INTERNET SEARCH ENGINES SYSTEM FEATURES, OPERATORS AND COMPARISONS [+]

محركات البحث المستخدمة في الانترنيت

Maisaa Ibrahem Abdul-Hussain [*]

Abstract

Internet search engine is a program designed to help find information stored on a computer system such as the World Wide Web (WWW), or a personal computer. Millions of people around the world use i.nternet search engines and use regularly updated indexes to operate quickly and efficiently. Internet search engines work an attempt to match search query with the content of web pages that is has stored, or cached, and indexed on its powerful servers in advance of search. This paper presents a comparison between two of the most popular internet search engines Google and Yahoo according to the system features, time spend in search, number of results and percentage of world searches. Search results show that Google employs a number of techniques to improve search quality including pagerank and anchor text

Keywords: Information Retrieval, Google, Search Engine, Web Search ,World Wide Web, Yahoo

المستخلص:

محرك البحث في الأنترنيت عباره عن برنامج يصمم للمساعده في ايجاد المعلومه المخزونه في نظام الكومبيوتر مثل الشبكه العنكبوتيه العالميه (www) او اي حاسبه شخصيه. محركات البحث تستخدم من قبل ملايين الناس حول العالم وتستخدم نظام فهرسة محدث للعمل بسرعه وكفاءه .عمل محركات البحث في الانترنيت محاولة لمطابقة الطلب مع محتوى صفحات الأنترنت المخزونة. هذا البحث يقدم نتائج المقارنة بين اثنين من اشهر محركات البحث Google وYahoo اعتمادا على خصائص النظام ،الوقت المستغرق،عدد النتائج والنسبة المئوية للبحث عالميا.تشير نتائج البحث ان Google يستخدم عدد من التقنيات لتحسين نوعية البحث تتضمن ال pagerank (,anchor text) مما يجعله افضل في مجال البحث والنتائج التي يظهرها.

Introduction:

Search engines do not really search the World Wide Web (WWW) directly. Each one searches a database of web pages that it has harvested and cached. When use a search engine, always searching a somewhat stale copy of the real web page [1][2].

Search engine technology has had to scale dramatically to keep up with the growth of the web. In 1994, one of the first web search engines, the World Wide Web Worm (WWWW) had an index of 110,000 web pages and web accessible documents، 1994, and the received an average of about 1500 queries.As of November 1997, the top search engines claim to index

[+] **Received on 13/4/2010 , Accepted on 4/5/2011 .**

[*] **Assist Lech./College of Science/ University of Baghdad**

from 2 million (WebCrawler) to 100 million web documents (from Search Engine Watch) and in November 1997, 20 million queries per day .With increasing number of users on the web. It is foreseeable that by the year 2000, a comprehensive index of the Web contains over an billion documents. At the same time, the number of queries search engines handle has grown incredibly too per day. It is likely that top search engines will handle hundreds of millions of queries per day by the year 2000 and in August 2006 the number of queries increase to 150 million [1].

The goal of the system is to address many of the problems, both in quality and scalability, introduced by scaling search engine technology to such extraordinary numbers [2,3].

When click on links provided in a search engine's search results, retrieve the current version of the page. Search engine databases are selected and built by computer robot programs called spiders. These "crawl" the web, finding pages for potential inclusion by following the links in the pages they already have in their database [4]. They cannot use imagination or enter terms in search boxes that they find on the web. If a web page is never linked from any other page, search engine spiders cannot find it. The only way a brand new page can get into a search engine is for other pages to link to it, or for a human to submit its URL (Uniform Resource Locator) for inclusion. All major search engines offer ways to do this after spider find pages; they pass them on to another computer program for "indexing." This program identifies the text, links, and other content in the page and stores it in the search engine database's files so that the database can be searched by keyword and whatever more advanced approaches are offered, and the page will be found if search matches its content many web pages are excluded from most search engines by policy. The contents of most of the searchable databases mounted on the web, such as library catalogs and article databases, are excluded because search engine spiders cannot access them Figure 1 [2]. shown the architecture of Search Engine. In this paper, two most popular search engines: Google and Yahoo are presented and compared. The reset of the paper is organized as follows. Section 2 briefly describes the background about the search engines and operators used to search the queries with illustrated examples. Section 3 demonstrates the Yahoo and the Google system features and operators. In Section 4, experimental results of the two search engines are presented and compared and the conclusion in section 5.

