RP
6
c.1

Search Engines & "Flying Search Engine"

By:

Mohamed Dbouk

Submitted in the partial fulfillment of the requirements
for the Degree Master of Science

Department of Computer Science
Lebanese American University
June 2005

BY90595

Search Engines & "Flying Search Engine"

Mohamed Dbouk

Project

Submitted in partial fulfilment of the requirements of the degree
of Master of Science at the Lebanese American University
Beirut, Lebanon
June, 2005

[redacted]

Dr Nash'at Mansour (Advisor)
Associate Professor of Computer Science
Lebanese American University

[redacted]

Dr. Faisal Abu Khuzam
Assistant Professor of Computer Science
Lebanese American University

I grant to the **LEBANESE AMERICAN UNIVERSITY** the right to use this work, irrespective of any copyright, for the University's own purpose without cost to the University or to its students, agents and employees. I further agree that the University may reproduce and provide single copies of the work, in any format other than in or from microforms, to the public for the cost of reproduction.

_____

ACKNOWLEDGMENTS

I would like to thank my advisor Dr. Nashaat Mansour for his guidance throughout my M.S. studies; as well as, for being in my project committee. Thanks are also due to Dr. Faisal Abu Khuzam for being in my project committee too.

Finally, I would like to thank my sister Nancy for her long support.

ABSTRACT

By

Mohamed Dbouk

Search engines search documents for specified keywords corresponding to a user's query and return a list of documents (web sites) where the keywords were found.
In this project we design and implement a special-purpose search engine called Flying Search Engine (FSE). The FSE yields relatively good results especially when the databases are small and specialized. Hence, we recommend the adoption of the FSE in small corporations such as family businesses, banks, and schooling.

Contents

Chapter 1: Introduction

File Transfer Protocol (FTP) in the early 1990, a system specifying the way of exchanging files over the internet, was adopted to store and retrieve files. Sharing a file thus meant setting up an FTP servers. The FTP however started its journey towards extinction when Archie was born. Archie was created by Alan Armitage a student at McGill University in Montreal. Archie created a database of files indexed from FTP sites across the internet into which a user can have access when its regular expression matcher retrieved the appropriate file corresponding to the user's query. Then came Matthew Gray's Wanderer the first robot on the web to count Web servers and it grew to capture URLs into which the URL database Wandex, the first web database, was created. Wanderer however, spiked a controversy at that time since it led to the downgrade of the Net because it accessed the same page hundreds of times in the same day. [11]

Inspired by Wanderer, Martin Koster created ALIWEB (Archie-Like Indexing of the Web) in October 1993. Similar to Archie, ALIWEB did not have a web searching robot yet it was the HTTP breakthrough. Net bandwidth was left index by wanderer since webmasters posted their own index information for each page they wanted listed.[11]

By December 1993, three robot powered search engines were launched; JumpStation, the World Wide Web Worm and the Respiratory-Based Software Engineering (RBSE) spider. JumpStation was a linear system matching keywords with the title and header of web pages. This design led to its halting breakdown since the JumpStation went limp as the web grew. In the same route, WWWWorm used regular expressions to search the index

that contained the titles and URLs of the pages it visited to which the order of results had no relevance. The RSBE tackled this flaw by applying a ranking system based on relevance to the keyword string.

After that a great number of categorized links and Web search engines started appearing. In mid 1993, six Stanford undergraduates released Architext- a search software for webmasters to use on their own websites. Their idea was utilizing statistical analysis of word relationships to allow for more competent searches through the wide amount of information on the Internet. Architext is now known by Excite for Web Servers.[11]

Unfortunately, these spiders needed the users to specify their queries because they lacked the intelligence of understanding their indexing content; thus no specificity, no hits.

EINet Galaxy known as Galaxy.com was created in response to this deficiency; this marked the appearance of browsable/searchable web directory. During this time people started to create pages to their favourite documents. David Filo and Jerry Yang, two Stanford University Ph.D. candidates, created their own pages in April 1994 to which they were rewarded by instant popularity. The collection of pages was called Yahoo! Yahoo! (www.yahoo.com) became a searchable directory to better organize data and aid in their flexibility of retrieval after the number of links grew and the pages began to receive thousands of hits a day. Yahoo! then was not classified as a search engine since entries were entered and categorized manually. Yahoo! has since blurred the distinction between an engine and a directory by automating some aspects of gathering and classification process. (LOGIKA Corporation, 2004)

1.1. Overview of web search engines

Search engines are three basic types:

Search Scripts, written in Perl or C searching sites of a few thousand pages. These don't constitute complex algorithms for optimal search and indexing and do not function if the site is over a thousand pages. Those are most beneficial for a home-made site to which a search box is similar to a gadget. Those scripts index modified pages on a termed basis in a few minutes or seconds producing a fixed output with a small layout options for the search results. Those scripts are CGIs so they do not store data in memory but instead parse it each time it is invoked and thus those search engines are relatively slow. Among such scripts you can find WebGlimpse (http://www.webglimpse.net/) or ht://Dig (http://www.htdig.org/).

The second type is Search Servers which are ISAPI or WAI applications but are sometimes mixed with CGI scripts. Those engines overcome the problem of indexing by constantly rereading from the hard disk.This search engine category is required for larger sites that need good searches since they demand more hardware power and more memory. This demand is due to the fact that Search Servers work as a permanently running servers and answer to multiple search requests. Alkaline is designed as a server persistent search engine and also Infoseek Ultraseek (http://www.ultraseek.com/ or Thunderstone Webinator (http://www.thunderstone.com/webinator/).

Search servers maintain indexes in RAM or use some internal swap mechanism. They have complex algorithms for searching and indexing, kept secret by their designers. Alkaline uses the concept of "cellular expansion", cells are fast and resistant to growing data, giving an interesting performance and opening the door for future research. Alkaline powered sites maintain an index of 500,000 pages with about 450,000 word forms and run on industry average Pentium III or Sun Ultra servers. It can handle from two to three search requests per second.

Finally, Distributed Servers target searching and indexing of the whole web. Large companies are competing for the best technology and for the most relevant search results. It is a plan of a parallel implementation of Alkaline for a cluster over a TCP/IP platform independent network and for IBM SP2. Already numerous tests have been made over a PVM network. For a distributed architecture the aim is for Alkaline to index 5-10 million pages running fast on a cluster of 32 PII PCs. Unlike Altavista, no plan set search limits to Alkaline depending on the price that is it is distributed as one single product for the same value no matter what the search is. Choosing Alkaline, requires also the choosing of a team that works for the future.

Distributed search servers perform both parallel indexing and searching which become faster with increased hardware power but depending on the network charge overhead. All major search engines use distributed architectures and can hit hundreds of requests per second.

A search engine is a coordinated set of programs that includes:

- A spider aka a "crawler" or a "bot" that goes to every page or representative pages on every Web site that wants to be searchable and reads it, using hypertext links on each page to discover and read a site's other pages
- A program that creates a huge index, a "catalog", from the pages that have been read
- A program that receives the search request, compares it to the entries in the index, and returns results back to the user

However, search engines are facing new limitations as searches are becoming more advanced and sites are rapidly increasing in number and complexity. The three major limitations are;

- Lack of the higher quality sources that are not found on the Web
- No concept of classification as found in library systems
- They are like an index of every word on every page in every book in a user's library

Even Google which as discussed later is the most popular and sought after search engine is facing problems; those are:

1. A limit set on how many terms that can be used in a search query (capped at ten)
2. A limit set on how many of the numerous supposed results can be viewed (usually less than 1000)

1. Google caps the number of terms in a search query at ten and thus we cannot use more than ten terms in any one search or we'll get the following message:

"Eleventhword" (and any subsequent words) was ignored because we limit queries to 10 words.

This is problematic for the following reason: sometimes we need to search for a specific item that is very common on Web sites but not in the desired context. So to help narrow the search, we need to enter all sorts of phrases to exclude but we are cornered with the ten words maximum criteria.

Consider the following example discussed on a user's chat room blog, if we try looking for a cloth foldable shopping cart, and not the digital virtual shopping

cart. This poses a problem since every e-commerce site has a "shopping cart" and so to find just the sites that sell the actual carts becomes quite a challenge.

If the search is broad and generic, Google will show for a search in Sociology for instance: Results 1 - 10 of about 2,100,000. However, trying to see one of the entries far down the list, we find that only the results below the 1000th are accessible, sometimes not even up to the 900th. So although the user is told that there are over two million results to this query, he can only view a very small portion of them.

Are there any alternatives for using a search engine? The alternative is exploring a structured directory of topics. An example of which is Yahoo which is the most widely-used directory on the Web. Yahoo however, also allows the use of its search engine. A number of Web portal sites offer both the search engine and directory approaches to finding information.

1.2. Objectives and Scope:

The objectives of this project are as follows:

- To provide brief definition and description of major Web Search Engines.
- To give a comparison and contrast of Web Search Engines.
- To List and explain the limitations of major Search Engines.
- To shed light on the concept and the advantages of specialized Search Engines.
- To develop and document a search engine "Flying S E".
- To present results of using Flying SE and compare the results to some major search engines.

1.3. Organization of this Report

Chapter 1 presents an overview of search engines as a whole in addition to introducing the report organization. Chapter 2 is a study of the major web search engines. This includes a description of a search engine with a focus on the spider first and the meta-searcher second. The other part of this chapter describes some of the most used search engines which include: Google, Yahoo, Ask Jeeves, All the web, AOL Search, Hotbot, Toema, and AltaVista. Experimental comparison of the latter search engines is studied and the dependency of some search engines on others. Also in this chapter the limitations of search engines are discussed and finally a view of the specialized search engines is briefed. Following we have chapter 3 which is a modest analysis of the architecture of a search engine. Chapter 4, consequently then is an examination of a software design of Flying Search Engines specifically their function and design enquiry. In the same notion, chapter 5 portrays some experimental results for the Flying search engine; as well as of its limitations. Finally, chapter 6 is a concise conclusion of the paper as a whole. References of this project are listed here as well. In the final part of this chapter we have the appendix where an account of the method of operation of the project design at hand is exposed counting the installation process together with the significant features.

Chapter 2: Major web search engines

2.1. What is a search engine?

Search Engine is computer software that assembles documents lists and contents usually on the WWW (World Wide Web). Thus it is a program that searches documents for specified keywords and returns a list of the documents where the keywords were found [13]. Although search engine is really a general class of programs, the term is often used to specifically describe systems like Alta Vista and Excite that enable users to search for documents on the World Wide Web and USENET newsgroups. Search engines respond to a user entry· a query, by searching for a match for the query through a list of documents called web sites when on the World Wide Web and displaying the corresponding matches. While some search engines include the starting portion of the text of Web pages in their lists, others include only the titles or addresses known as Universal Resource Locators, or URLS of Web pages. In addition, some search engines occur apart from the WWW, indexing documents on a local area network (LAN) or other system.

Search engines do not search the World Wide Web directly instead they search in databases of the full text of web pages residing on servers. Thus, using the search engine is searching older documents that become recent upon clicking on a desired link.[9]

Typically, a search engine works by sending out a spider to fetch as many documents as possible. Another program, called an indexer, then reads these documents and creates an index based on the words contained in each document. Each search engine uses a proprietary algorithm to create its indices such that, ideally, only meaningful results are returned for each query.

