Universidade do Porto

Faculdade de Engenharia

FEUP

Information Retrieval Techniques in Commercial Systems

Google, Teoma and WiseNut

Ana Isabel Pinto Correia
Sérgio Sobral Nunes

Information Storage and Retrieval
Master in Information Management, June 2002

Abstract

In this essay we are going to review three search engines: Google, Teoma and WiseNut. We will try to evidence the main information retrieval techniques currently in use by these services.

Within each service, an introduction is provided and the technical details are presented according to three sub-sections: Document Acquisition and Storage, Query Introduction and User Options and Results Selection and Presentation.

Some engines provide more technical details and have a greater press coverage, while others are newer and less information is available. So, in this work there are notable differences between each kind. Nevertheless, among all the information available, we need, within our capacities, to distinguish marketing propaganda from real facts.

In the conclusions our opinions, comparing these 3 systems, are stated along with some remarks about the challenges and threats that commercial search engines on the Web face.

As a final note, we would like to refer the fact that the Internet is in constant and fast evolution and, due to this, any work about it faces the risk of quickly becoming inaccurate. This paper reflects the situation that we found in June of 2002.

Google

Introduction

Google (www.google.com) is one of the most visited websites in the world [1] and, according to a recent rating [2] the most used search engine, leaving others well behind.

Google Inc., owner of Google search engine, was founded in 1998, by Larry Page and Sergey Brin, two former graduate students at the Stanford University. The systems was build based on a technology developed by the two students – PageRank, which analyses the links pointing to a page to calculate its relevance. [3]

Since 1998, the service has gained tremendous popularity among web users. According to Google's data more than 150 million searches per day [4] are answered and 3 billion web documents[1] are currently indexed [5]. The company has also gained notoriety due to its innovative and "fast moving" style. Along with it's search engine, Google Inc. currently offers a collection of diversified services.

Google Answers[2] is the latest service launched by the company, it is a bet in the "expert-advice" market. In a nutshell, the service gathers researchers that answer user's questions for a price, in real time. [6]

Among others and besides it's web search service, Google offers catalogs search, translation services, spell correction, an enterprise search device, an Application Programming Interface (API) to some of its search functions and access to news and headlines from around the world. [7]

News have been found about the possibility of an Initial Public Offering (IPO) in the near future. Their biggest challenge is, according to several analyst, the elaboration of a solid and well defined business model. [8]

IR Techniques

Document Acquisition and Storage

Google searches more than 3 billion Web documents, which include Web pages, images and Usenet postings [5]. Google uses a standalone Web crawler, distributed trough several machines, to create indexes and copies of the documents [13]. Google also offers a way to manually add a site to it's database [14].

Besides standard .html files, Google also indexes other file types including .pdf, .ps, .doc, .xls, .txt, .ppt, .rtf among others [15].

In it's hardware infrastructure Google uses hundreds of cheap PC running Linux [16] and DRAM for storage as opposed to hard disks. According to Eric Schmidt, "DRAM is 200,000 times more efficient when it comes to storing seekable data" and cheaper. [17].

A copy of each crawled page is stored in Google's repository. Indexes are created using sorted words, pointing to an inverted index file. Also, PageRank, anchor text and text formatting options (i.e.: font size, style) are stored along with each page. [12]

---

[1] Web pages, images and newsgroups postings.

[2] https://answers.google.com/answers/main

In essence, PageRank works by considering a link from page A to B as a weighted vote, being this weight calculated according to page's A own rank. The anchor text of a page is the text that is associated with all the links to that page. According to Google's founders, most of the time, this text is a more accurate description of the page than it's own contents. Finally, text formatting options are considered, making larger, bolder words more weighted than other words. [12], [39]

Google refreshes it's indexes more frequently for certain pages than the rest of the database. As of February 2002, 3 million pages were being refreshed on an almost daily bases. [18]

Query Introduction and User Options

Since it's foundation, Google has been steadily introducing new features.

Google uses Boolean search without nested expressions support and with some variations. By default, it automatically uses the AND operator between terms, the minus symbol can be used to perform a NOT function and the OR operation is supported (using OR in upper case). [19]

Google supports stop words, disregarding them unless a plus sign is used to force it's inclusion. [20]

Google supports phrase search using double quotes and, even when double quotes are not used, proximity of the word is considered. Google does not uses stemming, nor truncation, but allows the use of '*' as a wildcard in the middle of a phrase. For example, searching for "Search Engine" wields quite different results from "Search * Engine". [18]

Google has a myriad of other choices available trough it's advanced search page (i.e.: other interface languages, filtering by document type, searching specific document parts, date filter, search in a site, etc) [21].

Results Selection and Presentation

To select which documents are presented, Google combines a document's PageRank value, anchor text and proximity [12]. The ranking function has many parameters and figuring out the right values is, according to it's authors, "something of a black art" [12].