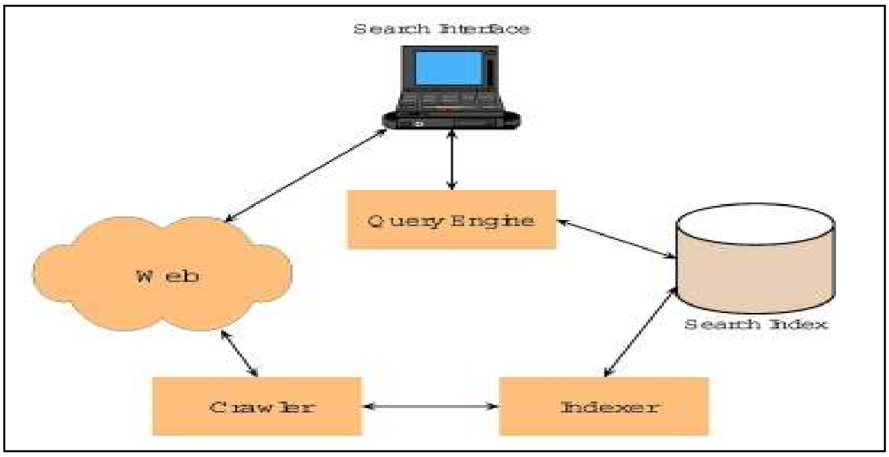


Figure (1): Architecture of Search Engine

Search Engines System Features and Operators :

Search engines do not always include all Web pages from a website. Usually they will only include a sample of pages – the ones they determine to be most valuable. Some of Web pages will be more important to have indexed [3]. A product information page is far more important to have indexed than a contact form, as it is more likely someone will search for products than contact form texts. Search engines do not always find the right pages to index by themselves; sometimes they need a little help and guidance. A system of standardized words ("operators") used to connect search terms. These include AND, OR, NOT and sometimes NEAR. AND requires all terms appear in a record. OR retrieves records with either term. NOT excludes terms. Parentheses may be used to sequence operations and group words. Always enclose terms joined by OR with parentheses Table 1 shown each operator, it is used and example of each one.

Table1: operators and examples

	Operator	**It's used**	**Example**
1.	+ (plus sign)	Use it to mark words that must appear in each Web page	+ computer science
2.	- (minus sign)	Use this to exclude pages containing a particular word	Computed science-data base
3.	““(quotation marks)	Indicates exact multiple-word phrases. Without quotation marks, the search engine may assume that the phrase is a list of separate query terms	"Jail House Rock"
4.	AND	Connects two search terms, both of which must appear in each Web page on the results list	Chantilly AND lace
5.	OR	Connects two words, at least one of which should appear on each Web page returned by the query	capital OR. Ringo OR Starr OR Starkey
6.	NOT	Is used much like the minus sign to exclude words. For some engines	AND Starr NOT Beatles
7.	( ) (parentheses)	With parentheses, Boolean really begins to look like alge bra. Use parentheses to connectgrouped terns	Beatles AND(Lennon AND McCartney)

Most popular Internet Search Engines:

This section presents the two most popular search engines Yahoo and Google and specifies the system features of each search engine .

A. Yahoo (www.yahoo.com) - (1994)

Yahoo is a web search engine, and is currently the third largest search engine on the web. Originally, Yahoo Search started as a web directory of other websites, organized in a hierarchy, as opposed to a searchable index of pages. In the late 1990s, Yahoo! evolved into a full-fledged portal with a search interface and, by 2007, a limited version of selection-based search [5].Yahoo Search, referred to as Yahoo provided Search interface, would send queries to a searchable index of pages supplemented with its directory of sites. The results were presented to the user under the Yahoo brand. Originally, none of the actual web crawling and storage/retrieval of data was done by Yahoo itself. Yahoo Search has been enhanced with a number of features that makes it easier for keyboard users to quickly reach various areas of the search application [3, 5]. Additionally, if visually impaired user and use a screen reader provided several auxiliary features and helpful messages to guide the search Yahoo provides results in seven categories. The first listed results under Web are from the search engine with the page title, a keyword in context extract (or directory description or meta description), the URL, file size, cache link, and a possibly more pages from this site link. The second and third items link to their image and video databases. The Yahoo directory results are available under the Directory heading. The local link goes to local and yellow page results. The News link goes to the Yahoo News database while the Products tab, introduced fall 2003, goes to the Yahoo Shopping search [4, 5].

