Why Google?

Ashish Kumar & Jayvardhan

This article discusses the growth of Internet in recent past, lists some of major search engines used for searching information, examines search engine Google, explains the working of Google and its effective use for searching pin-pointed information from the web.

Introduction

In an Information Society, information plays a pivotal role in every activity of human development. Internet in this information age is the most used medium of information generation, processing and dissemination. World wide web (www) of Internet provides the information on every subject one can imagine. The continuous growth of web pages has always been a problem to search the required information, as one cannot remember every web page. The solution of this problem is also provided by the Internet itself, with the name Search Engine.

Internet

Internet is a network of networks, where computers are connected to each other forming a Global Network; this network is continuously increasing, offering newer services to its users and increasing the usages of internet.

The growth of Internet usages is tremendous, in many parts of the world it is more than 1000% in last ten years, (Table:1 provides detail data of internet usages)

Table- 1 World internet usages

WORLD INTERNET USAGE AND POPULATION STATISTICS						
World Regions	Population (2009 Est.)	Internet Users Dec. 31, 2000	Internet Users Sep 30,2009	Penetration (% Population)	Growth 2000-2009	Users % of Table
Africa	991,002,342	4,514,400	67,371,700	6.8 %	1,392.4 %	3.9 %
Asia	3,808,070,503	114,304,000	738,257,230	19.4 %	545.9 %	42.6 %
Europe	803,850,858	105,096,093	418,029,796	52.0 %	297.8 %	24.1 %
Middle East	202,687,005	3,284,800	57,425,046	28.3 %	1,648.2 %	3.3 %
North America	340,831,831	108,096,800	252,908,000	74.2 %	134.0 %	14.6 %
Latin America/Caribbean	586,662,468	18,068,919	179,031,479	30.5 %	890.8 %	10.3 %
Oceania / Australia	34,700,201	7,620,480	20,970,490	60.4 %	175.2 %	1.2 %
WORLD TOTAL	6,767,805,208	360,985,492	1,733,993,741	25.6 %	380.3 %	100.0 %

Interestingly the growth is not only in western countries where this phenomenon started. In Table- 1 we can see that almost half of the internet users are from Asian countries.

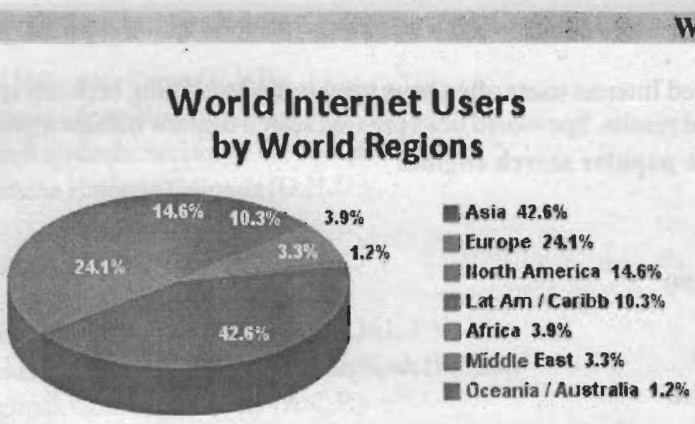

Fig. 1: World Internet Users

Information on Internet has doubled in every 9 to 14 months since it began in the late 1970s. In 1981 only 213 computers were connected to the Internet. By 2000 the number had grown to more than 400 million. The current number of people who use the Internet can only be estimated. The huge amount of information on Internet is also a problem as there is no classification Mechanism to Classify the Information. The most common way we know to find information on net is through Search Engine.

Search Engine

Search Engine is a system that helps users to find information stored on a personal computer, or a network of computers, such as the Internet. A user enters search terms, typically by typing a keyword or phrase, and the search engine retrieves a list of web sites, personal computer files, or documents, either by scanning the content stored on the computers or computer networks being searched or by *parsing* (analyzing) an index of their stored data.

These engines operate by buildingand regularly updatingan enormous index of Web pages and files. This is done with the help of a Web crawler, or spider, a kind of automated browser that perpetually trolls the Web, retrieving each page it finds. Pages are then indexed according to the words they contain, with special treatment given to words in titles and other headers. When a user inputs a query, the search engine then scans the index and retrieves a list of pages that seem to best fit what the user is looking for. Search engines often return results in fractions of a second.

Generally, when an engine displays a list of results, pages are ranked according to how many other sites link to those pages. The assumption is that the more useful a site is, the more often other sites will send users to it. Google pioneered this technique in the late 1990s with a technology called PageRank. But this is not the only way of ranking results. Dozens of other criteria are used, and these will vary from engine to engine.

