders 15 million articles from trade journalseg, finds 12,938 results for "periodontitis" organized in 25 folders, eg, Actinobacillus, Diabetes
Catalog-based retrieval systems	
Open Directory Project http://dmoz.org/	Human-edited directory of Web sites Volunteer editors provide a brief description of each directory entry Searches are limited to the information in its own database eg, 535 entries in the dentistry category
Yahoo! http://www.yahoo.com/	Hierarchical organized directory system Search engine powered by Google provides results in different categories Provides country- and language specific Yahoo!s Acts as portal with vast range of feature, such as email account, picture upload, auctions, etc. eg, 342 entries for dentistry

Table 2 (*continued*)

News group search	
Google Groups (formerly DejaNew) http://groups.google.com/	In beta version available Advanced search allows date-, group-, author- and language-specific searches eg, finds 1,220 posting for "periodontitis"
Meta search	
Ixquick Metasearch http://www.ixquick.com/	Supports natural language, Boolean, wildcard queries Ranking based on top ten results of each queried search engine Plain interface Specialized searches for news, MP3, pictures
MetaCrawler http://www.metacrawler.com/	Cluttered user interface Search results can only be accessed after scrolling (advertisement on page) "More like this" feature help to refine the search query Displays from which search engine the result originates
Media search	
Fast Multimedia Search http://www.multimedia.alltheweb.com/	Clean user interface WAP (wireless access protocol) access possible Thumbnail view of pictures on search result page eg, finds 279 pictures for "periodontitis"

The experimental search Clever, which was developed by IBM at its Almaden Research Center in San Jose, uses a slightly different approach to judge relevance based on links. The engine does not crawl the Web but uses indices built by other programs to discover useful Web sites. This initial result is compiled as a root set. Then, Clever looks for documents that link to and from the root results. Clever rates the Web page in the root set and the linked pages on the basis of how many other sites link to them. Pages that many Web site authors have chosen to link to are called "authorities" and are considered to be valuable sources of content. Web sites that link to many authorities are called "hubs" and are considered to be valuable reference tools. The technology is called hyperlink-induced topic search (HITS) [18,24,25].

The described concept of ranking by relevance has three major flaws. First, a group's judgment about relevance is applied to an individual's search interest, which might not necessarily correspond. Second, the assessment of relevance must not be mistaken for the Web page's quality. Third, the assumption that a hyperlink from page A to page B is a recommendation of page B is probably true in most cases, but not in all, such as on a Web page that criticizes the market behavior of Microsoft Corporation [26].

Example for search result ranking

Problem

Which different forms of localized periodontitis are known?

Search engine used

Google

URL of the search engine

http://www.google.com/

Query syntax

"localized periodontitis"

Result description

Google lists two links to localized juvenile periodontitis resources and continues with a link to a Web page about localized necrotizing ulcerative periodontitis. Additional links to pages about localized periodontal therapy are following. A link check finds that Google's first results are referenced by six other Web sites, which shows their popularity but not the validity of the content offered.

URL
http://jeffline.tju.edu/DHNet/cases/oralb/

Tips to improve the search effectiveness

- If the first few results are not fulfilling one's information need, the query might need to be rephrased
- Try using synonyms or related words

Searching by example

To overcome the problem of applying a group's judgment about relevance to an individual's information need, a technique called relevance feedback is used. This technique can improve recall dramatically by asking the system to find pages that are similar to the ones that match the searcher's needs. With this method the searcher only must recognize one item that is on the right track, rather than explicitly generating new search words. Several search engines use this concept by offering a link to "related pages" close to all search results listed by any given query. This feature also exists in MEDLINE, where it works well because of the high standardization of index terms.

Example for search by example

Problem

Which drugs can be used to treat localized juvenile periodontitis?

