Proceedings of the 6th WSEAS Int. Conf. on EVOLUTIONARY COMPUTING, Lisbon, Portugal, June 16-18, 2005 (pp291-296)

OPTIMIZING THE SEARCHING TECHNIQUES

Jatinder Ohri
Lecturer,Department of Computer ScienceKhalsa College, Amritsar

Rajesh Ohri
Department of Computer Science
Khalsa College, Amritsar

Karnpal Singh
Department of Computer Science
Khalsa College, Amritsar

Abstract

Information retrieval system display search results by various methods. This paper focuses on a model for displaying a list of search results by means of textual elements that utilize a new information limit that replaces the currently used information unit. The paper includes a short description of several studies that support the model. Most Internet search engines displays their information as a serially ordered list of results. In most cases this list includes the document title, URL, at times the first few lines of the document. The information as currently displayed to the user is incomplete and insufficiently focused on the search query. This requires to the user to actually read all the documents in the list with being able to discriminate.
With today's search engines most of the transactions yields a list of hundreds and even thousands of documents. While studies show that the average user only look at the first 10 to 20 search results. Finding the solution to this present a serious challenge to researchers in the field. This paper will suggest a way to locate the relevant document without having to read the listed documents.

1. Introduction

1.1 Search Engine

The first term we will define is search engine. Generally a search engine is any program that search a database and produce a list of results. To work at such an abstract level within this document would limit us to a very theoretical discussion. Therefore, for the purposes of this document, we will apply a more narrow definition of search engine as follows.

A system that uses automated techniques, such as robots and indexers to create indexes of the web, allows those indexes to be searched according to certain search criteria and delivers a set of results ordered by relevancy to those search criteria. Examples of search engines are AltaVista, Fast, and Goggle.
The term search engine is often used generally to describe both crawler based search engines and human power directories. These two types of search engines gather their listing in different ways and it is important to distinguish between them and their data gathering techniques.

1.2 Crawler-Based Search Engines

Crawler based search engine such as HotBot, create their listing automatically. They routinely "crawl" or "spider" the web, then a human editor will search through the results. This means that, unlike directories the site is likely to have several if not many pages listed with them. By correctly structuring web pages, crawler-based search engines find the web pages and determines how the site listed. Pages titles, body copy and other elements are all part of the criteria.

Human Powered Directories

A human-powered directory such as Yahoo depends on humans for its listings. A short description is submitted to the directory for entire site, or editors write one for sites they review. A search looks for matches only in the descriptions submitted.

Criteria that may be useful for improving a listing with search engines could have nothing to do with improving a listing in a directory. The only exception is that a good site, with good contents might be more likely to get reviewed for free than a poor site. Sites that are listed with directories, however, are more likely to be found crawler-based search engines and added to their listings for free.

1.3 Hybrid search engines or mixed results

In the webs early days, a search engine either presented crawler-based result or human-powered listing. Today, it extremely common for both type of result to be presented. Usually, a hybrid /search engine will favor one type of listings over another. For e.g. Yahoo is more likely to present human-powered listings. However, it does also present human powered listings. However it does also presents crawler-based results especially for more obscure queries.

1.4 Components of Crawler-based Search Engine

Crawler based search have three major elements. First is the Spider also called the Crawler. The spider visits a web page, reads it, and then follows links to other pages within the site. This is what it means when someone refers to a site being spidered or crawled. The spider returns to the site on a regular basis such as every month or two, to look the changes. Everything the spider finds goes into the second part of the search engine, the indes. The index sometimes called the catalogue is like a giant book containing a copy of every web page that the spider finds. If a web page changes then this book is updated with new information. Sometime it can take a while for new pages or changes that the spider finds to be added to the index. Thus, a web may have been spidered but not yet indexed. Until it is indexed, added to the index it is not available to those searching with the search engines.

Some Search engines index more web than others. Some search engines also index web pages more often than others. The result is that no search engines has the exact same collection of web pages to search through. That naturally produces differences, when comparing search results.

2. How Crawler-based search engines determines rankings

2.1 Keyword placement

Search engines will check for search keywords to appear near the top of a web page, such as in the headline or in the first few paragraphs of text. They assume that any page relevant to the topic will mention those words right from the beginning.

2.2 Frequency

A search engine will analyze how often keywords appear in relation to other words in a web page. Those with the higher frequency are often deemed more relevant than other web pages.

2.3 Meta Tags

Changing page title and adding Meta tags is not necessarily going to help a page do well for target keywords if the page has nothing to do with the topic. The keywords need to be reflected in the pages content. Not all engines read Meta tags. The search engines that do read Meta tags weight them differently.

2.4 Spamming

Search engine may also penalize pages or exclude them from the index, if they detect search engine "Spamming". An example is when a word is repeated hundred of times on a page, to increase the frequency and propel the page higher in the listings. Such engines watch for common Spamming methods in a variety of ways, including following up on complaints from their users.
Crawler-based search engines have plenty of experience now with webmaster who constantly write their web pages in an attempt to gain better ranking. Some sophisticated web masters may even go to great lengths to "reverse engineer" the location/frequency systems used by a particular search engine. Because of this, all major search engines now also make use of "of the page" ranking criteria.