B. Google (www.google.com) – (1997) – search engine.

Google search is a web search engine owned by Google Inc. and is the most-used search engine on the Web. Google receives several hundred million queries each day through its various services. Google search was originally developed by Larry Page and Sergey Brin in 1997.Google Search provides more than 22 special features beyond the original word-search capability. These include synonyms, weather forecasts, time zones, stock quotes, maps, earthquake data, movie show times, airports, home listings, and sports scores [2,6]. Figure 2[6] shown the architecture of Google search engine.

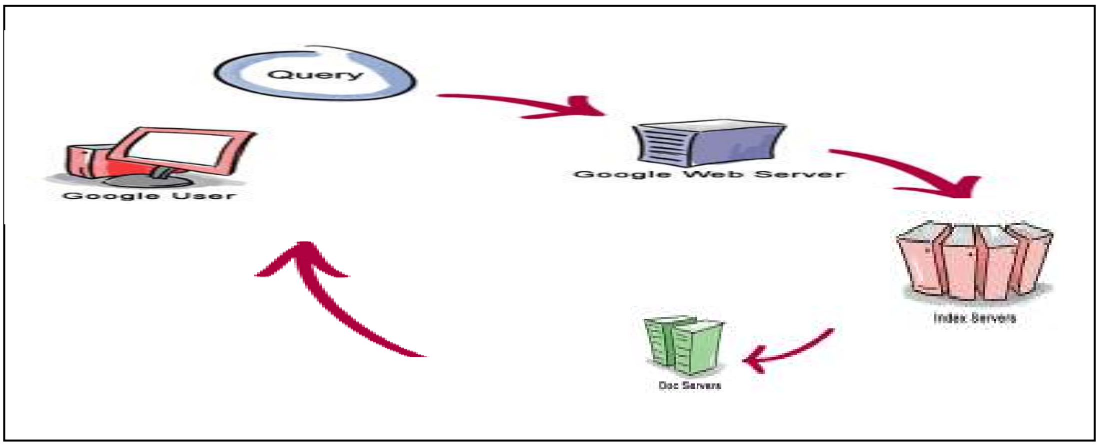


Figure 2: Architecture of a Google Search Engine

Google search engine has two important features that help it produce high precision results. First, it makes use of the link structure of the Web to calculate a quality ranking for each web page. This ranking is called PageRank . Second, Anchor text [7, 8]

1. PageRank

PageRank is a link analysis algorithm, named after Larry Page, used by the Google Internet search engine that assigns a numerical weighting to each element of a 518 million of these hyperlinks of documents, such as the WWW with the purpose of "measuring" its relative importance within the set. The algorithm may be applied to any collection of entities with reciprocal quotations and references. The numerical weight that it assigns to any given element E is also called the PageRank of E and denoted by PR(E).Page Rank is an excellent way to prioritize the results of web keyword searches. For most popular subjects, a simple text matching search that is restricted to web page titles performs admirably when PageRank prioritizes the results. For the type of full text searches in the main Google system, PageRank also helps a great deal [2,8].

PageRank is defined as follows:

Assume page A has pages T1...Tn which point to it (i.e., are citations). The parameter d is a damping factor which can be set between 0 and 1. We usually set d to 0.85. There are more details about d. Also C(A) is defined as the number of links going out of page A. The PageRank of a page A is given as follows:

$$PR(A) = (1-d) + d\,(PR(T1)/C(T1) + ... + PR(Tn)/C(Tn)) \ldots.. [2]$$

Note that the PageRanks form a probability distribution over web pages, so the sum of all web pages' PageRanks will be one. PageRank or PR(A) can be calculated using a simple iterative algorithm, and corresponds to the principal eigenvector of the normalized link matrix of the web. In addition, a PageRank for 26 million web pages can be computed in a few hours on a medium size workstation.

2. Anchor Text

The text of links is treated in a special way in Google search engine. Most search engines associate the text of a link with the page that the link is on. In addition, we associate it with the page the link points to. This has several advantages. First, anchors often provide more accurate descriptions of web pages than the pages themselves. Second, anchors may exist for documents which cannot be indexed by a text-based search engine, such as images, programs, and databases. This makes it possible to return web pages that have not actually been crawled [2]. Note that pages that have not been crawled can cause problems, since they are never checked for validity before being returned to the user. In this case, the search engine can even return a page that never actually existed, but had hyperlinks pointing to it. However, it is possible to sort the results, so that this particular problem rarely happens. This idea of propagating anchor text to the page it refers to was implemented in the (www) especially because it helps search non-text information, and expands the search coverage with fewer downloaded documents [9,10].