2.1.1. What is a Spider?

A spider (also called robot, softbot, wanderer, crawler, and fish) is a computer program that automatically monitors documents i.e. Web pages on the World Wide Web (WWW). Web pages include at least one link to another Web page and could include hundreds of links.[13] A spider works through this structure by starting at one Web page and continuing by following every link on a Web page and then following every link in the new Web pages. Although, they are said to "crawl" the web to choose the pages to be included, they stay in the same place. They follow the links already stored in their databases; they cannot type a URL or judge for appropriate web pages. This follows that if a search engine is not linked to any page in the database, the search engine cannot find it. So if a new page is to be included in the search engine database, a request to the search companies is to be submitted.

When a spider find the specific page it passes them for "indexing" by another computer. The text is identified and stored in the database.

Some spiders save the Uniform Resource Locator (URL) of every Web page they visit. Search engines use those spiders to build indexes of Web pages to be accessed to search for information on a particular topic. Indexing spiders often also store the title, part or the complete text of a Web page for detailed searches.

In order to update lists or provide lists of new Web pages, some spiders store only URLs of web pages not listed yet. To correct lists, URLs that are no longer valid are noted to correct the lists.

2.2.2. What is a Meta-Searcher?

Meta Search engines search more than a search engine and a subject directory at the same time. They then compile the results in an appropriate display. They sometimes arrange the results in a uniform list. Some even offer the ability of search refinement, customization of search engines and directories and even the time spent on the search. Usually, those utilities run as server-side applications but some need to be downloaded and installed on the computer.[13]

MetaCrawler functions by reformatting the search engine output from the various engines that it indexes it onto one concise page. Throughout MetaCrawler's history, the search engine companies that it worked with did not entirely approve of this procedure. The most common complaint was that the advertising banners that the search engines had on their sites were not appearing when a user employed MetaCrawler. This meant that their ads were not reaching the intended audience, reducing their ad revenues. The move to go2net heralded MetaCrawler's concession to these concerns.

Now MetaCrawler displays the ads from each search site right above the results. MetaCrawler users were not thrilled by this change because it increased the time it took for the result page to download. However, skilful design of the result pages now causes the text to load first, calming the restless native users.

2.2. Major Web Search Engines:

Searching Features

Alternative/Inclusive Default

.An inclusive default search engine treats two separate words like there is an AND between them. An alternative default search engine treats the two words however as if there is an OR between them. Thus there can be a gross difference between the results of an inclusive and an alternative search engine.

- Inclusive default search engines include Google, HotBot, and Lycos.
- Alternative search engines include AltaVista and Excite.

Many search engines leaves the choice to the user to designate an alternative or inclusive search engine by using connectors such as OR, AND and the + sign.

Keyword/Concept Default

Some search engines adapt automatic concept as a default. Advanced researchers utilize keyword searching where the string of characters typed is searched for thus a concept default might be inappropriate in their case. Concept searching does not only looks for the character string but for other word forms, synonyms and other words that statistically appear with the typed word.

Table 1 shows some of the features of some major search engines,

Table 1. Major SE features[9]

	AllTheWeb	AltaVista	Google	HotBot	MSN	Teoma
Boolean Operators	and, or, andnot, rank, +, -	+, - AND, OR, AND NOT	+, -, OR	AND, OR, NOT, +, -	AND, OR, NOT +, -	OR, +, -
Nesting of Boolean Operators	()	()	()	()	()	No
Boolean Default	AND	AND	AND	AND	AND	AND
Phrase Search	" "	" "	" "	" "	" "	" "
Title Field	title:	title:	intitle: allintitle:	title, domain, media, feature	title:	intitle:
URL Field	url:	url:	inurl: allinurl:	url	via pull down menu	inurl:
Other Fields	site: language: filesize: filetype:	anchor: applet: object: domain: filetype: host: image: like: link: text:	site: filetype: link: intext: inanchor: daterange: related:	domain: feature: linkdomain:	feature: domain: linkdomain: scriptlanguage:	site: geoloc: lang: last: inlink:
Truncation & Wildcard	No	*	No	*	*	No
Case Sensitive	No	Yes	No	Yes	Yes	No
Special Features	News stories, pictures, video clips, MP3 and ftp files Language search PDF files	Images, Audio, Video, Directory, News, Language Search, Translation Service PDF files	Images, Usenet Groups, News, Open Directory, Cached links, Spell Check, Translation, Similar Pages PDF files	Images, Video, Audio, Javascript, MP3, Date, Language, Directory	Document Directory Depth, Images, Videos, Audio, MP3, ActiveX, Shockwave, JavaScript, VBScript, Acrobat, Java Applets, Language Search, Region Search	Refine feature, Resources section of results

2.2.1. Google

Google is a top choice for those searching the web. Google provides the option to find more than web pages. Using on the top of the search box on the Google home page, a user can seek out images from across the web; discussions that are taking place on Usenet newsgroups, locate news information or perform product searching. Using the More link provides access to human-compiled information from the Open Directory.[2]

Google also offers cached links that lets the user see older versions of recently changed ones. It offers excellent spell checking, easy access to dictionary definitions, integration of stock quotes, street maps, telephone numbers and more. The Google Toolbar provides easy access to Google and its features directly from the Internet Explorer browser.

Google also operates its own advertising programs. The cost-per-click AdWords program places ads on Google as well as some of Google's partners. Google is also a provider of unpaid editorial results to some other search engines.

Google is the number one among all the search engines competitors relative to privacy issues.[9]

- Google was the first search engine to use a cookie that expires in 2038. Immortal cookies are now commonplace among search engines; Google was the one who set the standard though because no one bothered to challenge them. This cookie places a unique ID number on the user's hard disk. Anytime a searcher is on a Google page, he would get a

Google cookie if he doesn't already have one. If he has one, Google reads and records the unique ID number.

- Google records for all searches the cookie ID, Internet IP address, the time and date, search terms, and browser configuration. Google is more and more customizing results based on user's IP number that is "IP delivery based on geolocation."
- Google has no policies concerning the retention of data since it is able to easily access all the user information it collects and save.
- Google's free toolbar, if the advanced features are enabled, Explorer phones home with every page surfed. It reads the cookie too, and sends along the last search terms used in the toolbar. Google's toolbar updates to new versions without asking. That is if a toolbar installed, Google essentially has complete access to the hard disk every time the user calls home. Unlike most software vendors including Microsoft which ask if the user desires an updated version, Google does not.

2.2.2. Yahoo

Launched in 1994, Yahoo used to be a directory where human editors organize web sites into categories. However, in October 2002, Yahoo shifted to crawler-based listings for its main results. [2]

In addition to excellent search results, users can click on tabs above the search box on the Yahoo home page to seek images, Yellow Page listings or use Yahoo's excellent shopping search engine. Users can visit the Yahoo Search home page, where even more specialized search options are offered. The Yahoo Directory still survives, however. When offered, these will take

the researcher to a list of web sites that have been reviewed and approved by a human editor. It's also possible to do a pure search of just the human-compiled Yahoo Directory. To do this, search from the Yahoo Directory home page, as opposed to the regular Yahoo.com home page.

Sites pay a fee to be included in the Yahoo Directory's commercial listings, though they must meet editor approval before being accepted. Non-commercial content is accepted for free. Yahoo's content acquisition program also offers paid inclusion, where sites can also pay to be included in Yahoo's crawler-based results. The CAP program also brings in content from non-profit organizations for free. Like Google, Yahoo sells paid placement advertising links that appear on its own site and which are distributed to others. These are sold through Overture.

2.2.3. Ask Jeeves

Ask Jeeves initially gained fame in 1998 and 1999 as being the "natural language" search engine that lets the user search by asking questions and responded with what seemed to be the right answer to everything.

In reality, technology wasn't behind Ask Jeeves performance success. Behind the scenes, the company at one point had about 100 editors who monitored search logs. They then went out onto the web and located what seemed to be the best sites to match the most popular queries.[9]

Today, Ask Jeeves instead depends on crawler-based technology to provide results to its users. These results come from the Teoma search engine that it owns.

Ask Jeeves is doing innovative things with invisible tabs and with what it calls Smart Search. Ask Jeeves also owns now closed Direct Hit service.

2.2.4. AllTheWeb.com

AllTheWeb is powered by Yahoo; however the search is more customizable and entertaining search. In addition to web search, news, picture, video, MP3 and FTP search are also offered.

2.2.5. AOL Search

AOL Search provides users with editorial listings that come with Google's crawler-based index. The "internal" version of AOL Search provides links to content only available within the AOL online service. In this way, you can search AOL and the entire web at the same time. The "external" version lacks these links. Many of Google's features such as "cached" pages are not offered by AOL Search and thus AOL is only preferred by registered AOL users.

2.2.6. HotBot

HotBot provides easy access to the web's three major crawler-based search engines: Yahoo, Google and Teoma. Yet it is not a meta search engine and thus cannot blend the results from all of these crawlers together. Nevertheless, it's a fast, easy way to get different web search "opinions" in one place.

2.2.7. Teoma

Teoma is a crawler-based search engine owned by Ask Jeeves. Google and Yahoo, their rival crawler-competitors, have a larger index of the web. Teoma's high relevancy however with respect to popular queries does put it at an advantage as of that of Google and Yahoo. Its "Refine" feature offers

suggested topics to explore after a search is much sought for by some searchers. The "Resources" section of results is also unique, pointing users to page that specifically serve as link resources about various topics. Teoma also provides some results to the Toema web site.

2.2.8. AltaVista

AltaVista opened in December 1995 and for several years was the "Google" of its day, in terms of providing relevant results and having a loyal group of users that loved the service.

However, an attempt to turn AltaVista into a portal site in 1998 saw the company lose track of the importance of search. Over time, relevancy dropped, as did the freshness of AltaVista's listings and the crawler's coverage of the web.

Today, AltaVista is once again focused on search. Results come from Yahoo, and tabs above the search box let you go beyond web search to find images, MP3/Audio, Video, human category listings and news results.

2.3. Experimental Comparison

The analysis below [8] indicates that the search wars are wide open. Google currently provides the highest quality content and best matches to the search query in its results, but is only slightly ahead of Yahoo!'s search engine. MSN's beta release trails Yahoo! by a small amount in the numbers. From the results, it also appears that both Yahoo! and MSN are sorely lacking the index size that Google has. This competitive advantage may be the difference in Google's ability to provide more relevant results.

An analysis of the PageRank & measurable backlinks to the sites reveals similar trends between the 3, but Google's massively larger index and ability to crawl more web pages more frequently may be the key to their current success. Yahoo! and MSN still have only a fraction of the pages in cache that Google does. A rough guess based on the results would be that Yahoo!'s index is approximately 1/2 the size of Google's, while Microsoft's is probably less than 1/4 Google's size.

The measures of content quality and match quality were made as objectively as possible and show that while Google on average provides the best results, there is still a very long way to go towards providing an excellent search engine for the web. The search term "HDTV Comparison", for example, produced not 1 result among 30 (the top 10 results at each search engine) that I considered excellent.

Spamming each of the search engines is still surprisingly easy. Many results at all 3 search engines were clearly little more than traffic grabbing pages with no real, unique content. The search engine that deals with this issue most effectively will gain a significant advantage for the future, as sites & pages like these multiply on the web into the millions.