Search engine used

Google

URL of the search engine

http://www.google.com/

Query syntax

"drug susceptibility localized juvenile periodontitis"; similar pages feature

Result description

Whereas Google's first result is a Canadian Material Safety Data Sheet for infectious substances and the second result goes to a rather general article of the *Mount Sinai Journal of Medicine*, "Periodontal Disease: An Overview for Physicians," the third result is a link to a full-text article of the *Journal of Periodontology*, "Position Paper: Systemic Antibiotics in Periodontitis." Because this resource is an almost perfect match for the original problem, we click on the "similar pages" link. Google suggests a total of 24 resources, including the one used as sample resource. Unfortunately, all offered results go only to the home pages of Web sites and not to the actual pages about drug susceptibility.

URL
http://www.perio.org/resources-products/pdf/21-Antibiotics.pdf

Tip to improve the search effectiveness

- Search carefully for the best match before using the similar pages feature

Meta search engines

Meta search engines draw on the strength of many search engines. A meta search engine submits a searcher's query to many different search engines and then organizes and displays the results in a uniform format. Because coverage of the Internet by search engines is generally low, the combination of several of them increases the coverage, assuming that all search engines cover slightly different parts of the Internet. Meta search engines

Example for meta search engine

Problem

How do I treat a case of localized juvenile periodontitis?

Search engine used

MetaCrawler

URL of the search engine

http://www.metacrawler.com/

Query syntax

''localized juvenile periodontitis''

Result description

MetaCrawler queries three search engines: AltaVista, DirectHit, and Internet Keywords. Each displayed result is associated with the search engine where it was found. The first result for this search is a Web page of the Department of Microbiology, University of Pennsylvania School of Dental Medicine, which lists bibliographic information about articles published by Joseph DiRienzo, PhD. The page was updated the last time in 1998 and contains no information about the problem of treating juvenile periodontitis. The following results are similarly poorly fitting and even leave the field of dentistry, such as the fourth result, ''Localized B2B emarketplace.'' Only the ninth result links to a case report by dental group of periodontists. The resulting home page links to a case report entitled ''Treatment Approach to Localized Juvenile Periodontitis with Long-Term Follow-up.''

URL
http://www.parkaveperio.com/case/

Tips to improve the search effectiveness

- Disregard distracting advertisement
- Omit widely used terms in the query, such as ''treatment'' or ''diagnosis''
- Do not limit the review to the first result because of poor ranking capabilities
- Try using synonyms or related words

usually eliminate duplicates that are retrieved from different search engines. Meta search engines also provide another indicator for relevance. If several primary search engines provided the particular result, it might be relevant. In essence, meta search engines combine several other search engines and collate the results.

Advanced filtering

The simplest approach for filtering is to filter content by keywords. Most search engines do so by applying a searcher's query options to their existing database. An advanced approach of filtering is the use of collaborative-filtering techniques that mine user-access patterns, Web logs, preferences and profiles to tailor the content provided at specific sites [27]. One example is Direct Hit, which tracks users through Web searches and builds a popularity list based on the user's behavior on these sites. When clicking on a result at Direct Hit's search result page, searchers are detoured through an adjunct to the search engine, which captures and stores the searcher's choices. Gary Cullis, a former patent agent and a graduate of Harvard Law School, emphasizes that the efforts of searchers are actually a byproduct that traditional search engines are not capturing [24]. Privacy advocates argue, however, that capturing and storing users' choices violates privacy, even if the capturing is anonymously. Another concern is that users typically do not navigate deeply into the result set and do not add appropriately to the popularity list.

Northern Light uses another filtering technology. This approach tries to narrow the scope of queries to yield results that are more relevant. When submitting a keyword search, users can fill out electronic forms to specify, for example, that they want only information that relates to a certain industry or a certain geographic location [18].

Natural language

Search services are beginning to work with natural-language queries to make them easier to use. For example, with Ask Jeeves, instead of typing in one or more keywords, users who are looking for the current temperature in Boston would type "How is the current temperature in Boston?" The service would then refer them to a site that provides weather updates for Boston. Unlike other search engines that are almost entirely automated, Ask Jeeves requires the work of editors. To scale the workload of editors, Ask Jeeves solicits sponsoring of certain topics. For instance, the health channel on the Ask Jeeves site is sponsored by OnHealth.com (acquired by WebMD). OnHealth.com's primary goal for providing the labor of the editors is to drive traffic to OnHealth.com's Web site [28]. Ethical problems might arise from the fact that a sole sponsor edits health care resources—a fact that most searchers might not even be aware of.