Results are presented including the file type (when different from HTML), the title, URL, a page extract near the search terms, the file size, a link to search for similar pages, a translation option for some languages and, when the page is cached, a link to it's copy in the repository (a unique feature among the main search engines). Results are clustered by server with two visible results and a link to "More results from *server*".

Google helps users by correcting misspelled words in their search queries using, not a predetermined dictionary, but it's own index of the entire Web [8].

Google's visual interface is one of the simplest and, according to many, one of the reasons to Google's success, "it's simple and it works". [22]

Teoma

Introduction

Teoma was founded in 2000 in Piscataway, New Jersey by a team of scientists from Rutgers University, headed by Professor Apostolos Gerasoulis. In late 2001 it was acquired by Ask Jeeves, Inc.[3] and currently provides search results for their service – www.ask.com. [9] Several news have appeared referring Teoma as Google's newest contender. [10], [11]

Teoma was re-launched in April 2002 [23].

IR Techniques

Document Acquisition and Storage

Little information was found about the crawling process and storage details of Teoma.

We were able to find that, according to Teoma's VP of search technology, the "ideal index size will be between 350 million and 500 million pages" because there aren't 2 billion[4] useful pages on the Web [24].

Teoma has no free submission to it's database, it's a paid service [25]. So, this adds to one of Teoma's main flaws, it's database size, as a recent study [26] reflects. Also, Teoma only seems to index standard html pages, ignoring other file types.

When compared, Teoma's database freshness stands quite below Google's one [27].

Query Introduction and User Options

Teoma has few advanced search capabilities.

It offers phrase search using double quotes or using a check box available in the page. Teoma's help page states that, using this feature, it returns "results which exactly or closely match the given phrase", so not all phrase matches will by accurate.

It also supports a very limited form of Boolean search. By default it uses an AND operator between terms but this option can be overridden using a minus sign before the term to exclude from the query (negation). [29]

Teoma ignores stop words unless these are preceded by a plus sign.

Teoma does not offers any filtering capabilities (i.e.: adult pages).

Results Selection and Presentation

Teoma uses a striped down, Google style, interface.

Teoma separates results in 4 different sections: 'Sponsored Results', 'Results', 'Refine' and 'Resources'. 'Sponsored results' are gathered from Overture's[5] paid ads and appear, when available, in the top left side of the page.

---

[3] http://www.irconnect.com/askj/

[4] Google's index size at the time.

Teoma uses link patterns to try to identify Web communities around a determined subject. According to it's help page [30], Teoma dynamically clusters results into actual communities as they exist on the Web.

Having these communities defined Teoma is able to define 'subjects', which appear in the 'Refine' section. For example, searching for 'photography' returns refinements as 'digital photography', 'nature photography', 'history, museum' among others. This is presented at the top right side of the page. Following these links, Teoma filters the results to pages within the selected community.

The 'Resources' are pages that Teoma identifies as expert within a subject Web community, generally hub pages (sometimes called metasites). These links are displayed in the bottom right side. [31]

Finally, the bulk of the results page, available in the bottom left side, is the 'Results' section. The pages are ranked using a algorithm that considers the number of same-subject page that reference them. In a nutshell, it takes in consideration the links within a Web community, not the entire World Wide Web (as Google does). For example, when searching for golf pages, only pages that belong to this subject "vote" [32]. Teoma calls this technology "Subject-Specific Popularity " that "counts the links that count". Along with each result in the main section the page's title and URL are displayed and the description of the site is shown (gathered from the site's meta tags).

[5] http://www.overture.com

WiseNut

Introduction

WiseNut Inc., owner of www.wisenut.com search engine, was founded in 1999 by Yeogirl Yun. The search engine was launched in September 2001 and has been growing at a steady pace, being well placed in some independent reviews. [26], [28].

Nevertheless, it still hasn't reach "prime time", not figuring in the top ten most used search engines list [2].

In March 2002 LookSmart acquired WiseNut Inc. [33]

IR Techniques

Document Acquisition and Storage

WiseNut has distinguish itself by putting together a very large database since it's beginnings. Currently, according to Search Engine Showdown [28], it's the second largest after Google's. On the other hand, the "freshness" of it's pages makes it fall to the bottom of the list [27].

WiseNut's CEO claims that it's crawler (named ZyBorg) can "index up to 50 million pages a day, and process millions of documents in a single hour, indexing every word on every page – all this with just a dozen or so off-the-shelf servers" [34].

Unlike Google, WiseNut only indexes standard html pages.

Query Introduction and User Options

WiseNut lacks almost all advanced search capabilities found in other search engines. By default it uses an AND operator between the terms in the query. It supports term exclusion using a minus sign before the words and phrase search using double quotes. [35]

Stop words can be forced into the query using a plus sign before the word.