The key to short-term success for both Yahoo! & MSN will be finding a way to build their index fast enough to compete with Google's giant head-start.

SEO Analysis [8]

For an SEO professional, the results below indicate that Google's trend of favouring "authority" sites as defined by the hilltop algorithm continues unabated. With the exception of several odd standouts, Google, Yahoo! & MSN all give great credence to sites that have many thousands of backlinks, a high PageRank & a large number of pages on the site. The era of small, niche sites is falling to the wayside as the search engines look for the dominance of major web presences like news organizations, government websites & large commercial endeavors.

For MSN specifically, the results are very hard to pick apart. It appears that MSN may be using some additional pieces of ranking technology in their algorithm, possibly further questioning the value of non-relevant links. However, the current size of their index in comparison with Yahoo! & Google makes it exceptionally difficult to determine which pieces they consider important and which pieces are simply left out because they have not yet been spidered. The only piece which stuck out in my analysis, although it was not included in my statistics, is that MSN does not appear to have Google's marked preference for older sites.

Survey Data & Key

This survey comprises four relatively common searches at the three major search engines, including MSN's new Beta search engine release. Each search was conducted with the default settings at each of the following URLs:

http://beta.search.msn.com

http://www.google.com

http://search.yahoo.com

The tables contain five measurements for each result in the SERPs.[8]

1. PR - This is the Google PageRank according to the toolbar for the homepage of the site
2. YLinkD - This is the number of results according to a query run at Yahoo! for the number of links to the top level domain (linkdomain:url.com -site:url.com).
3. MSNLink - This is the number of results according to a query run at MSN's Beta Search engine for the number of links to the top level domain (link:url.com).
4. Size - This value is taken from a search at Google for the number of pages in the index (site:url.com).
5. Qual - This is a subjective measure of the quality of the page's content. The score is X/10 based on the terms listed in the quality of content criteria chart.
6. Match - This is a subjective measure of the closeness of the match in the site's content to the search terms entered in the query.

Table 2. Survey comparing msn, Yahoo!, and Google [8]

msn Search (beta)	PR	YLinkD	MSNLink	Size	Qual	Match
MSN Beta Average for 4 Searches	5.0	907,704	110,017	58,997	5.4	4.2
US National Holidays - *1303 searches in October**						
Average	5.4	32,868	6532	46,803	5.7	3.3
1. http://www.villas4all.com/american-holidays.php	5	1200	1924	754	5/10	6/10
2. http://www.discoverfrance.net/France/DF_holidays.shtml	6	2600	3484	760	7/10	2/10
3. http://usinfo.state.gov/usa/infousa/facts/factover/holidays.htm	8	245,000	52,055	231,000	7/10	6/10
4. http://falcon.jmu.edu/~ramseyil/holidays.htm	6	76,600	159	158,000	7/10	4/10
5. http://www.prttravel.com/National_Holidays.htm	3	334	765	24	5/10	4/10
6. http://www.asiatraveltips.com/NationalHolidays.shtml	6	1,170	1887	25,500	3/10	1/10
7. http://www.nzstay.com/new-zealand-holidays.htm	5	686	962	370	6/10	2/10
8. http://www.allhotelsitaly.com/italy/holidays.htm	5	172	1668	171	4/10	2/10
9. http://www.us-parks.com/	5	898	2285	51,200	7/10	1/10
10. http://www.granisle.com/holidays.html	5	24	138	256	6/10	5/10

HDTV Comparison - *619 searches in October**

Average	3.8	3039	1604	91,948	4.2	3.7
1. http://www.bigpicture-hdtv.com/compare.html	5	1560	963	14	6/10	4/10
2. http://www.about-hdtv.com/hdtv-comparison.htm	0	5	271	83	2/10	4/10
3. http://www.about-hdtv.com/	0	5	271	83	2/10	1/10
4. http://www.ee.washington.edu/conselec/CE/kuhn/hdtv/95x5.htm	7	26,600	2672	916,000	9/10	5/10
5. http://www.twcsc.com/hdtv.htm	5	380	2244	293	4/10	4/10
6. http://www.eegent.com/hfeature.htm	6	57	124	87	4/10	4/10
7. http://www.twckc.com/services/hdtv/compare.asp	5	380	1359	905	4/10	4/10
8. http://satellitesense.com/	3	15	1515	556	4/10	1/10
9. http://www.plasmatvbuyingguide.com/...html	5	1310	6627	611	7/10	7/10
10. http://www.onlinemarketingmall.com/lcd-tv/hdtv-comparison.htm	2	86	2	851	0/10	3/10

George Plimpton - *593 searches in October**

Average	5.7	281,787	36,040	48,420	5.6	3.4
1. http://www.siteway.com/illustrations_georgeplimpton.php	5	783	285	533	6/10	5/10
2. http://www.parisreview.com/audio/audio.htm	7	5140	1542	84	6/10	4/10
3. http://www.gadflyonline.com/archive-plimpton.html	4	579	120	519	6/10	3/10
4. http://time.com/time/community/...plimptontime100.html	8	848,000	105,158	190,000	6/10	3/10
5. http://www.pagitica.com/extras/plimptondialog.html	4	123	186	85	6/10	3/10
6. http://www.historybound.com/...Code=BIOHISTMEM	4	29	3	255	4/10	2/10
7. http://www.lectures.org/plimpton.html	6	253	457	198	6/10	5/10
8. http://www.salon.com/audio/2000/10/04/johnson_plimpton/	8	1.93mil	238,795	279,000	6/10	3/10
9. http://www.milkmag.org/plimpton.htm	5	465	364	227	4/10	4/10
10. http://www.museumofhoaxes.com/siddfinch.html	6	32,500	13,494	13,300	6/10	2/10

Order Digital Prints Online - *184 searches in October**

Average	5.1	3.31mil	1891	48,817	6	6.5
1. http://www.chatsworthsalesphoto.com/	2	9	88	102	7/10	8/10
2. http://www.chatsworthsalesphoto.com/digital	2	9	88	102	7/10	6/10
3. http://www.freewebs.com/digital-photo-prints/	7	1.62mil	7754	196,000	2/10	6/10
4. http://www.picturetrail.com/webpages/makeprints.shtml	6	311,000	7780	53,400	6/10	8/10
5. http://www.adobe.com/digitalimag/tips/phsalonlsvc/main.html	10	15.6mil	1.97mil	119,000	7/10	7/10

6. http://www.adobe.com/digitalimag/photoshop_services.html	10	15.6mil	1.97mil	119,000	6/10	5/10
7. http://www.edmonsonweddings.com/Shutterfly-Order-Prints.php	3	89	2274	351	6/10	5/10
8. http://www.greatnationalcamera.com/	4	69	866	160	6/10	6/10
9. http://www.photocounter.ca/digital/online.html	3	6	3	10	6/10	7/10
10. http://www.cboplmaging.com/orderonline.html	4	20	60	47	7/10	7/10

Google

	PR	YLinkD	MSNLink	Size	Qual	Match
Google Average for 4 Searches	6.9	1.94mil	176,191	474,746	5.8	6.2
US National Holidays - *1303 searches in October**						
Average	6.4	83,673	12,469	43,825	6.2	6.7
1. www.villas4all.com/american-holidays.php	5	1200	1924	754	5/10	6/10
2. aa.usno.navy.mil/faq/docs/holidays.html	8	356,000	15,395	1.09mil	7/10	8/10
3. usinfo.state.gov/usa/infousa/facts/factover/holidays.htm	8	245,000	52,055	233,000	8/10	8/10
4. www.vpcalendar.net/Holiday_Dates/2000_2005.html	7	1410	1525	80	7/10	8/10
5. www.madmanmike.com/us_events_dates.html	5	196	307	149	8/10	8/10
6. www.opm.gov/fedhol/	9	178,000	43,824	184,000	7/10	9/10
7. www.smart.net/~mmontes/ushols.html	7	27,500	523	981	8/10	8/10
8. www.usflag.org/flag.holidays.html	6	25,700	7867	92	7/10	6/10
9. www.maladyspoetry.com/poetry1.htm	4	174	349	199	5/10	6/10
10. www.global7network.com/.../russian_holidays.asp	5	1550	927	19,000	0/10	0/10
HDTV Comparison - *619 searches in October**						
Average	7	3.06mil	61,832	96,721	5.9	5.3
1. www.plasmatvbuyingguide.com/dtv-hdtv-comparison19.html	5	1310	6627	611	7/10	7/10
2. www.bigpicture-hdtv.com/compare.html	5	1560	942	14	6/10	4/10
3. www.mysimon.com/HDTV_Sets/4007-6485_8-0.html	9	8.88mil	110,124	386,000	6/10	5/10
4. www.mysimon.com/...4007-6487_8-0.html	9	8.88mil	110,124	386,000	6/10	4/10
5. www.proactiveelectronics.com/...Hdtv	5	1470	573	7090	3/10	3/10
6. www.epinions.com/Flat_Panel_Televisions-Plasma-HDTV	7	1.81mil	43,626	4.79mil	6/10	7/10
7. www.epinions.com/...hdtv_ready	7	1.81mil	43,626	4.79mil	6/10	7/10

8. www.howstuffworks.com/hdtv.htm	8	517,000	211,041	174,000	6/10	5/10
9. shopper.cnet.com/...orderby%3D90%26sort%3D	8	8.7mil	77,081	1.97	6/10	7/10
10. www.directv.com/DTVAPP/learn/FAQ_DTVAdvanced_HDTV.jsp	7	61,500	14,556	13,500	7/10	4/10

George Plimpton - *593 searches in October**

Average	7.9	2.13mil	273,972	274,501	5.5	5.1
1. www.georgeplimpton.com/	4	113	138	11	0/10	3/10
2. www.imdb.com/name/nm0687321/	9	4.93mil	433,689	1.79mil	7/10	5/10
3. www.nytimes.com/2003/09/26/obituaries/26CND-PLIM.html	9	4.79mil	700,428	520,000	8/10	7/10
4. www.randomhouse.com/boldtype/1297/plimpton/	7	273,000	28,144	270,000	6/10	4/10
5. en.wikipedia.org/wiki/George_Plimpton	8	2.88mil	147,695	1.62mil	6/10	8/10
6. www.time.com/...plimptontime100.html	8	848,000	103,092	190,000	6/10	3/10
7. espn.go.com/page2/s/thompson/030929.html	9	1.76mil	176,402	946,000	6/10	5/10
8. www.npr.org/display_pages/features/feature_1447605.html	8	845,000	247,031	187,000	5/10	5/10
9. www.npr.org/display_pages/features/feature_1406696.html	8	845,000	247,031	187,000	5/10	4/10
10. www.usatoday.com/.../2003-09-26-plimpton-obit_x.htm	9	4.17mil	656,078	445,000	6/10	7/10