Example for the use of filter options

Problem

Is reimbursement for surgical gingival curettage per quadrant (D4220) covered by any federal or state medical assistance program in the United States?

Search engine used

Northern Light's Power Search

URL of the search engine

http://www.northernlight.com/power.html

Query syntax

Search for "localized juvenile periodontitis." Limit subjects to health and medicine. Limit documents to government Web sites, documents written in English, Web sites from the United States. Select date range, start date: 01/01/2001. Sort results by date and time.

Result description

Northern Light offers two results: one is the Workers' Compensation Medical Diagnosis Codes, which do not fit our information need. The second one is a provider services manual about periodontitis by the Rhode Island Department of Human Services, which includes a description of treatments indicated in juvenile periodontitis cases.

URL
http://www.dhs.state.ri.us/dhs/heacre/provsvcs/manuals/dental/perio.htm

Tips to improve the search effectiveness

- Restrict the retrieval only as much as necessary
- Limit country-specific questions geographically, such as insurance coverage
- Use "save this search as an alert" for an automatic notification via e-mail in case the search result changes over time (free Northern Light search alert account necessary)

Example for natural language query

Problem

How do I treat my patient whose clinical diagnosis is localized juvenile periodontitis?

Search engine used

Ask Jeeves

URL of the search engine

http://www.ask.com/

Query syntax

"How do I treat a localized juvenile periodontitis?"

Result description

Ask Jeeves offers answers to a set of questions that are similar to the given one. The first question comes close to the original one: "Where can I learn about the dental or oral condition periodontitis?" It is much broader, however (links to drkoop.com's Medical Encyclopedia). The second question already misses the main idea: "Where can I learn about the diabetes complications of gum disease?" Ask Jeeves also offers a list of answers under the section "People with similar questions have found these sites relevant." The first result fits the problem best: it is a link to a clinical case study of localized juvenile periodontitis by Denise M. Bowen, RDH, MS, Idaho State University, and Jane Forrest, RDH, EdD, Director, National Center for Dental Hygiene Research. The case study is part of the National Center for Dental Hygiene Research Web site and is sponsored by Oral-B.

URL
http://jeffline.tju.edu/DHNet/cases/oralb/

Tips to improve the search effectiveness

- Check out section "people with similar questions…"
- Use for broader and more general content questions
- Use simple questions without subclauses

Searchbots

Unlike Internet search engines that run on central servers, searchbots are software applications that run on a searcher's computer. The task of the searchbot is to search comprehensively on the user's behalf by querying various information stores, such as Internet search engines, catalogs, newsgroup archives, and other topic-specific resources. One of the best-known examples is Copernic (currently, version 2001 as Basic version for free, Plus and Pro version for purchase at www.copernic.com) by Copernic Technologies, Inc. This searchbot retrieves information from the Web, newsgroups, and e-mail directories using sources such as AltaVista, Excite, Yahoo!, Infoseek, Lycos, Open Text, and Deja.com. The results are displayed in order of relevancy; duplicates are filtered out, and links to search results are validated. Copernic offers access to 93 categories and provides access to approximately 1000 search engines and directories. The resources can be accessed in real time, or search tasks can be scheduled to be performed at a later time. The searchbot can perform searches on a regular interval and send an e-mail notification if a search result has changed. More about searchbots can be found at www.agentland.com.

Picture search

There are several attempts to solve the problem that most Web searches are only targeted toward text [29,30]. Finding a certain picture on the Web is currently only possible by actually searching for text and then hoping for a picture associated with the textual information presented. Few engines allow a broad search for pictures beyond their own multimedia archive. Examples of more advanced picture search engines are the meta search engine ixquick (http://www.ixquick.com) and Google's Image Search (http://images.google.com).