The Comparison Results:

Some common techniques will work in any search engine. However, in this very competitive industry, search engines also strive to offer unique features. In this section presented and compared between the two searches engines (Yahoo and Google). The comparison includes theoretical comparison and practical comparison in Table 2 the theoretical comparison contains the *differences features* of each of them and Table 3 contain the *features shared* by two Internet Search Engines Yahoo and Google.

.

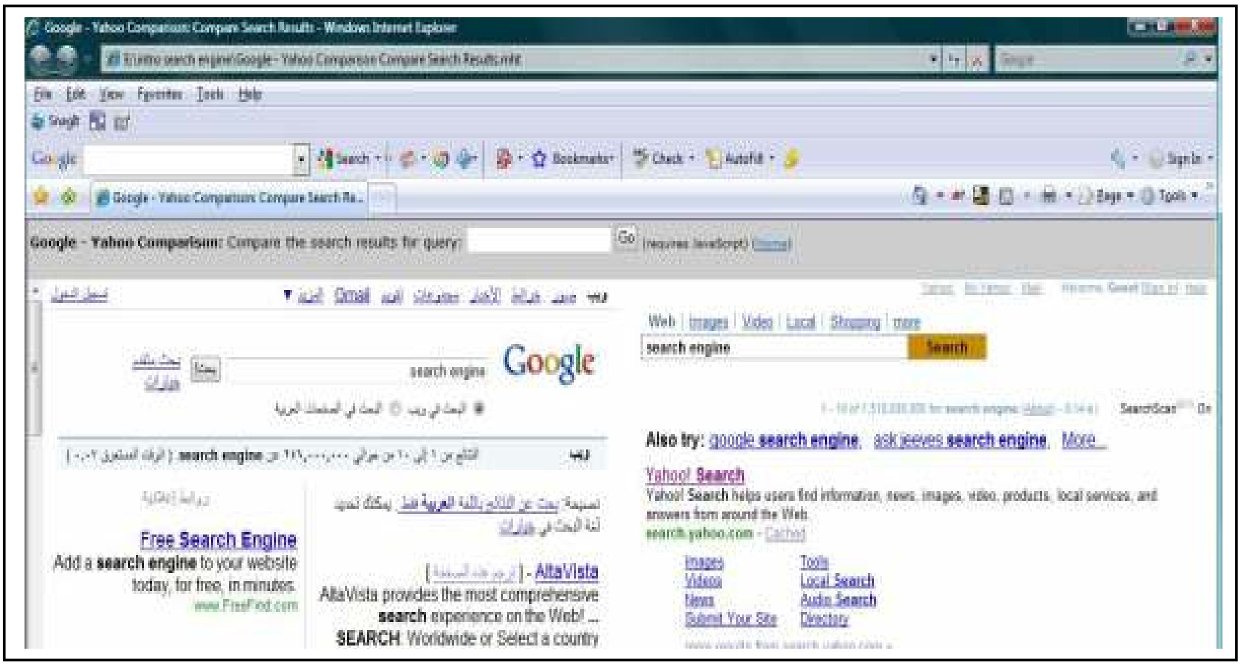
Table 2: Differences features

Search Engine	Google www.google.com	Yahoo! Search search.yahoo.com
Links to help	Google help	Yahoo! Help
Size, type	IMMENSE. Size not disclosed in any way that allows comparison. Probably the biggest.	HUGE. Claims over 20 billion total "web objects.".
Noteworthy features	PageRank™ system includes hundreds of factors, emphasizing pages most heavily linked from other pages.Many additional databases including Book Search, Scholar (journal articles), Blog Search, Patents, Images, etc .	Shortcuts give quick access to dictionary, synonyms, patents, traffic,stocks,encyclopedia, and more .
Results Ranking	Based on page popularity measured in links to it from other pages: high rank if a lot of other pages link to it .	Automatic Fuzzy AND .
Truncation, Stemming	No truncation. Stems some words. Search variant endings and synonyms separately, separating with OR (capitalized): airline OR airlines .	Neither. Search with OR as in Google.
Translation	Yes, in "Translate this page" link following some pages. To and sometimes from English and major European languages and Chinese, Japanese, Korean. Ues its own translation software with user feedback .	Available as a separate service.
Search size limit	32 words	16 words
Clustering results from same site	2 results from a site Link to show more .	1 results from a sit Link to show more.
Stemming	* Stems some words * (+) turns stemming of * No stemming with in phrases in quotes	No