Order Digital Prints Online - *184 searches in October**

Average	6.3	2491990	45,488	32,935	5.7	7.7
1. www.shutterfly.com/learn/order_digital_prints.jsp	7	180,000	165,681	23,200	6/10	8/10
2. www.shutterfly.com/learn/digital_prints.jsp	7	180,000	165,681	23,200	6/10	8/10
3. www.microsoft.com/.../orderprints.mspx	10	24mil	3.11mil	1.52mil	6/10	8/10
4. www.colormailer.com/	5	29,500	1161	735	5/10	8/10
5. www.icdphotos.com/	5	186	547	5340	5/10	7/10
6. www.chebucto.ns.ca/~rakerman/digiphoto.html	6	77,900	5025	54,200	9/10	9/10
7. www.bonusprint.co.uk/pages/digital_prints.htm?OnpageA=1	5	192	156	190	5/10	7/10
8. photos.wanadoo.co.uk/	6	136,000	9032	40,100	5/10	7/10
9. www.mcbaincamera.com/digitalprints.htm	4	117	144	388	4/10	7/10
10. www.kodak.com/eknec/PageQuerier.jhtml?...	8	316,000	107,457	182,000	6/10	8/10

YAHOO!	PR	YLinkD	MSNLink	Size	Qual	Match
Yahoo! Average for 4 Searches	6.5	5.6mil	198,358	807,658	5.3	5.3
US National Holidays - *1303 searches in October**						
Average	6.1	3.07mil	20,586	27,205	5	4.6
1. www.villas4all.com/american-holidays.php	5	1200	1898	754	5/10	6/10
2. www.usemb.se/Holidays/celebrate	6	17,200	333	3520	5/10	6/10
3. usinfo.state.gov/usa/infousa/facts/factover/holidays.htm	8	245,000	52,162	233,000	8/10	8/10
4. www.patriotism.org/page2.html	6	879	123	35	5/10	5/10
5. www.asiatraveltips.com/NationalHolidays.shtml	6	1170	1929	25,400	3/10	1/10
6. patriotic.kith.us/holidays.html	3	139	7	243	5/10	4/10
7. www.timechange.com/us_holidays.htm	6	702	308	8910	5/10	3/10
8. www.geocities.com/geraljo1/us_nationalholidays.htm	8	28.6mil	129,672	2.45mil	4/10	2/10
9. www.h1bresources.com/html/nationalholidays.shtml	5	524	433	189	5/10	6/10
10. dir.yahoo.com/Society_and_Culture/Holidays_and_Observances	8	1.86mil	19,000	3.15mil	5/10	5/10
HDTV Comparison - *619 searches in October**						
Average	5.8	1.61mil	10,464	95,057	4.6	4
1. http://www.nextag.com/...1400000	7	67,200	11,074	1.2mil	6/10	5/10
2. www.shopping.com/xGS-hdtv_comparison_brands	8	4.07mil	31,414	949,000	3/10	3/10
3. www.plasmatvbuyingguide.com/dtv-hdtv-comparison19.html	5	1310	6627	611	7/10	7/10
4. www.bigpicture-hdtv.com/compare.html	5	1560	942	14	6/10	4/10
5. www.digitalmegamall.com/hdtv.htm	3	36	4	21	4/10	5/10
6. http://www.proactivewm.com/...&subcat=HDTV	4	697	112	34	3/10	4/10
7. amdzone.pricegrabber.com	7	5.45mil	17,438	3.28mil	5/10	1/10
8. www.twckc.com/services/hdtv/compare.asp	5	583	1328	807	4/10	4/10
9. www.dealtime.com/xGS-hdtv~FD-0~DL-0~SK-0~CLT-DNL01	8	6.55mil	35,575	1.43mil	4/10	3/10
10. www.eegent.com/hfeature.htm	6	57	128	87	4/10	4/10
George Plimpton - *593 searches in October**						
Average	8	13.7mil	412,880	176,250	6.7	5.7
1. www.nytimes.com/2003/09/27/obituaries/27PLIM.html	9	4.91mil	700,428	520,000	8/10	7/10

2. us.imdb.com/Name?Plimpton,+George	8	2.17mil	433,689	1.79mil	7/10	5/10
3. www.newyorker.com/talk/content? 031006ta_talk_remnick	7	334,000	37,614	24,500	8/10	7/10
4. www.randomhouse.com/boldtype/1297/plimpton	7	274,000	28,144	270,000	6/10	4/10
5. slate.msn.com/id/2089012	8	956,000	87,998	145,000	9/10	7/10
6. www.npr.org/display_pages/features/feature_1447605.html	8	853,000	247,031	187,000	5/10	5/10
7. http://www.amazon.com/...2801400	8	60.2mil	823,927	5.81mil	6/10	5/10
8. www.amazon.com/exec/obidos/ASIN/0871135035	8	60.2mil	823,927	5.81mil	6/10	4/10
9. www.usatoday.com/.../2003-09-26-plimpton-obit_x.htm	9	4.25mil	656,078	445,000	6/10	7/10
10. www.msnbc.com/news/972260.asp?0cv=CB20	8	2.99mil	289,973	171,000	6/10	6/10
Order Digital Prints Online - *184 searches in October**						
Average	6.2	4.03mil	38,499	11,118	5	6.7
1. www.ofoto.com	6	141,000	11,749	52,600	5/10	7/10
2. www.shutterfly.com/about/com_overview.jsp	7	177,000	165,681	23,200	6/10	7/10
3. http://www.microsoft.com/.../orderprints.mspx	10	23.8mil	3.11mil	1.52mil	6/10	8/10
4. www.shutterfly.com	7	177,000	177,000	23,200	6/10	7/10
5. www.photoaccess.com	6	16,800	1774	1050	4/10	4/10
6. www.filmworks.com	7	1370	4675	485	5/10	7/10
7. online-photo-albums.bousperlavila.com	X	701	5	X	1/10	7/10
8. www.snapfish.com	7	112,000	11,814	10,600	5/10	7/10
9. www.cbopimaging.com/orderonline.html	4	18	60	47	7/10	7/10
10. store.yahoo.com/lindenhillsphoto/ordiprandhog.html	8	15.9mil	12,241	2.77mil	5/10	6/10

* # of Searches was taken from Overture's Search Term Suggestion Tool

2.4. Which powers which?

Some search engines in addition to running their own search engine sites may turn to third party search providers for their listings. This maybe confusing for those who wonder which companies are the major competitors in the powering market search. For this reason I have included the below chart to explain the major providers.[10]

Chart Key

Search Providers: These are listed at the top of each column.

Search Engines: These are listed at the beginning of each row, in order of share of searches;

- Dark Orange: search engines with 30 percent or greater share.
- Light Orange: search engines with 15 percent or greater share.
- Light Blue: search engines with 0.5 percent share or greater share.
- Gray: search engines with less than 0.5 percent share. They are shown only because of the name recognition they may have among serious searchers.

Main: Indicates that a search provider provides the "main" editorial results to a particular search engine, the most dominant listings that will be seen.

Paid: Indicates that a search provider provides paid placement listings to a particular search engine.

Backup: Indicates that a search provider provides the "backup" results that appear in cases where a search engine's main results fail to find good matches.

Option: If shown in the notes section, Indicates that information from this source is made available either on results pages or in other ways, though the prominence of the information may not be high.

Dates: Where shown, dates indicate when a particular partnership is due for renewal. Dates are shown in MM/DD/YY or similar format.

Table 3. Chart showing which engine powers which [10]

Search Engine (Read Down)	Provider: Google	Provider: Yahoo/Overture	Notes
Google	Main & Paid		Open Directory an option
Yahoo		Main & Paid	
MSN		Main & Paid (12/05 & 6/05)	LookSmart an option on home page
AOL	Main & Paid (est. 10/05+)		AOL-owned Open Directory an option
Ask Jeeves	Paid (9/05)		Main from Ask-owned Teoma. Paid can end as early as 9/04
InfoSpace	Runs several meta search engines. Dogpile is most popular, representative of others. Google (2006), Yahoo (3/06), many small providers have distribution deals.		
Lycos	Paid	Backup	Main from LookSmart; Open Directory an option
AltaVista		Main & Paid	Open Directory an option; owned by Yahoo

AllTheWeb		Main & Paid	Owned by Yahoo
HotBot	Paid	Main	Backup from Google & Ask; Owned by Lycos
Netscape	Main & Paid (est. 10/05+)		Owned by AOL; Open Directory an option
Teoma	Paid (Sept 05)		Main from Teoma; owned by Ask; Paid can end as early as 9/04
LookSmart	LookSmart provides its own Main & Paid		

2.5: Some Limitations of a General search Engine

There is not one Web Search Tool that comes without limitations and this is why an overview of these should be considered;

- Search spiders or crawlers, as explained, do not crawl the Web in real time. The major search services are improving turnaround on recrawling and adding pages, but in general, expect to wait many days before a keyword search will return a recent page.
- If a site or page is not linked to or submitted by someone (Webmaster, page author, etc.), it will not be accessible from a search engine.

- Simply because one, 1,000, or even more pages from a site are available does not mean that the engine makes every page of an entire site searchable.[2][9]

2.6. Specialized Search Engines?

General search engines may not handle files that are in the .pdf format in contrast to the specialized Web crawlers. This is why users should turn to these effective tolls in utilizing the open Web. Well-known specialized Web search engines include Psychcrawler, PoliticalInformation.Com, and Inomics.Com, each of which focuses on a specific subject (psychology, political science, and economics, respectively). Site-specific engines refer to the search engines that many sites make available to cover their own material.

The general search tools often do crawl material that can also be found using a specialized, focused, and site-specific search engine. However, in some cases, the general search engines may not cover this material as well as the specialized ones. Coverage of this material by general search engines like Excite or AllTheWeb may be spottier than the specialized search tools;[2]

- Time Lag. Spiders visit pages unannounced. Material changed or added between the dates when the spider last crawled the content remains invisible. News material is a good illustration. A normal page from the CNN site is technically searchable from any general-purpose engine being unsearchable for a period of time before.
- Depth of Crawl. Simply because a search engine makes a certain number of pages of a site accessible does not mean that it has crawled the entire site. Some engines only take a certain amount of material and then move on.
- Each Search Engine Database Is Unique. As the work of Greg Notess makes clear, each search engine database differs. What one search

engine may have in its database another search engine may lack and what one can access another search engine cannot.

- Dead-End Pages. A site-specific engine can crawl every page sitting on an entire server and make the page searchable. This is in contrast with a normal search tool which will not find for example a basic HTML page on a server not linked from any other page and not submitted.

Bottom line: several reasons encourage a user to go for specialized searches:

1. Smaller, more targeted databases make for greater precision though lower recall.
2. These resources often offer human interaction, with a knowledgeable editor telling the crawler where to go, how often to return, and how deep to crawl.
3. Finally, some of these specialized engines provide extra functionality, such as constant, even daily, updating and limiting options for search strategies.

In Google, the web crawling is done by several distributed crawlers. The crawlers fetch URL lists sent by a URLserver. The web pages that are fetched are then sent to the storeserver to which they are compressed and stored into repository. Every web page has an associated ID number called a docID,an ID number associated with every web page,is assigned whenever a new URL is parsed out of a web page. The indexer then reads the repository, uncompresses the documents, and parses them. Each document is converted into a set of word occurrences called hits. The hits record the word, position in document, an approximation of font size, and capitalization. These hits are then arranged into set of "barrels", creating a partially sorted forward index. The indexer performs another important function. It parses out all the links in every web page and stores important information about them in an anchors file. This file containing information determines where each link points from and to, and the text of the link.[3]

Relative URLs are then converted into absolute URLs and in turn into docIDs. This is possible after the URLresolver reads the anchor file. The URLresolver puts the anchor text into the forward index. The forward index is associated with the docID that the anchor points to. Pairs of docIDs are then generated into a database of links used to compute PageRanks for all the documents.