Finally, some options can be set using the preferences page [36]. Among other things, the user can filter results to one or more of 25 languages, customize the results page and turn-on adult filtering.

Results Ranking and Presentation

WiseNut tries to automatically generate semantic related searches to create "communities", known as WiseGuide categories [37]. It's algorithms uses link pattern and search analysis.

In the top of the results page, WiseGuide categories, for the given query, are presented. This tool can be used to narrow down the search by subject, helping the user to deal with, for example, problems of polysemy. It can be particularly useful when dealing with large result sets.

When ranking results, WiseNut strategy is also similar to Teoma's, it calculates the page relevance by summing up the references made by pages within the same subject (search-in-context).

WiseNut clusters results by site and displays a link to show more results from that site. Each results presents the page title, the URL, an extract of the content near the searched terms and a "Sneak-a-Peek" option. The later will open a sub-windows with the contents of the actual site. This is not a cached version, but an actual connection to the selected page. This is quite an interesting option which unfortunately only works in Microsoft Internet Explorer. [38]

Conclusion

In this essay we've tried to give a brief overview of the information retrieval techniques used by 3 search engines. We've analyzed Google, the most used search engine and 2 other, very recent, services – Teoma and WiseNut.

Google was one of the first to use popularity rankings to define the relevance of the pages (see [39]). Teoma and WiseNut try to extend this concept but addressing the notion of semantic networks. As we write, Google is launching a new service – Google Labs[6], and seems to be working in this area with it's Google Sets[7]. In this beta service, the context relations between words are explored.

Google has enormous popularity among Internet users and has always been very focused in the quality of it's results. Innovation is also a key aspect of Google's way of doing business, almost every month a new service is launched. So, although we were able to get good results (with some unique pointers in it) searching with Teoma and WiseNut, it will be difficult for them to steal Google's throne.

One of the main reasons is that today, unlike what was happening in 1998 (when Google appeared), Google users are satisfied with their results. So, it's less likely that the average Internet users will look for other alternatives. People are happy with Google.

Nevertheless, Teoma, who is more strongly betting on web search, is likely to become one of the major search engines in the future. WiseNut, on the other hand, is launching several products for intranets and dedicated services, while it's site tends to serve as a show case for it's technology.

As a final remark, we would like to mention the need for search engines to find an objective business model. This will be their biggest challenge. Google is planning an IPO but it's profits model is still not clear. As a remark, Overture, a paid-per-placement search engine, is expected to post $126 million in revenue for the first quarter, while Google estimates it's sales at $15 million to $25 million [8].

[6] http://labs.google.com

[7] http://labs.google.com/sets

References

[1] Global Top 50 Web and Digital Media Properties. *Jupiter Media Metrix*. (December 2001). http://www.jmm.com/xp/jmm/press/globalTop50WebProperties.xml 12-05-2002 16:00.

[2] Jupiter Media Metrix Search Engine Ratings. *Search Engine Watch*. (April 29, 2002). http://searchenginewatch.com/reports/mediametrix.html 12-05-2002 16:00.

[3] Google Corporate Information: Google History. *Google*. (2001). http://www.google.com/corporate/history.html 12-05-2002 16:00.

[4] Google Press Center: Company Overview. *Google*. (2001). http://www.google.com/press/overview.html 13-05-2002 10:00.

[5] Google Offers Immediate Access to 3 Billion Web Documents. *Google*. (2001). http://www.google.com/press/pressrel/3billion.html 13-05-2002 10:00.

[6] Google gives some advice...for a price. *CNET*. (April 19, 2002) http://news.com.com/2100-1023-887360.html 13-05-2002 10:00.

[7] Google Press Room: Google Reviewer's Guide. *Google*. (2001). http://www.google.com/press/guide/index.html 13-05-2002 11:00.

[8] Google's Toughest Search Is for a Business Model. *The New York Times*. (April 8, 2002). http://www.nytimes.com/2002/04/08/technology/ebusiness/08GOOG.html?pagewanted=all&position=top 13-05-2002 11:00.

[9] Teoma Search: Development Team and History. *Teoma*. (2002). http://static.wc.teoma.com/docs/teoma/about/developmentTeamHistory.html 17-05-2002 10:00.

[10] Teoma preps relaunch, wants to be Google-beater. *The Register*. (March 29, 2002). http://www.theregister.co.uk/content/6/24642.html 17-05-2002 10:00.

[11] Search start-ups seek Google's throne. *CNET*. (August 28, 2002) http://news.com.com/2100-1023-272230.html 17-05-2002 10:00.

[12] Brin, Sergey, Page, Lawrence – The Anatomy of a Large-Scale Hypertextual Web Search Engine. *Proceedings of the 7th International World Wide Web Conference*, 1998.

[13] Googlebot: Google's Web Crawler. *Google*. (2002). http://www.google.com/bot.html 18-05-2002 10:00.