Table 3: Features shared by two Internet Search Engines

Features shared by two INTERNET SEARCH ENGINES	Examples
Default AND between words	No need to type AND
Double quotes " " makes a phrase search, exactly as typed	"search engines"
OR (must be capitalized) to allow any or either word or phrase in quotes .	web OR internet "search engines" OR "subject directories"
Common words ignored + before forces them to be searched" " words enclosed in a phrase will be searched	Little consistency what words are common. Look at search results to find out what was searched.+all +in +a day's work"all in a day's work"
Minus (-) excludes. Cannot be used in combination with full Booleansearches with AND ,NOT, OR, ( )	pink –floyd "search engine" tutorial –site:com

The practical comparison results obtained from (Google-Yahoo comparison) web site as shown in Figure 7 and after search on a different terms (query).Table 3 shown the number of results and time spend of each query in Yahoo and Google .Table 4 shows the comparison between Yahoo and Google according to the Email system features and speed



Google - Yahoo Comparison: Compare the search results for query: Go (requires JavaScript) (Home)

search engine

Also try: google search engine, ask jeeves search engine, More...

Yahoo! Search
Yahoo! Search helps users find information, news, images, video, products, local services, and answers from around the Web.
search.yahoo.com - Cached

Free Search Engine
Add a search engine to your website today, for free, in minutes.
www.FreeFind.com

AltaVista
AltaVista provides the most comprehensive search experience on the Web! ... SEARCH: Worldwide or Select a country

Figure 7 :(Google-Yahoo Comparison)

Table 4: Results and Time spend comparison

Seq.	Queries	Google		Yahoo	
		No. of result	Time spend in second	No. of result	Time spend in second
1.	Search Engine	246,000,000	0,07 s	150,000,000	0.14 s
2.	introduction to search engine	85,700,000	0,07 s	75,800,000	0.28 s
3.	windows(programming) api in vb	14,800,000	0,07 s	9,010,000	0.10 s
4.	Blue Sky	85,000,000	0.08 s	226,000,000	0.30 s
5.	Windows server 2003	33,600,000	0.19 s	118,000,000	0.25

Table 5: Mail comparison

	Google mail(Hotmail)	Yahoo mail
1. Free email	Has 5GB	Has unlimited storage on free accounts
2.free accounts expire	Expires after 120 days of no activity	expires after 4 months of no activity
3. Searching for old emails	Is much better in yahoo	Is much simpler and quicker.

4. Number of users	193.3 million	273.1 million
5. Send ability	25 MB attachment	10 MB attachment

Conclusion:

1-Google employs a number of techniques to improve search quality including page rank and anchor text Google is a complete architecture for gathering web pages, indexing them, and performing search queries over them. Google result is more than yahoo and the time spend is lower than yahoo is.

2-The Percentage of world searches Between Leading Search Engine Providers in May 2010 70.13% for Google and 25.89% for Yahoo.

Reference

1- Nancy Paulson Computer Resource Teacher "Search Engine Basics" Douglas County School District May 2001.

2- The Anatomy of a Large-Scale Hyper-textual Web Search Engine Sergey Brin and Lawrence Page ,Department of Computer Science , Stanford University 2007.

3- Position Technologies, Inc " Tutorial for Search Engines and Directories " Copyright 2002-2007.

4- Junghoo Cho, Hector Garcia-Molina, Lawrence Page," Efficient Crawling Through URL Ordering" Department of Computer Science ,Stanford University.

5- Yahoo (2006). The Yahoo search API, yahoo, inc. http://developer.yahoo.com/search/.

6- Page Rank Citation Ranking: Bringing Order to the Web Lawrence Page, Sergey Brin, Rajeev Motwani und Terry Winograd Department of Computer Science Stanford University.

7- Steven Garcia " AN INTRODUCTION TO SEARCH ENGINE ARCHITECTURE" School of Computer Science and Information Technology, RMIT University,2004

8- Google (2003). "The Google API", google, inc. http://code.google.com/apis/. Graham, L. and Metaxas, P. T. (2003).

9- Lars Backstrom Jon Kleinberg "Spatial Variation in Search Engine Queries" Dept.of Computer Science Cornell University Ithaca, NY 14853.

10- Dr. Inderjeet Mani "A Study of Search Engine Technologies" imani@mitre.org