In a simple overview, the sorter takes the barrels, which are sorted by docID and resorts them by wordID generating the inverted index. Little temporary space is needed for this operation knowing that it is done in place. The sorter also produces a list of wordIDs and offsets into the inverted index. A program called DumpLexicon generates a new lexicon to be used by the searcher taking this wordIDs list together with the lexicon produced by the indexer.

Using the lexicon built by DumpLexicon together with the inverted index and the PageRanks, the searcher run by a web server answers queries.

Major Data Structures

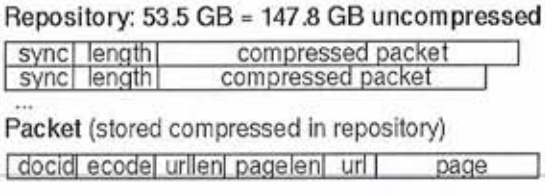
A little cost is what is needed for a large data document to be crawled, indexed, and searched since Google's data structures are optimized. A disk seek requires about 10 ms to complete so Google is designed to avoid disk seeks whenever possible. This yields a considerable influence on the design of the data structures. [3]

BigFiles

BigFiles are virtual files spanning multiple file systems (allocated among each other automatically) and are addressable by 64 bit integers. Since the operating systems do not provide enough for allocation and deallocation of file descriptions, the BigFiles package handles this task . BigFiles also support rudimentary compression options.

Repository

The repository contains the full HTML of every web page. Each page is compressed using zlib. Zlib's speed over a significant improvement in compression offered by bzip was behind the choice. The compression rate of bzip was approximately 4 to 1 on the repository as compared to zlib's 3 to 1 compression. In the repository, the documents are stored one after the other and are prefixed by docID, length, and URL. Data consistency and development is much easier since the repository requires no other data structures to be used in order to access it. We can rebuild all the other data structures from only the repository and a file which lists crawler errors.



Repository Data Structure adapted from Barroso,Dean et al.

Document Index

The document index is a fixed width ISAM (Index sequential access mode) index, ordered by docID and keeps information about each document. The information stored in each entry includes the current document status, a pointer into the repository, a document checksum, and statistics. If the document has been crawled, it also contains a pointer into a variable width file called docinfo which contains its URL and title. Otherwise the pointer points into the URLlist which contains just the URL. This design decision leads to a reasonably compact data structure and to the ability to fetch a record in one disk seek during a search.

A file which is a list of URL checksums with their corresponding docIDs and is sorted by checksum is used to convert URLs into docIDs. The URL's checksum is computed and a binary search is performed on the checksums file to find its docID. URLs may be converted into docIDs in batch by doing a merge with this file, a technique the URLresolver uses to turn URLs into docIDs. This batch mode of update is crucial because otherwise we must perform one seek for every link which assuming one disk would take more than a month for our 322 million link dataset.

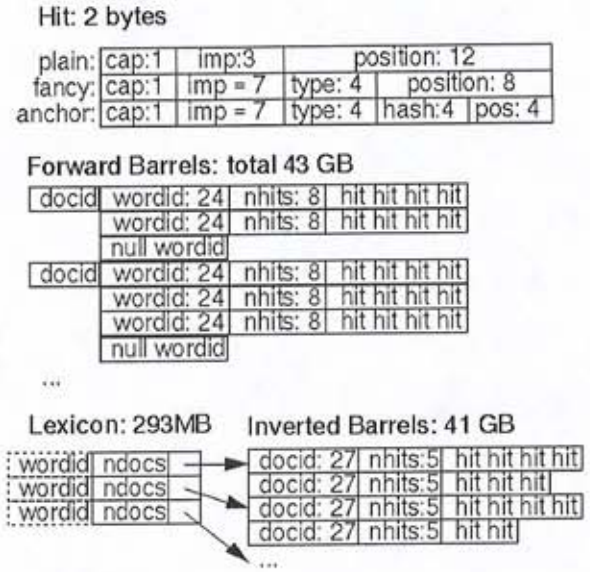
Lexicon

The current Lexicon fit in memory for a reasonable price. The Lexicon in memory can be kept on a 256MB main memory machine. It contains 14 million words. It is implemented in two parts, a list of concatenated together words, separated by nulls with some auxiliary information and a hash table of pointers.

Hit Lists

A hit list corresponds to a list of occurrences of a particular word in a document including such information such as position, font and capitalization. Representing hit lists efficiently is very important for they account for the majority of space used in forward and inverted indexes. There are several alternatives for encoding position, font and capitalization: simple encoding (a triple of integers), a compact encoding (a hand optimized allocation of bits), and Huffman coding. The best is the hand optimizing compact encoding for it requires less space than the simple encoding and less bit manipulation than Huffman manipulation.

The compact encoding uses two byes for every hit. We have two types of hits: fancy and plain hits. Fancy hits are those occurring in a URL, title, anchor text or meta tag. Plain hits consist of a capitalization bit, font size, and 12bits of word position in a document. Font size is represented relative to the rest of the document using three bits. We use font size relative to the rest of the document because when searching, you do not want to rank otherwise identical documents differently just because one of the documents is in a larger font. A fancy hit consists of a capitalization bit, the font size set to 7 to indicate it is a fancy hit, 4 bits to encode the type of fancy hit, and 8 bits of position. For anchor hits, the 8 bits of position are split into 4 bits for position in anchor and 4 bits for a hash of the docID the anchor occurs in.



Forward and Reverse Indexes and the Lexicon adapted from Barroso,Dean et al.

The length of the hit list is combined with the worded in the forward index and the docID in the inverted index paving the way for saving space and thus limiting it to 5 and 8 bits respectively. An escape code is used in those bits if the code is longer than would fit in that many bits. The next two bytes contain the actual length.

Forward Index

The forward index is stored in a number of barrels to which each of the barrels holds a range of wordID's. Instead of sorting actual wordID's, each wordID is is stored as relative difference from the minimum WordID stored in the barrel. Thus we can leave 8 bits for the hit list length since we use just 24 bits for the wordID's in the unsorted barrels.

Inverted Index

Like the forward index, the inverted index consists of the same barrels only here they are processed by a sorter. The lexicon contains a pointer into the barrels that every valid wordID falls into. It points to a doclist of docID's ,representing all the occurrences of that word in all documents, together with

their corresponding hit lists. An important issue here is the order of the docID's in the doclist. The best option is keeping two sets of inverted barrels. One set is for the hit lists that include a title or anchor hits and another for all hit lists. Subsequently, we can check first the first set of barrels and if that yields not enough matches we check the larger ones.[3]

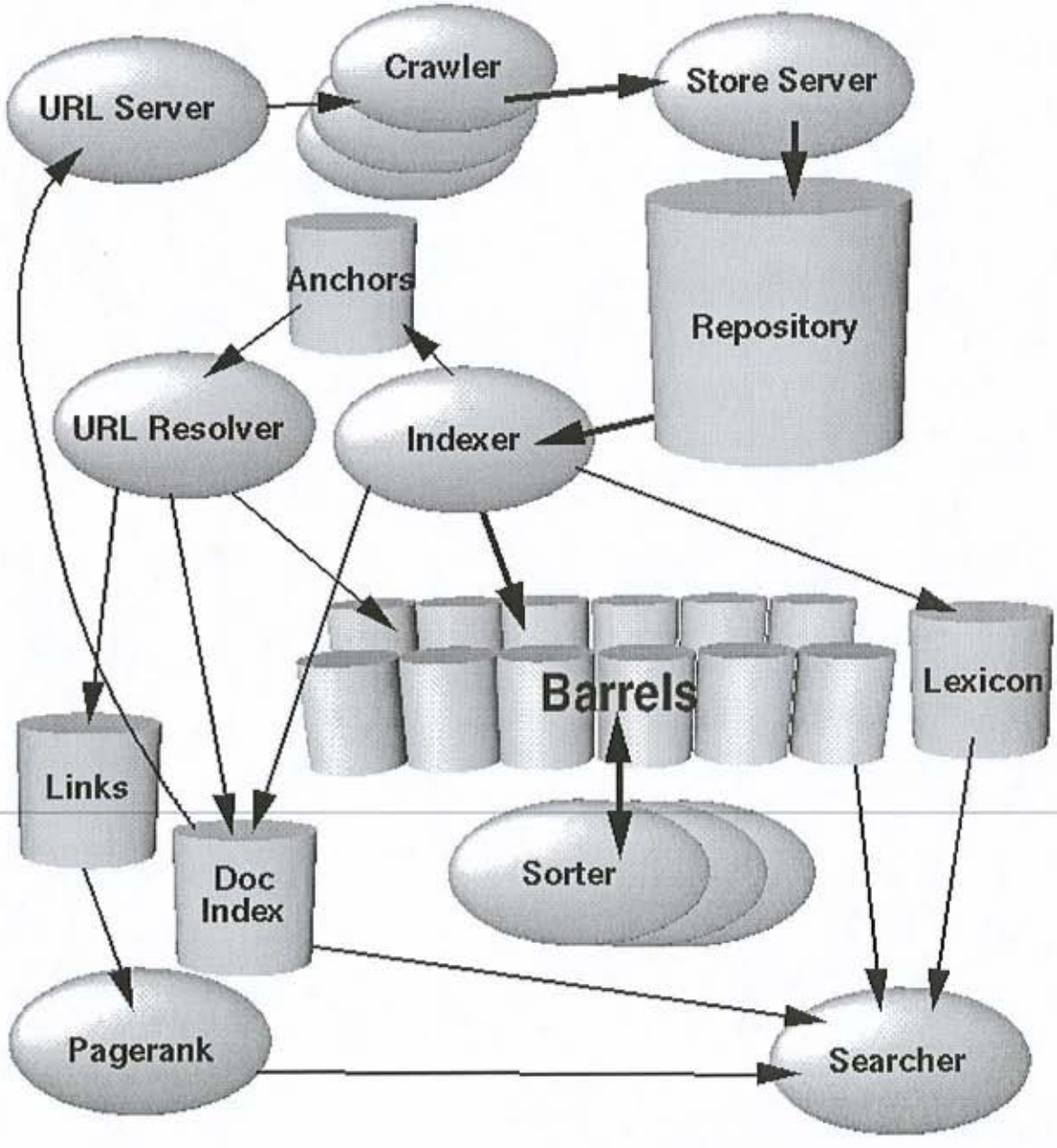


Figure 1. Search engine Architecture Diagram

Chapter 4: Software design of Flying S E

As for the reasons stated before comes the need for a simple specialized search engine. In my project I developed a search engine which can be used as a precise as well as fast and efficient mean to find special pages

4.1. What does it do?

Flying S E is like a directory where human editors organize web sites into categories, users visit the search engine where they have to do registration listing information about themselves and their site including the category, and sub categories. Users can even upload the main HTML pages the engine will take the categories and the other sub-categories from the meta-name. However, the page won't be published until the administrator of the search engine checks the site, and makes sure of the right categorization and the he has to approve it. As soon as the site is approved it will be available for the searchers.

In addition to excellent search results, users can view directly the logo of the site next to the information about it. So in another words, searches will take the researcher to a list of web sites that have been reviewed and approved by a human editor. One important feature of the search engine is the site that will be visited most by the researcher will get a higher rank which means, if a user chooses a certain site from a list of results the site will be listed in first results of the next search, depending on the number of visits the more the site is visited the higher it get in the list.