Newsgroup search

If Web searches are unsuccessful, other information resources on the Internet may be useful. Reading what was posted on newsgroups and mailing lists is like checking out people's past conversations about any given topic. Newsgroups and discussion lists can provide valuable advice. Although the validity of these information resources is just as uncertain as that of Web sites, personal opinions can lead information searchers to the right resources or help them to ask the right questions. In the past, accessing newsgroups was a complicated technical challenge. Currently, all newsgroups and most mailing lists are available on the Web and can be accessed by using a regular Web browser. Some companies have specialized in archiving the content of all newsgroups online and making them searchable (eg, Dejanew.com, which was recently acquired by Google).

Example for using a searchbot

Problem

How do I treat a case of localized juvenile periodontitis?

Search engine

Search in Copernic 2001 Pro (http://www.copernic.com/)

Query syntax

"localized juvenile periodontitis"; search for exact phrase; remove broken links

Result description

Copernic found 37 results from different search engines, which were reduced to 29 results after removing the ones whose URLs were not valid anymore. A score allowed the searcher to see which documents were more relevant based on Copernic's evaluation. A single click on the results, which are displayed with a short description, shows a thumbnail overview of the page. The first result was a clinical case from jeffline. The second highest rank was an article by L.C. Bueno, et al entitled "Relationship Between Conversion of Localized Juvenile Periodontitis-Susceptible Children From Health to Disease and *Actinobacillus actinomycetemcomitans* Leukotoxin Promoter Structure" available as abstract only at the given URL. Overall, the remaining search results seemed to be well suited for the information need.

URL
http://jeffline.tju.edu/DHNet/cases/oralb/;
http://www.perio.org/jo urnal/abstracts/Sept98/998.html

Tips to improve the search effectiveness

- Must be purchased and installed on own computer and requires some time before using
- Use the advanced search features, such as removal of dead links, to be more efficient and automatic search updates and scheduling function
- Try using synonyms or related words

Example for picture search

Problem

Compare their own clinical findings with picture of a published case report.

Search engine used

Fast Multimedia Search

URL of the search engine

http://www.multimedia.alltheweb.com/pt

Query syntax

Search for images; ''localized juvenile periodontitis''

Result description

One image was found and displayed as thumbnail. Image-specific data are shown, such as image size, file size, image format, last modified, and text snippet surrounding the image. The result page provides a link directly to the image and a separate link to the Web page that contains the image. The resulting page is entitled ''Reconstructive Osseous Surgery'' by Dr. Paulo Camargo. It is part of an accredited continuing dental education course offered by the University of California, Los Angeles, Periodontics Information Center.

URL
http://www.dent.ucla.edu:81/pic/members/ros/ros.01.html

Tips to improve the search effectiveness

- Limit the image format to images one can display on the computer
- Use the advanced search feature if one wants to search, for instance, for an exact phrase
- Select images if one wants to omit video or other media in the result section
- Try using synonyms or related words

Example for search for personal conversation

Problem

What is the indication for surgery in cases of localized juvenile periodontitis?

Search engine used

Google's Groups

URL of the search engine

http://groups.google.com/

Query syntax

"localized juvenile periodontitis and surgery indication"

Result description

Google correctly recognizes that the related newsgroup is science->medicine->dentistry (sci.med.dentistry) and displays this at the top of the search result screen. The first posting found by Google was originally posted by a dentist to the newsgroup sci.med.dentistry in May 1997, and is a reply to another dentist's question.

The comprehensive two-page answer starts with

> What is the point of the surgery? To eliminate microbes? Will antibiotics do the same—eliminate microbes?

I do surgery primarily in one of two indications

...
(">" indicates the original question)
Although this is only a personal opinion, it still gives an idea of what other dentists think about the problem.

Tips to improve the search effectiveness

- Be specific when formulating the query
- Use the advanced group search feature to limit the search results, for instance, to more recent postings
- Try using synonyms or related words

Current problems

The Web lacks separation of practitioner-oriented and consumer-oriented information. This can be problematic in a domain such as health care, in which the source and quality of information are important. Whereas traditional database vendors, such as drug reference resources, focus on providing high-quality commercial information, the Web allows anyone to publish anything. Although this may be an advantage in political and other spheres, it can be a disadvantage in health care because practitioners and educators base decisions and education, respectively, on the highest-quality scientific information [4,31].