3. Off the page factors

3.1 Link Analysis.

By analyzing how pages link to each other, a search engine can both determine what a page is about and whether that page is deemed to be "important" and thus deserving of a ranking boost. Link analysis is about "popularity" not volume. A website should link with quality web pages with related topics.
Sophisticated techniques are used to screen out attempts by webmasters to build "artificial" links designed to boost their rankings.

3.2 Click through Measurement.

A search engine may watch the results a user selects for a particular search, then eventually drop high-ranking pages that aren't attracting clicks, while promoting lower-ranking pages that do pull in visitors. As with link analysis, systems are used to compensate for artificial links generated by eager webmasters.
Content HTML text should appear on each page. Sometimes sites present large sections of copy via graphics. It may be visually appealing, but search engines can't read those graphics. That means they miss out on text that might make the site more relevant. Some of the search engines will index ALT text and comment information, along with Meta tags.

3.3 Submitting a listing to a Directory

Prior to attempting to submit a site, a written 25 word or less description of the entire web site should be developed. That description should make use of the two or three key terms that will return results. The target keywords/phrases should always be at least two or more words. Usually, too many sites will be relevant for a single word increases the rankings within a search engine.

4. Important:

In order to achieve high rankings, follow the search engine rules. Keep the content useful, improve the link popularity and monitor the search engine positing for improvement opportunities. Most search engines where the main results come from crawling the web will also provide human-powered "directory" results in some way. For example, in a search at Goggle, "category" links that lead to human-complied information often appear at the very top of the search results page.

For search engines where the main results come from human work, it's common for them to have a "backup" or "fallthrough" partnership with a crawler-based search engine. For example, if a search at Yahoo fails to find a match in Yahoo's own human-compiled information, then matches from Goggle provide answers.

4.1. A model for displaying textual search results

This section will define a hierarchical structure containing three levels for displaying search results. Search results can be displayed from textual databases by relying on tow basic principles; visualization of the results, and the use of textual components to design the list of results. This focuses solely on the use of textual components to display search information and external document information.

4.2. Results based on internal document information

In this category, a number of techniques are used, most of which include information components related to the search topic. Following is a description of the various of the various methods.

4.3. Significant sentences

Significant sentences can be descriptive sentences based on defined paragraphs in the document, for example: Abstract, introduction, Conclusion. Alternatively, sentences relevant to the search to the search query can be used, which include the terms that were the reason for the document being chosen.

4.4. Significant words

Significant words in the document are intrinsic descriptive, such as keywords or frequently repeated words. The document's author determines keywords, or they can be produced automatically. Frequently repeated words that are computer generated (including Stop List operation) can yield results that are similar but less exact.

4.5 Information from HTML tags

The language tags can provide us with information about the document. For example, paragraph or subtitle headings can be located by using the <H> tags, and can even be used to generate a table of contents. <META> tags contain information about the document as recorded by the document author, such as: abstract, keyboards and others. A certain amount of "noise" must be taken into account with these tags because of commercial rating considerations.

4.6. Additional information

Additional information can be generated from the actual document; for example, when a
Document includes citations from other documents, the titles of the cited documents can be used, assuming that they have a subject in common.

4.7 Results based on external document information

This category utilizes a number of methods that include information components based on the document's subject filed and not contained within the actual document. A description of these methods follows.

5. Document classification

This method displays the category with which the document is associated. Search engines that manually define document categories (such as Yahoo) can be used for this purpose. It is also possible to create categories with the aid of computerized algorithms, and the subject association of the document can be established by clustering all the search results

5.1. Citing documents

This refers to a situation in which one document cites another, where both have a subject in common. The citing documents can be located directly via the Internet, or by using a subject-oriented database such as the Science Citation Index. When the citing documents are located, either their titles or, alternatively, their cited paragraphs can be used.

5.2 Information from the database

The database in which the document is located can provide an indication of the document's subject in a number of ways. Subject oriented databases usually specify the database subject field. An attempt can be made to determine the database subject filed from the titles of additional documents contained in the database.

Conclusion

The objective of the studies was to examine some of the components of the search results display model. We found that, in addition to the alternative information unit must include lines by search context, keywords, and an indication of the document category. Authors of article and database administrators can benefit by including the suggested information components in each document using standardized means.

References

1. Amento, B., Hill, W., Terveen, L., Hix, D., Ju, P., An empirical

Evaluation of User interfaces for Topic Management of Web Sites, *Proceedings of CHI'99*, ACM Press, Pittsburgh PA, May 1999, 552-559.2.

2. Baldonado, M., Winograd, T., SenseMaker: an information-exploration interface supporting the contextual evolution of a user's interests, Proceedings of CHI97, 1997, 11-18
3. Drori, O., The User Interface in Text Retrieval Systems, SIGCHI bulletin, New York: ACM, July 1998,30 (3), 26-29.
4. Egan, D., Remde J., Landauer, T., Lochbaum, C., Gomez, L., Behavioral Evaluation and Analysis of a Hypertext Browser, Proceedings of CHI'89, New York: ACM, 1989, 205-210.
5. Luhn, H., Keyword in Context Index for Technical Literature, *American Documentation*, XI (4), 1960,288-295.
6. Zamir, O., Etzioni, O., Grouper: A Dynamic Clustering Interface to Web Search Results, WWW8 Proceedings, Toronto: WWW, 1999