4.2 Design of Flying S E

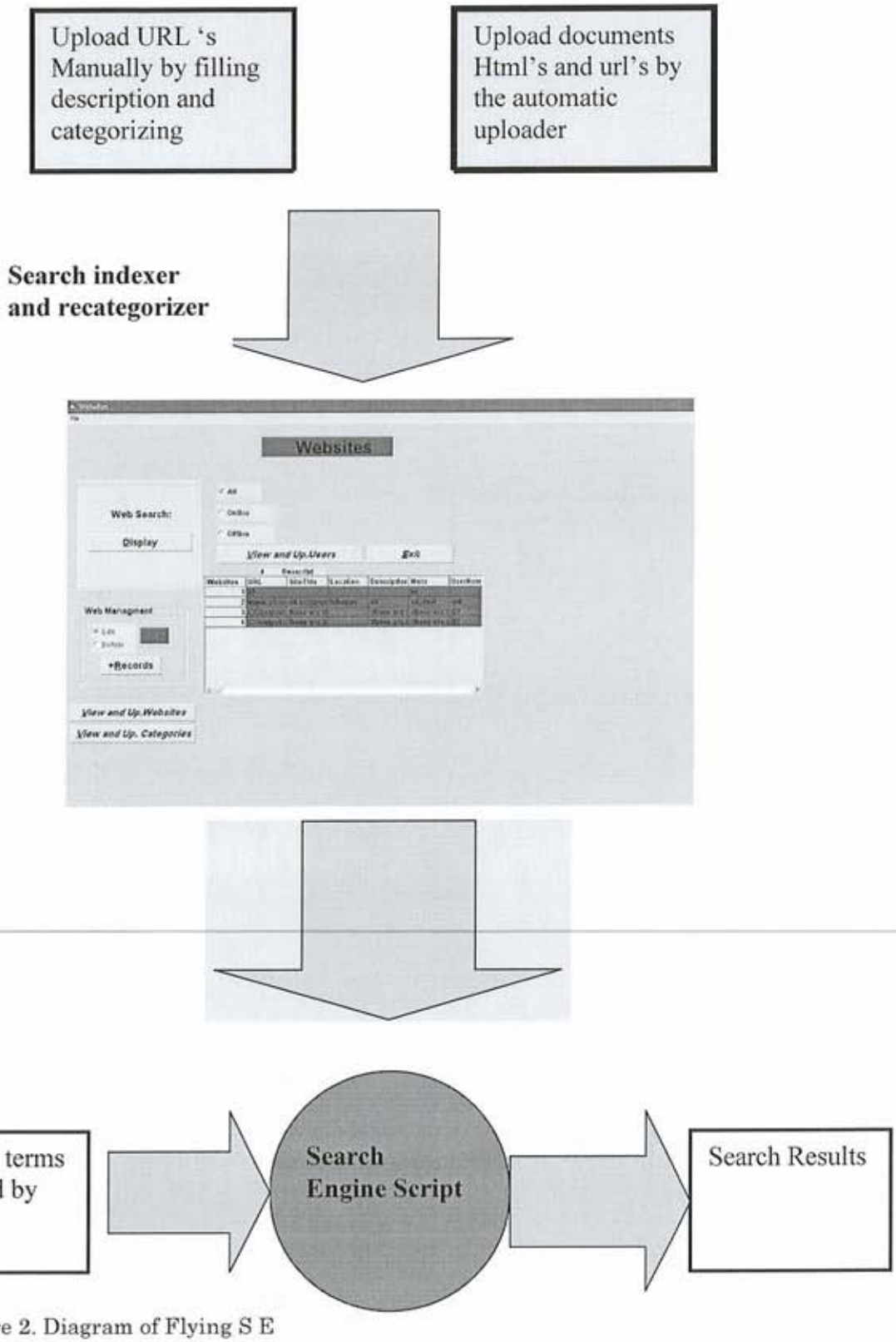


Figure 2. Diagram of Flying S E

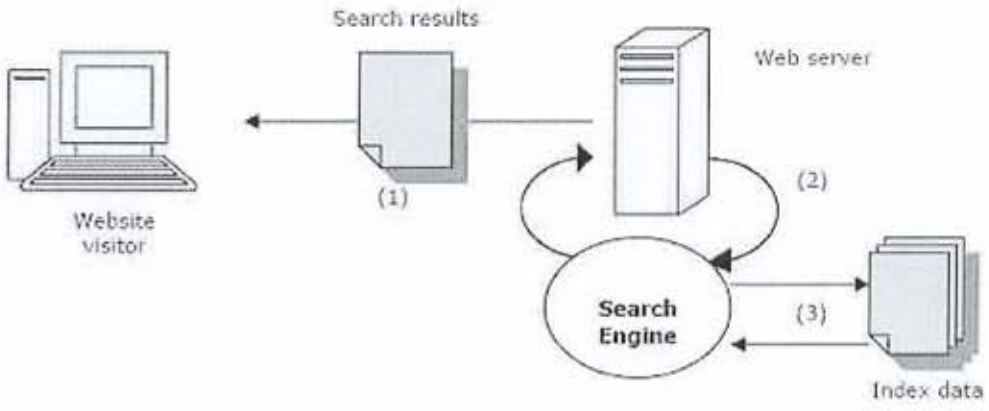
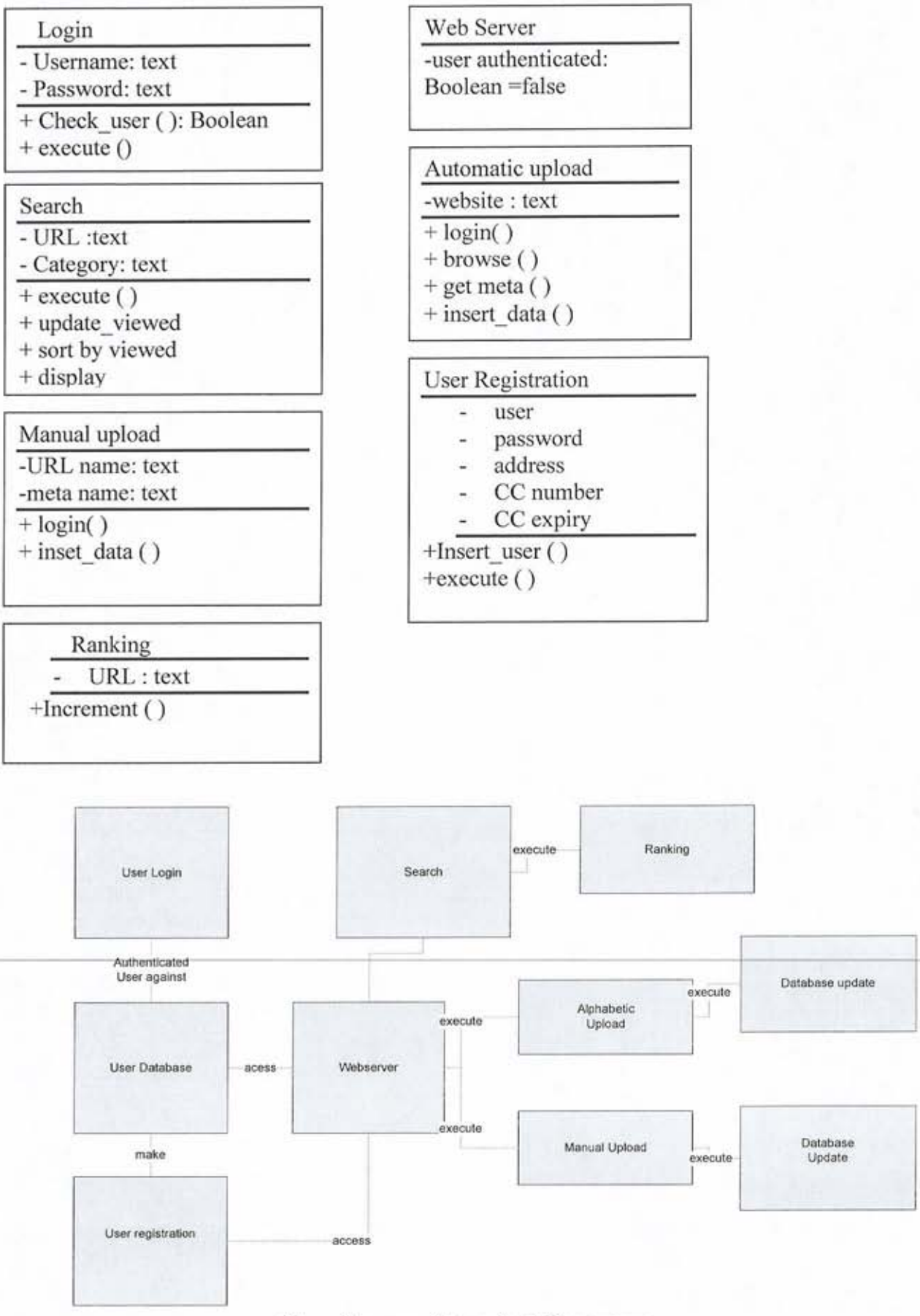


Figure 3. Flying S E Structure

Flying Search Engine works as figure 2 shows. First users have to register to be valid users to be able to upload their websites. The sites are uploaded either by filling manually the URL and the descriptions of the page, or by uploading the pages automatically by the upload function

After the pages are uploaded the administrator checks the webpage the URL name the link, the descriptions, categories, and the subcategories. If the website is approved the site is made available online by the visual basic application.



Class diagram with navigability arrows

Chapter 5: Experimental Results

5.1. Results

In this section we present some experimental results comparing MSN, Yahoo!, Google, and the Flying S E. First I uploaded some sites with data relative to my search to my search engine and did some ranking on them. Like I put 20 pages that has Lebanon in their meta tag as well as related to some different subjects which I planned to use in my queries.

The 5 queries I used are:

1 .Lebanon

2. Civil war

3. Lebanese hotels

4. Best tourist sites in Lebanon

5. Lebanon war lords

And below a list of some of the sites I used in The Flying S E.

www.**lebanon**.com
www.lebanon-**hotels**.com
www.tourismeliban.com/
www.fortunecity.com/meltingpot/stelmo/615/youth.htm
www.middle-east-online.com/english/?id=13196
www.memoryforthefuture.com/english/about
www.theage.com.au/news/World/.../2005/03/04/1109700678652.html
www.csmonitor.com/2002/0125/p07s01-wome.html
www.en.wikipedia.org/wiki/**Lebanese_Civil_War**
www.worldhistory.com/wiki/L/**Lebanese-Civil-War**.htm
www.lexicorient.com/e.o/leb_civ_**war**.htm
www.rimbaud.freeserve.co.uk/lebanon_civ**war**.htm
www.wikimirror.com/**Lebanese_Civil_War**
www.tanbourit.com/lebanon_**war**.htm
www.lonelyplanet.com/**destinations**/middle_east/**lebanon**
www.**lebanon**panorama.com
www.middleeastuk.com/**destinations/Lebanon**
www.xotic**destinations**.com/**Lebanon** Photo Tour.htm
www.finance.gov.lb/main/aboutus/**CountryProfile/countryprofile**.htm
www.un.org/esa/earthsummit/leban-cp.htm

www.cia.gov/cia/publications/factbook/geos/le.html
www.news.bbc.co.uk/2/hi/middle_east/**country_profiles**/791071.stm

Some explanation of the terms we used in the following tables,

Good sites mean sites that are relevant to the search query.