The complete coverage of all existing documents is another important factor that characterizes the quality of retrieval systems. Unfortunately, technical and political reasons restrict search engines from finding sites [23]. Most Internet search engines harness only a fraction of the indexable Web (less than 30%, according to one study [27,32]). Some later studies specify that no search engine indexes more than 16% of the indexable Web [33].

Whereas researchers try to learn more about the underlying algorithms of how search engines, catalogs, and agents work, the unscrupulous can use this knowledge to manipulate the ranking heuristics. Relevancy (or keyword) spamming lets Web page designers trick the algorithm into giving their pages a higher ranking. For example, ranking spammers often stuff keywords into invisible text and tiny text hidden from most Web users but visible to spiders, such as text brims with repeated instances of keywords, which elevates a site's ranking. This ranking warfare has created an impossible situation. Search engine operators do not fully and truthfully disclose the underlying algorithms that govern indexing, searching, and ranking because they fear that spammers will use this knowledge to trick them. Ethical Web page designers legitimately must know how to indicate relevancy to the ranking algorithm, however, so that their pages are listed in response to genuinely relevant searches.

Beyond the challenge of second-guessing ranking algorithms, there may yet be another, more certain method of getting results. Some Web site producers simply buy a higher ranking, despite the indignant protests of several major search engine representatives that they do not sell search positions. In a much-publicized move, however, Alta Vista and DoubleClick invited advertisers to bid for position in their top slots. Yahoo! sells prominence indirectly, allowing Web owners to pay for express indexing, which moves their pages ahead in the 6-month queue. Google calls the highly priced first spots on their search result page "premium sponsorship." Another method for buying prominence lets Web owners buy keywords that, when searched for, display the search results and the owners' banner ads. Amazon Books, for example, has a comprehensive arrangement of this type with Yahoo!, as does Barnes & Noble with Lycos [23].

The ethical question behind these trends is whether the market shall decide which information resources get listed in search engines by introducing a commercial bias.

Outlook

Information economics

Because everybody can be a publisher on the Web at almost no cost, the information available is growing exponentially. The amount of information that can be accessed by all users grows linearly at best, however, following an article by Coiera [34]. He suggests that the consequence of this ever-expanding information marketplace for information producers is that their success increasingly depends on their ability to compete for the attention of information consumers. Considering the growing amount of information that must be filtered and evaluated, information retrieval will become more and more time consuming. The importance of information retrieval systems grows for the information producer, such as Web site creators, and for information consumers.

Semantic Web

Semantic [Greek] means the understanding of language. The Web originally was built for human consumption, and although everything on it is machine-readable, these data are not machine-understandable [35]. Web search engines effectively retrieve entire documents, but they are imprecise because they do not exploit and retrieve the semantic Web document content. To address this problem, the World Wide Web Consortium (W3C), the leader in the technical evolution of the Web, developed a technical specification for a semantic Web, called resource description framework (RDF). This resource description framework allows us to retrieve more precise information because documents are manually structured, for instance, with the extensible markup language, which describes the data elements. Only the description of the data can help us to achieve a more flexible and precise knowledge representation and better information retrieval. The underlying extensible markup language as format allows resource providers to exchange information among different systems [36]. To illustrate this theoretical approach, here is an example: when using a semantic search, the search engine understands the meaning of the term in relation to other terms. For instance, a semantic search engine "knows" the double meaning of the term "root" and would ask a searcher who entered the search term "root" whether the intent is to search for a part of a plant or for a part of a tooth. Precondition for a semantic search is not only to add information about a document as metadata but also to give the search engine an understanding of how the terms relate to each other. For instance, a "root" is an essential part of a "tooth," whereas a "tooth" is a part of the "body," and so forth.