[14] Google: Add your URL to Google. *Google*. (2002) http://www.google.com/addurl.html 18-05-2002 10:00.

[15] Google: FAQ – File Types. *Google*. (2002). http://www.google.com/help/faq_filetypes.html 18-05-2002 11:00.

[16] Google Bets The Ranch On Linux. *TechWeb*. (May 30, 2000). http://content.techweb.com/wire/story/TWB20000530S0011 18-05-2002 11:00.

[17] Three Minutes With Google's Eric Schmidt. *PCWorld.com*. (January 30, 2002). http://www.pcworld.com/news/article/0,aid,81685,00.asp 18-05-2002 11:00.

[18] Review of Google. *Search Engine Showdown*. (2002). http://www.searchengineshowdown.com/features/google/ 18-05-2002 11:00.

[19] Boolean Searching on Google. *Search Engine Showdown*. (November 2, 2000). http://www.searchengineshowdown.com/features/google/googleboolean.html 19-05-2002 11:00.

[20] The Basics of Google Search. *Google*. (2002). http://www.google.com/help/basics.html 19-05-2002 12:00.

[21] Google Advanced Search. *Google*. (2002). http://www.google.com/advanced_search 19-05-2002 12:00.

[22] Searching for Google's success. *CNET*. (September 30, 2001). http://news.com.com/2009-1023-273704.html 19-05-2002 12:00.

[23] Press Release: Advanced Search Engine Teoma.com Launches. *Ask Jeeves, Inc.*. (April 2, 2002). http://www.irconnect.com/askj/pages/news_releases.mhtml?d=25648 19-05-2002 12:00.

[24] Teoma preps relaunch, wants to be Google-beater. *The Register*. (March 29, 2002). http://www.theregister.co.uk/content/6/24642.html 19-05-2002 12:00.

[25] Ask Jeeves: Submit a Site. *Ask Jeeves*. (2002). http://static.wc.ask.com/docs/addjeeves/Submit.html 19-05-2002 12:00.

[26] Relative Size Showdown. *Search Engine Showdown*. (March, 2002). http://www.searchengineshowdown.com/stats/size.shtml 19-05-2002 12:00.

[27] Freshness Showdown. *Search Engine Showdown*. (April, 2002). http://www.searchengineshowdown.com/stats/freshness.shtml 19-05-2002 12:00.

[28] Database Total Size Estimates. *Search Engine Showdown*. (March, 2002). http://www.searchengineshowdown.com/stats/sizeest.shtml 19-05-2002 12:00.

[29] Teoma Search: Search Tips. *Teoma*. (2002). http://static.wc.teoma.com/docs/teoma/about/searchTips.html 19-05-2002 12:00.

[30] Teoma Search: Search with Authority. *Teoma*. (2002). http://static.wc.teoma.com/docs/teoma/about/searchWithAuthority.html 19-05-2002 12:00.

[31] Review of Teoma. *Search Engine Showdown*. (2002). http://www.searchengineshowdown.com/features/teoma/review.html 19-05-2002 12:00.

[32] Is this the rival to Google? *The Register*. (July 24, 2001). http://www.theregister.co.uk/content/6/20614.html 19-05-2002 14:00.

[33] LookSmart Strengthens Leadership Position in Search Targeted Marketing With Acquisition of WiseNut, Inc. *LookSmart, Ltd*. (2002). http://www.shareholder.com/looksmart/news/20020312-74579.cfm 19-05-2002 14:00.

[34] WiseNut's Online Search Goes Into Full Production. *About.com: Web Search*. (September 5, 2001) http://websearch.about.com/library/searchtips/bltotd010905.htm 19-05-2002 14:00.

[35] WiseNut: WiseSearch. *WiseNut, Inc.*. (2001). http://www.wisenut.com/wisesearch/wisesearch.html 19-05-2002 14:00.

[36] WiseNut: Preferences. *WiseNut, Inc.*. (2001). http://www.wisenut.com/preferences 19-05-2002 14:00.

[37] WiseNut: What's WiseGuide. *WiseNut, Inc.*. (2001). http://www.wisenut.com/help/help_wiseguide.html 19-05-2002 14:00.

[38] Review of WiseNut. *Search Engine Showdown*. (2002). http://www.searchengineshowdown.com/features/wisenut/review.html 19-05-2002 14:00.

[39] Page, Lawrence, Brin, Sergey et al. – The PageRank Citation Ranking: Bringing order to the Web. *Technical Report, Computer Science Department, Stanford University*, 1998.

Bibliography

Kobayashi, Mei; Takeda, Koichi – Information Retrieval on the Web. *ACM Computing Surveys*. 32:2 (2000) 144-171.

Baeza-Yates R.; Ribeiro-Neto B. – *Modern Information Retrieval*. New York: Addison Wesley, 1999. ISBN 0-201-39829-X.