Bad sites mean sites that are not completely relevant to the search query

Duplicate sites are sites which are present more than once.

Table #4: Total Results

Queries	MSN	Yahoo	Google	Flying S E
1	22,417,949	42,700,000	46,100,000	20
2	19,469,609	43,500,000	44,600,000	5
3	2,646,686	4,790,000	3,110,000	7
4	112,034	203,000	298,000	1
5	116,950	113,000	145,000	0

Table #5: Total Good Sites (relative to search) in Top 20

Queries	MSN	Yahoo	Google	Flying S E
1	10	13	9	20
2	20	20	20	3
3	8	8	13	2
4	19	18	20	1
5	8	7	11	0

Table #6: Bad Sites /Duplicates in the Top 20

Queries	MSN	Yahoo	Google	My Search engine
1	10/0	7/0	11/0	0/0
2	0/1	0/0	0/2	2/0
3	7/1	11/1	7/0	5/0
4	0/1	2/0	0/0	1/0
5	12/0	13/0	9/0	0/0

The results are good with general searches. However, Flying Search Engine needs more work because it has many limitations which will be discussed in the following section For instance, when we search for Lebanon, the first query, we notice that we get 20 hits; however, when we try to look for complex queries, the quality and quantity of the results decreases. With respect to the ranking of the pages the ranking method we used depends on the number of times the user opens a certain page. That means every time a page is visited its ranking gets higher by one in the search results so the more searches and visits are done the better the relevance of the results should get.

5.2 Limitations of Flying S E

Flying Search Engine has many limitations. We'll give a list of some of these limitations:

- A page made available on the Web won't be available in the database of the search engine until the administrator approved it.
- Small database since pages are uploaded by interested users no crawler
- Pages might be missed since the user might not include all descriptions of his site.
- Ranking method might be prejudiced because it is controlled by the administrator then people might enter sites that are not relevant to their search so changing the ranking
- Ranking method is not strong enough, not based like Google on how many links point to a certain website or how often it is updated or how important it is; however it is based on the users and the administrator.
- Approving sites, checking the database, uploading site, all is done by users which will demand a lot of time as the database becomes larger. Moreover, it might become nearly impossible to handle and to check dead sites by the administrators which may result in the bad sites.

Chapter 6: Conclusion

Flying Search Engine is a prototype search engine implemented in ASP and Access database. Experimental results show that it is able to do some acceptable results with some simple queries. However it has some limitations for example the limited database; however, that makes it more suitable to be a specialized search engine to search special topics or areas like searching for music, books, universities...

Summing up, Flying Search Engine project is a modest shot at developing a search system specialized for private sectors such as universities and privately-owned companies. Generalizing the program for the public requires years of research in addition to a mega monetary capital. At the time being, the preferred search engine seems to be for most people to be Google followed by Yahoo and the case appears to remain so in the near future for the variety of reasons explained through the paper.

Further work can be done by improving the ASP file uploading algorithm so it can be a smart crawler. That means it uploads the web pages as well as to get the URL's pointed to from these pages to be added to the search engine database.

References :

[1] Abiteboul S., Quass D. , McHugh J., Widom J., and Wiener J., "The Lorel Query Language for Semistructured Data." Technical report, Stanford University, California, 1996.

[2] Baker J., "What Makes a Search Engine Good?" Berkeley University, USA. 2003.

[3] Barroso L.A. , Jeffery D., et al., "Web Search for a Planet: The Google Cluster Architecture." 2400 Bayshore Parkway, Mountain View, CA 94043; USA,2003.

[4] Belkin, N., C. Cool, Croft W. B., Callan, "The effect of multiple query representations on information retrieval system performance," in *Proceedings of the Sixteenth Annual International ACM-SIGIR Conference*. Ed. by Robert Korfhage, Edie Rasmussen and Peter Willett. Pittsburgh, Pennsylvania, June 27-July 1, 1993, p. 339-346

[5] Brin S. and Page L., "The Anatomy of a Large-Scale Hypertextual Web Search Engine." Technical report, Stanford Univeristy, California 1997.

[6] Brin S., Motwani R., Page L., and Winograd T., "What can you do with a Web in your Pocket?" *IEEE Data Engineering*, 21, June 1998.

[7] Buneman P., Davidson S., and Hillebrand G. "A Query Language and Optimization Techniques for Unstructured Data". Technical report, AT&T Research, 1996.

[8] Fishkin Rand , Fishkin Si , "MSN's Beta Search vs. Yahoo! & Google." Seattle, WA 98105.

(http://www.seomoz.org/articles/msn-beta-vs-yahoo-google.php)

[9] Franklin C. , « How Internet Search Engines Work ? »
Johns Hopkins University, Meryland, Baltimore ; USA, 2004

[10] Sullivan Danny, "The Search Engine Report." April 2005.
Http://searchenginewatch.com/sereport/article.php/350894

[11] "The History of Search Engines," (http:/www.galaxy.com/info/history)

[12] "TPC Benchmark C Full Disclosure Report for IBM Server xSeries 440 using Microsoft SQL Server 2000 Enterprise Edition and Microsoft Windows .NET Datacenter Server 2003, TPC-C Version 5.0,"
(http://www.tpc.org/results/FDR/TPCC/ibm.x4408way.c5.fdr.02110801.pdf)

[13] "Search Engines." Microsoft Encarta Encyclopedia 2005.

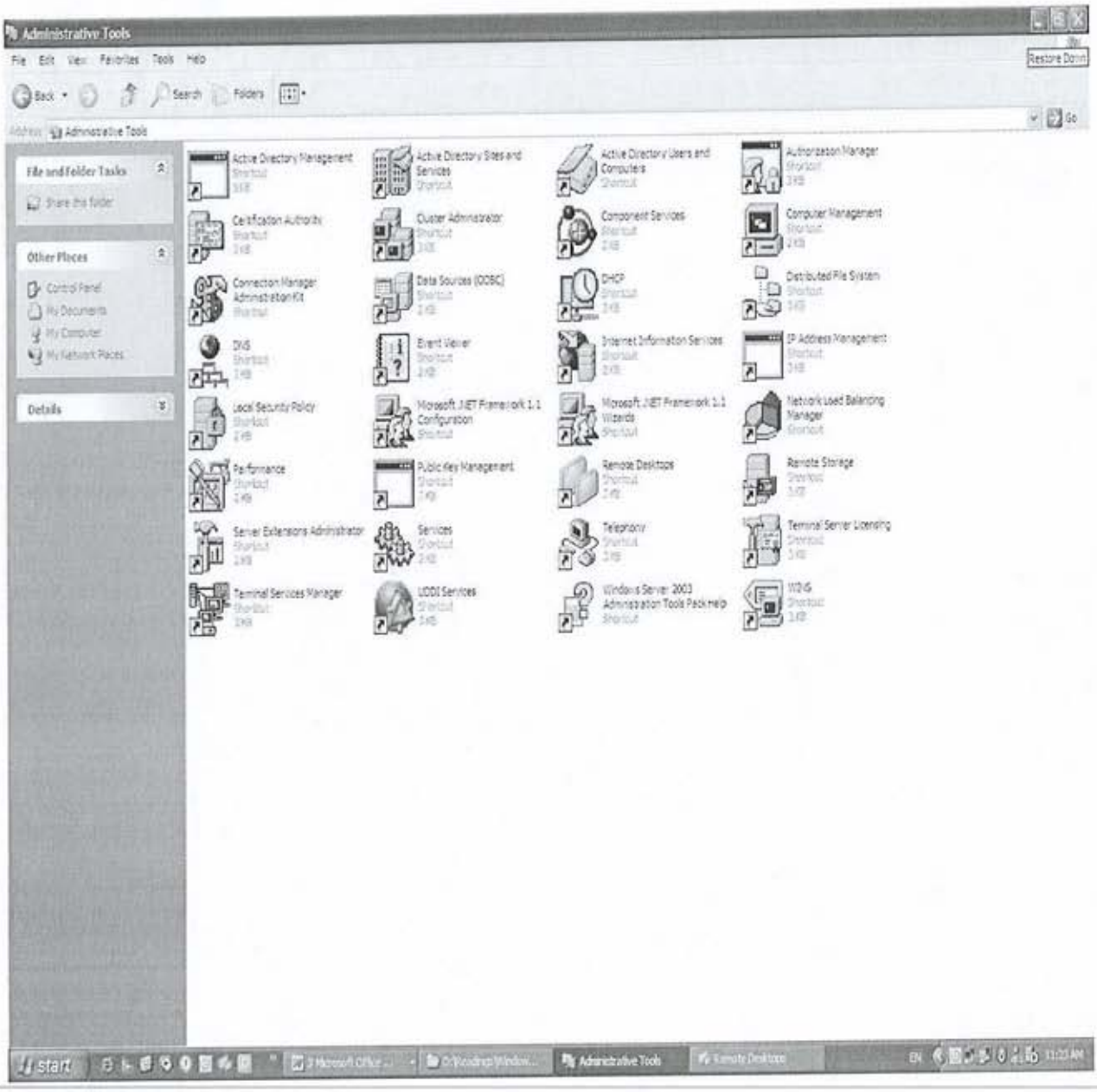
Appendix

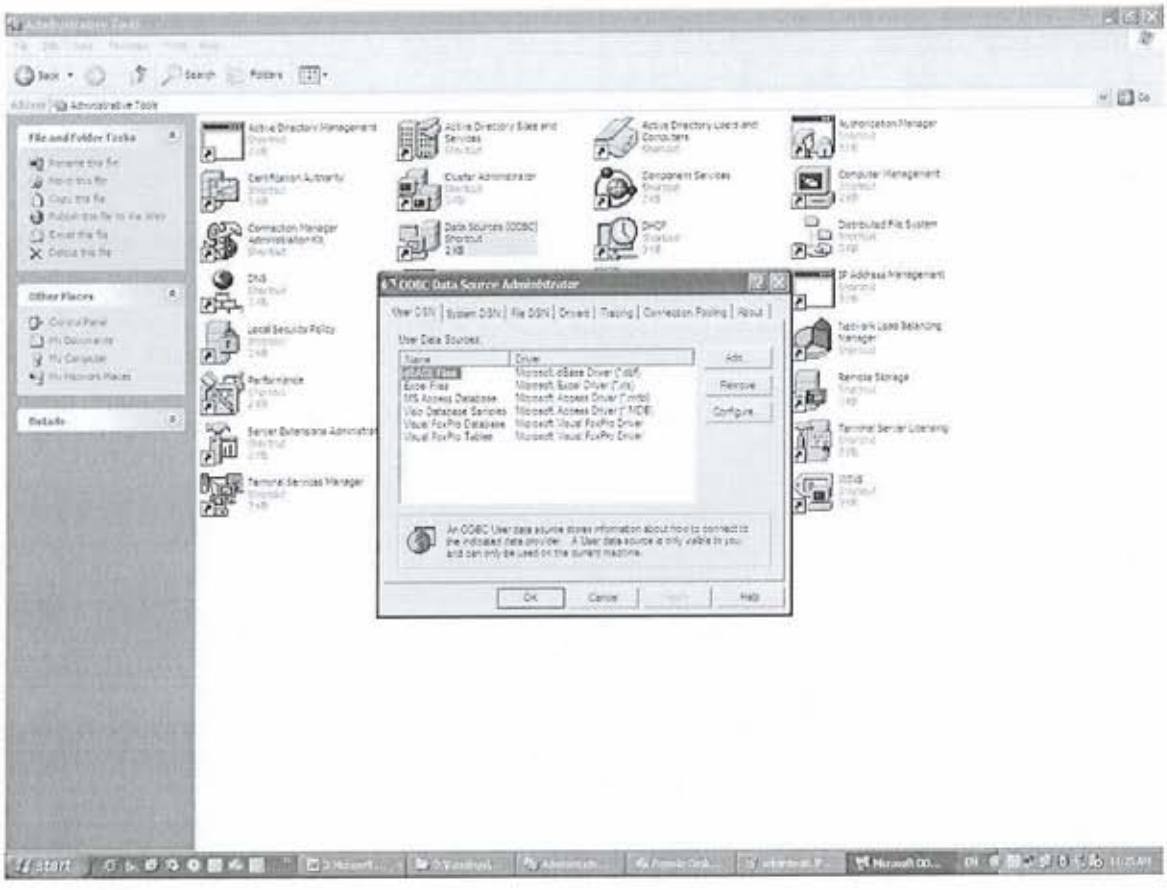
1. How does it Work?