Intelligent agents

When we try to characterize an intelligent electronic agent, we use the following qualifiers: autonomous, goal-driven, reactive, social, customized, adaptive, mobile, strives to be believable. Currently available agents usually lack one or more of these qualifiers. For instance, almost no agents are social, which means they do not communicate well with other agents.

The use of intelligent agents for information retrieval could help to prepare a personalized view of the Web. This personalization happens behind the scenes after the agent has learned its user's preferences. Sophisticated agents can learn on their own by following the searcher's example. They can watch a searcher browse during a session, discover what the searcher might be interested in, and tailor their behavior accordingly [37].

InfoFinder agent

Systems such as InfoFinder learn profiles of user interests from sample documents that users submit while browsing, without surveying them as to their interest in a set of sample documents. This makes InfoFinder significantly easier to use than other preference-learning agents. InfoFinder learns general profiles from the documents by heuristically extracting phrases that are likely to represent the document's topic. InfoFinder's learning algorithm generates a search tree, which the system then translates into a Boolean search string for submission to a generic search engine [16].

WebWatcher project

WebWatcher, developed by Carnegie Mellon School of Computer Science, is a "tour guide" agent for the Web. Once a searcher tells it what kind of information he or she seeks, it accompanies the searcher from page to page as he or she browses the Web, highlighting hyperlinks that it believes are of interest. Its strategy for giving advice is learned from feedback from earlier tours [38].

Genetic algorithms

An intelligent agent usually collects documents from the Web based on the keywords provided by the searcher. Then it extracts data features, such as common words or phrases, and uses them as input to an algorithm for creating a searcher profile. The agent tries to discover Web pages similar to those generated from the profile by using search engines or tapping into a collection of Web pages that is based on other searchers' recommendations. The agent compares the newly discovered Web pages to the searcher's profile and presents the most relevant ones to the searcher. Finally, the agent uses the searcher's feedback to adapt the profile in an algorithm that is called genetic because branches of searches that were not relevant are extinct [39]. Webnaut, one prototype example of an intelligent software agent, collects information from the Internet and filters it according to a profile of searcher interest. Webnaut uses a genetic algorithm, which enables it to learn the searcher's interest and adapt as the searcher's interest changes over time. This learning process is driven by searcher feedback [39].

Summary

The Internet is not designed for efficient information retrieval. Searchers experience difficulties when trying to find information quickly. Once they find the information, they often cannot assess the validity of the information or its origin. Various Internet search engines and information retrieval systems try to facilitate this process to make the searching more efficient. Searchers use these services, which are usually free, to satisfy their information needs. Knowing when to use which search engine and how to use it is crucial to finding the right information in a timely fashion.

Acknowledgments

The author gratefully acknowledges the contributions of Titus Schleyer, DMD, PhD, School of Dental Medicine, University of Pittsburgh. The author also would like to thank Ms. Kelly Anne Dolan, Temple University Writing Center, for her thoughtful review of the manuscript.

References

[1] Dyson E. Release 2.1: a design for living in the digital age. New York: Broadway Books; 1998.

[2] Bush V. As we may think. July 1945. Available at: http://www.theatlantic.com/unbound/flashbks/computer/bushf.htm. Accessed on: July 9, 2002.

[3] Berners-Lee T. Information management: a proposal. Available at: http://www.w3.org/History/1989/proposal.html.

[4] Hersh WR, Brown KE, Donohoe LC, Campbell EM, Horacek AE. CliniWeb: managing clinical information on the world wide web. J Am Med Inform Assoc 1996;3:273–80.

[5] Lyman P, Varian HR, Dunn J, Strygin A, Swearingen K. How much information? Available at: http://www.sims.berkeley.edu/research/projects/how-much-info/. Accessed on: July 9, 2002.

[6] Survey NI. Cyber dialogue: the future's bright for online health industry. Available at: http://www.nua.ie/surveys/index.cgi?f=VS&art_id=905355995. Accessed on: July 9, 2002.

[7] Wyatt J. Use and sources of medical knowledge. Lancet 1991;338:1368–73.

[8] Tuominen K, Crouse BJ. Use of the world wide web by family practitioners. Minnesota Medicine 1996;79:43–6.