Many times, search results also include what are called sponsored links, links that are ranked high in the search results or are prominently displayed because third-party companies pay a fee to the search engine. More often than not, sponsored links are labeled as such, but

inexperienced Internet users often have trouble distinguishing between sponsored pages and unsponsored results. Sponsored links provide search engines with their primary source

Some of the popular search engines

1. Google
2. Yahoo!
3. Search.com
4. Lycos
5. AltaVista
6. All the web
7. MSN
8. AOL
9. Webcrawler
10. Excite

Comparison of Google with other search engines

In 2006 Google ranked as the world's most comprehensive search engine, providing Web searches in more than 100 countries. Google was the most frequently visited Web site for searches. Google also offered a variety of other services for computer users, including e-mail, blog space, and tools for searching computer hard drives.

Google has been in the first place in Search Engines on the web, latest report of August 2009 of Search Engine Watch reveals that Google still has almost 65% share of total search engine users and ranked top in US.

Table.2: Top 10 Search Providers for August 2009, Ranked By Searches (US)

Search Provider	Searches	Month-on-Month Growth(%)	Share of Searches (%)
Google	6,986,580	2.6	64.6
Yahoo	1,726,060	-4.2	16.0
MSN/Bing	1,156,415	22.1	10.7
AOL	333,231	1.8	3.1
ASK.com	186,270	2.9	1.7
My Web	128,432	0.5	1.2
Comcast	50,328	-21.6	0.5
Yellow Pages	37,923	2.7	0.4
NextTag	31,830	0.4	0.3
Local.com	16,314	2.9	0.2
Total	10,812,734	2.9	100