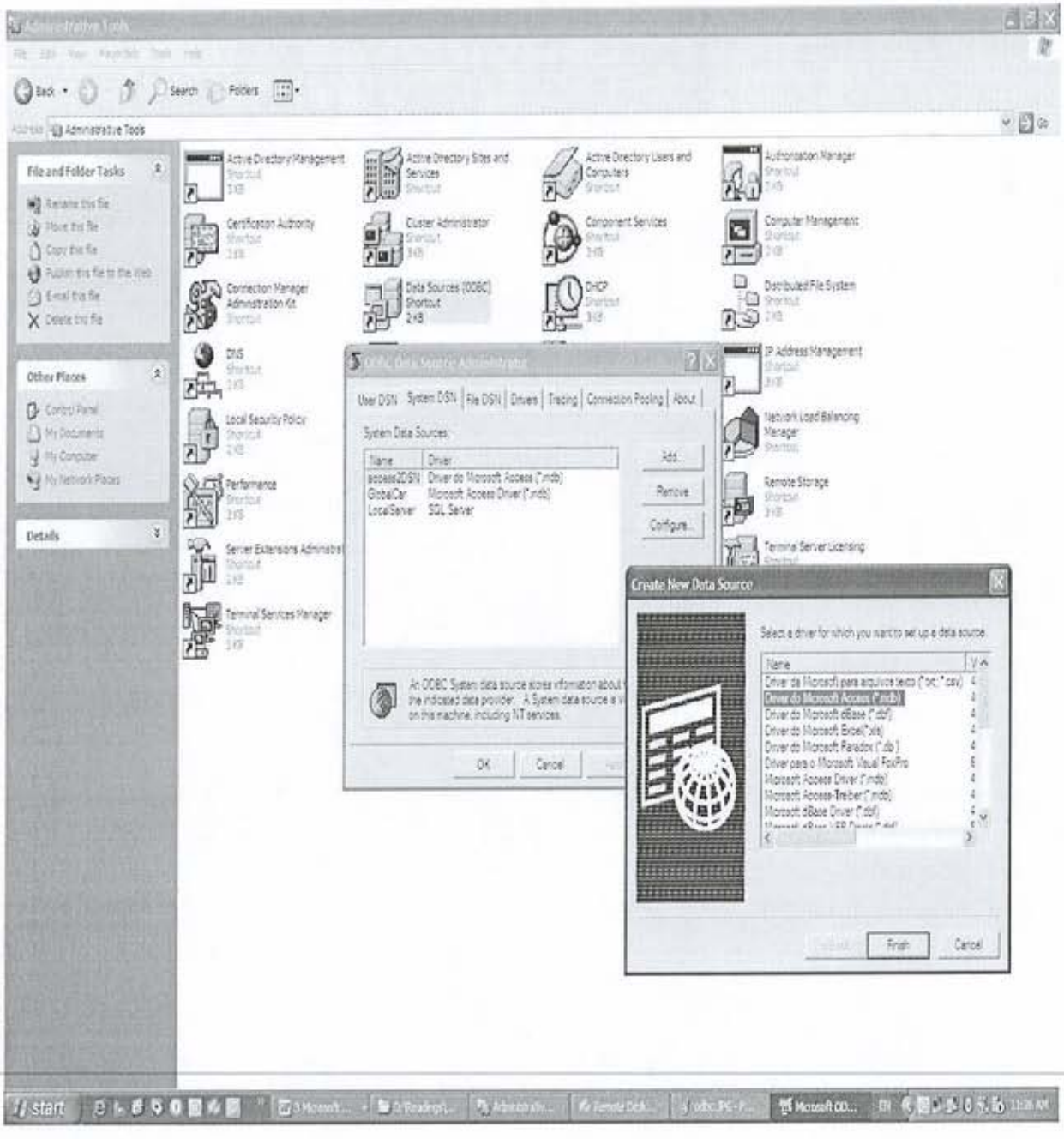
The search engine has a database which is divided into categories and subcategories. When the researcher looks for a certain page he'll enter a word or a description for the page from these categories. Then all the pages will appear with the particular description by the most visited first.

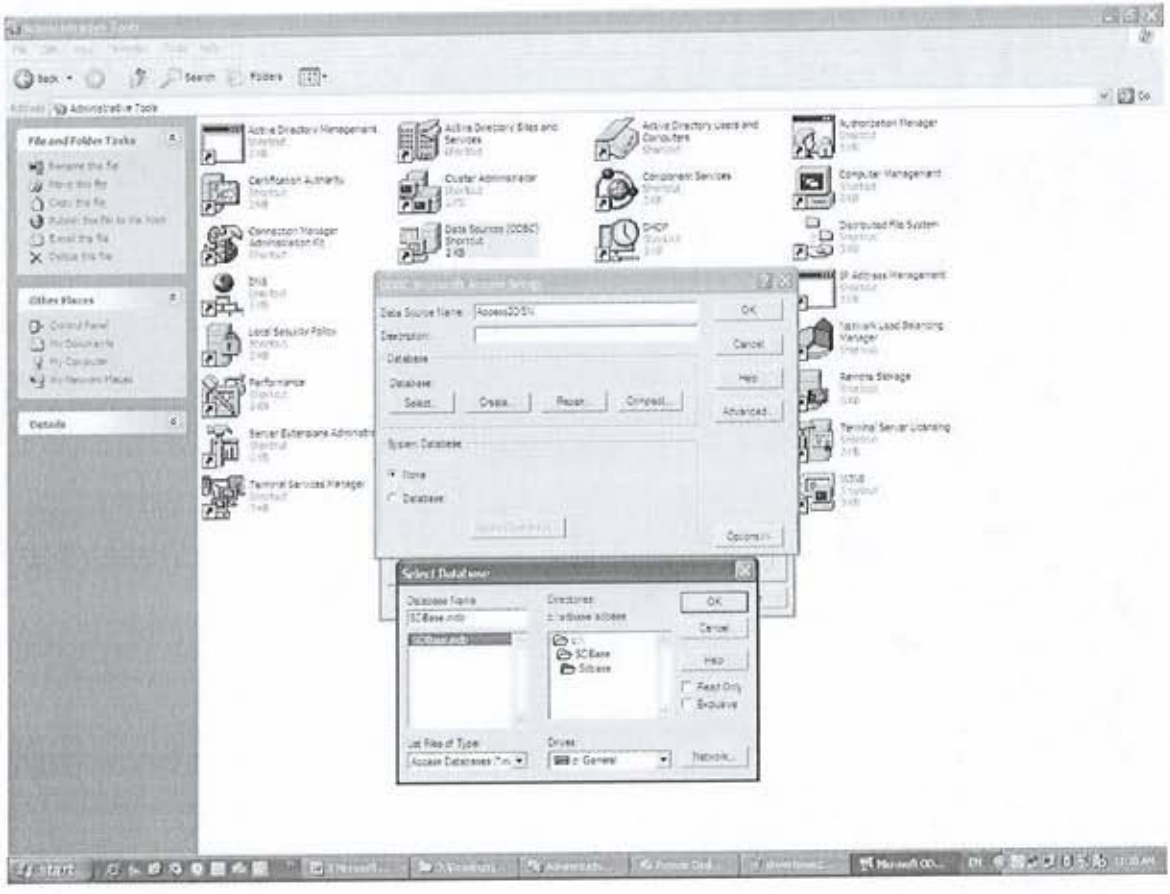
2. Installation and usage

Create a folder called sdbase, then a subfolder called sdbase too, then copy the sdbase.mdb file to the destination sdbase\sdbase and make sure that the sdbase file is not read only. Then create the odbc link as shown in the following pictures:


Administrative Tools
File Edit View Favorites Tools Help
Back Search Folders
Address Administrative Tools
File and Folder Tasks
Share this folder
Other Places
Control Panel
My Documents
My Computer
My Network Places
Details
Active Directory Management
Active Directory Sites and Services
Active Directory Users and Computers
Authorization Manager
Certification Authority
Cluster Administrator
Component Services
Computer Management
Connection Manager Administration Kit
Data Sources (ODBC)
DHCP
Distributed File System
DNS
Event Viewer
Internet Information Services
IP Address Management
Local Security Policy
Microsoft .NET Framework 1.1 Configuration
Microsoft .NET Framework 1.1 Wizards
Network Load Balancing Manager
Performance
Public Key Management
Remote Desktops
Remote Storage
Server Extensions Administrator
Services
Telephony
Terminal Server Licensing
Terminal Services Manager
UDDI Services
Windows Server 2003 Administration Tools Pack Help
WINS
start


ODBC Data Source Administrator
User DSN
System DSN
File DSN
Drivers
Tracing
Connection Pooling
About
User Data Sources:
Name
Driver
Add...
Remove
Configure
OK
Cancel
Help


Administrative Tools
Active Directory Management
Active Directory Sites and Services
Active Directory Users and Computers
Authorization Manager
Certification Authority
Cluster Administrator
Component Services
Computer Management
Connection Manager Administration Kit
Data Sources (ODBC)
DHCP
Distributed File System
DNS
IP Address Management
Local Security Policy
Network Load Balancing Manager
Performance
Remote Storage
Terminal Server Licensing
Terminal Services Manager
User DSN
System DSN
File DSN
Drivers
Tracing
Connection Pooling
About
System Data Sources:
Name
Driver
access2DSN
Driver do Microsoft Access (*.mdb)
GlobalCar
Microsoft Access Driver (*.mdb)
LocalServer
SQL Server
Add...
Remove
Configure...
OK
Cancel
Create New Data Source
Select a driver for which you want to set up a data source.
Driver do Microsoft Access (*.mdb)
Finish
Cancel
start



After finishing with the database installation, make sure thee iis is installed. Copy the html and asp files to the system.

Make default.asp as the main page in the iis administration page.

Now the program is ready to be used.