[9] Hersh E, Hickam D. Applicability and quality of information for answering clinical questions on the web. JAMA 1998;280:1307–8.

[10] Hersh WR, Hickam DH. How well do physicians use electronic information retrieval systems? JAMA 2001;280:1347–52.

[11] Hersh W, Chute C. A world of knowledge at your fingertips: the promise, reality, and future directions of on-line information retrieval. Lake Buena Vista (FL): Hanley & Belfus; 1999.

[12] Nonaka I. A dynamic theory of organizational knowledge creation. Organization Science 1994;5:14–37.
[13] Huber G. Organizational learning: the contributing processes and the literatures. Organization Science 1991;2:88–115.
[14] Hersh W, Chute C. Information retrieval at the millennium. Orlando (FL): Hanley & Belfus; 1998.
[15] Lawton G. Knowledge management: ready for prime time? Computer 2001;34:12–4.
[16] Krulwich B, Burkey C. The infoFinder agent: learning user interests through heuristic phrase extraction. IEEE Expert 1997;12:22–7.
[17] Gudivada VN, Raghavan VV, Grosky WI, Kasanagottu R. Information retrieval on the world wide web. IEEE Internet Computing 1997;1:58–68.
[18] Greenberg I, Garber L. Searching for new search technologies. Computer 1999;32:4–11.
[19] Schleyer T, Spallek H, Spallek G. The global village of dentistry: Internet, intranet, online services for dental professionals. 1st edition. Berlin: Quintessence; 1998.
[20] Suarez HH, Hao X, Chang IF, Masys DR. Searching for information on the Internet using the UMLS and medical world search. Nashville (TN): Hanley & Belfus; 1997.
[21] Redman PM, Kelly JA, Albright ED, Anderson PF, Mulder C, Schnell EH. Common ground: the HealthWeb project as a model for Internet collaboration. Bull Med Libr Assoc 1997;85:325–30.
[22] Anderson PF, Allee N. Megasite. Available at: http://www.lib.umich.edu/megasite.
[23] Introna L, Nissenbaum H. Defining the web: the politics of search engines. Computer 2000;33(1):54–62.
[24] Frauenfelder M. The future of search engines. The industry standard 1998;34–41.
[25] Kerstetter J. IBM launches a clever search. PC Week 1999;31.
[26] Henzinger MR. Hyperlink analysis for the web. IEEE Internet Computing 2001;5:45–50.
[27] Ramakrishnan N. PIPE: web personalization by partial evaluation. IEEE Internet Computing 2000;4:21–31.
[28] Ince JF. Searching for profits. Upside 2000;92–104.
[29] Lew MS. Next-generation web searches for visual content. Computer 2000;33:46–53.
[30] Favela J, Meza V. Image-retrieval agent: integrating image content and text. IEEE Intelligent Systems 2000;14:36–9.
[31] Anderson PF, Allee N. The quality connection: health information megasites and search engines. In: Guide to health care on the Internet: ensuring the quality of health care information on the Net. New York (NY): Faulkner & Gray; 2000. p. 159–88.
[32] Lawrence S, Giles CL. Searching the world wide web. Science 1998;280:98–100.
[33] Lawrence S, Giles CL. Accessibility of information on the web. Nature 1999;400:107–9.
[34] Coiera E. Information economics and the Internet. J Am Med Inform Assoc 2000;7: 215–21.
[35] W3C. Resource description framework (RDF) model and syntax specification. Available at: http://www.w3.org/TR/REC-rdf-syntax/.
[36] Martin P, Ekulund PW. Knowledge retrieval and the world wide web. IEEE Intelligent Systems 2000;15:18–25.
[37] Dragan RV. Future agent software. PC Magazine 1997;16:190–6.
[38] Mladenic D. Personal web watcher, project page. Available at: http://www-2.cs.cmu.edu/afs/cs/project/theo-4/text-learning/www/pww/.
[39] Nick ZZ, Themis P. Web search using a genetic algorithm. Virtual Marketplaces 2001; 5:18–26.