How Google works

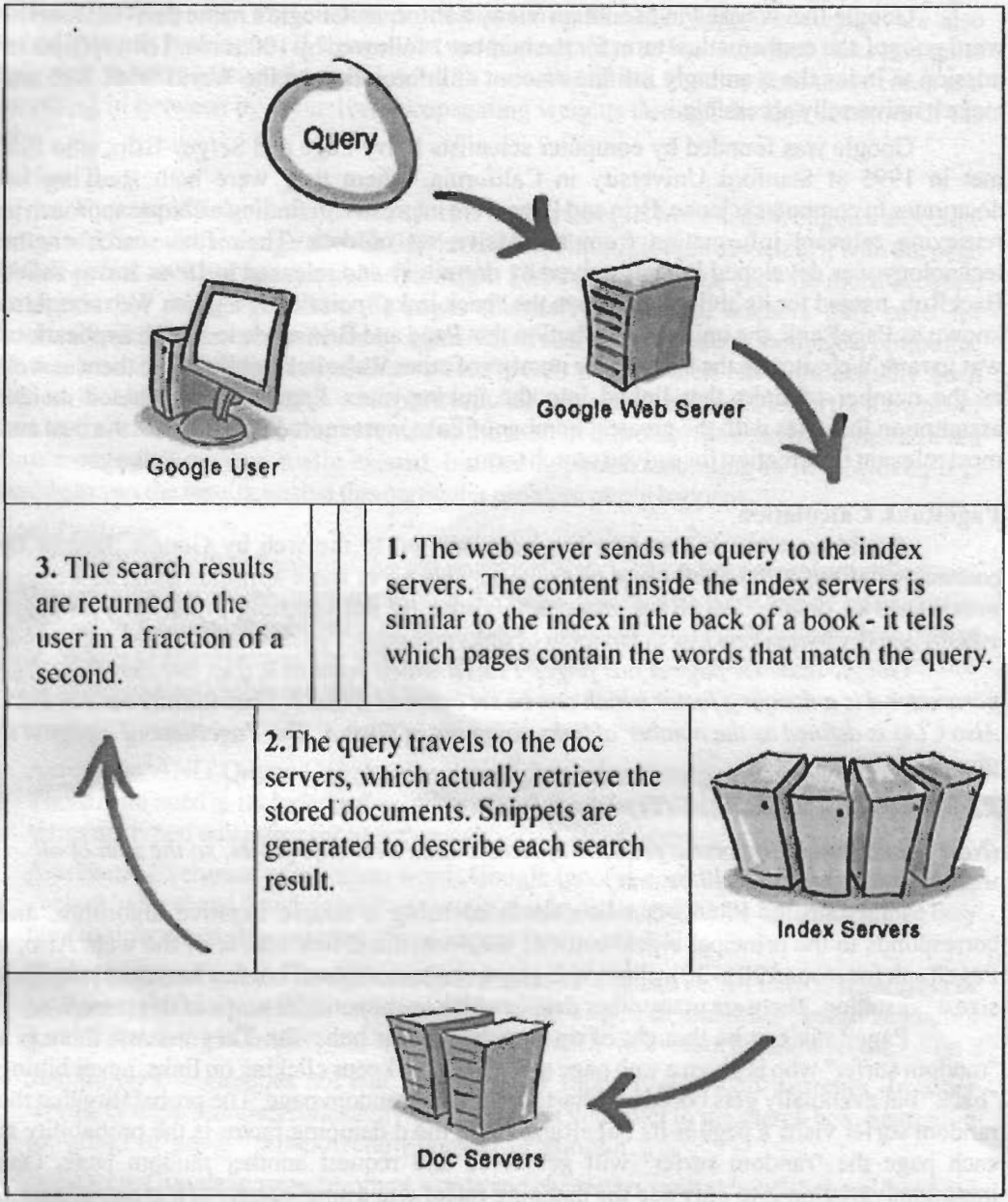

Fig.2: Different Steps of Google Search

The History of Google

Google Inc. is based in Mountain View, California. Google's name derives from the word *googol*, the mathematical term for the number 1 followed by 100 zeros. Term reflects its mission to index the seemingly infinite amount of information on the World Wide Web and make it universally accessible.

Google was founded by computer scientists Larry Page and Sergey Brin, who first met in 1995 at Stanford University in California, where they were both studying for doctorates in computer science. Brin and Page were interested in finding a unique approach to retrieving relevant information from a massive set of data. Their first search engine technology was developed in their university dormitory and released in 1996. It was called BackRub, named for its ability to analyze the "back links" pointing to a given Web site. Also known as PageRank, the unique contribution that Page and Brin made to search applications was to rank Web sites on the basis of the number of other Web sites that linked to them as well as the number of links that linked into the linking sites. PageRank was based on the assumption that sites with the greatest number of links were more likely to offer the best and most relevant information for a given search term.

PageRank Calculation

Academic citation literature has been applied to the web by Google, largely by counting citations or backlinks to a given page. This gives some approximation of a page's importance or quality. PageRank extends this idea by not counting links from all pages equally, and by normalizing by the number of links on a page. PageRank is defined as follows:

Google assumes 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. They usually set d to 0.85. 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:

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

(Note: 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. Also, a PageRank for around 20 to 30 million web pages can be computed in a few hours on a medium size workstation. There are many other details which are beyond the scope of this paper.

PageRank can be thought of as a model of user behavior. They assume there is a "random surfer" who is given a web page at random and keeps clicking on links, never hitting "back" but eventually gets bored and starts on another random page. The probability that the random surfer visits a page is its PageRank. And, the d damping factor is the probability at each page the "random surfer" will get bored and request another random page. One important variation is to only add the damping factor d to a single page, or a group of pages. This allows for personalization and can make it nearly impossible to deliberately mislead the system in order to get a higher ranking.

Another justification is that a page can have a high PageRank if there are many pages that point to it, or if there are some pages that point to it and have a high PageRank. Intuitively,

pages that are well cited from many places around the web are worth looking at. Also, pages that have perhaps only one citation from something like the Yahoo! homepage are also generally worth looking at. If a page was not high quality, or was a broken link, it is quite likely that Yahoo's homepage would not link to it. PageRank handles both these cases and everything in between by recursively propagating weights through the link structure of the web.

Anchor Text

The text of links is treated in a special way in Google. Most search engines associate the text of a link with the page that the link is on. In addition, Google 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 which have not actually been crawled. 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.

Other Features

Aside from PageRank and the use of anchor text, Google has several other features. First, it has location information for all hits and so it makes extensive use of proximity in search. Second, Google keeps track of some visual presentation details such as font size of words. Words in a larger or bolder font are weighted higher than other words. Third, full raw HTML of pages is available in a repository.

Features of Google:

1. Automatic 'AND' Query: Google only returns pages that include all of your search terms. There is no need to include 'and' between terms. Keep in mind that the order in which the terms are typed will affect the search results.
2. Automatic Exclusion of common word: Google ignores common words and characters such as 'where' and 'how', as well as certain single digits and single letters, because they tend to slow down your search without improving the results.
3. Capitalisation of words: Google searches are not case sensitive. All letters, regardless of how you type them, will be understood as lower case. For example. Search 'mahatma gandhi', 'Mahatma Gandhi', and 'mHaTmA gHaNdHi' will return the same result.
4. Stemming: Google does not use 'stemming' or support 'wild card' searches. In other words, Google searches for exactly the words that you enter in the search box. Searching for 'googl' of 'googl*' will not return 'googler' or 'googlin'.
5. '+' Search: Google ignores common words and characters such as 'what' 'where' 'how' etc if the words are essential to getting the result you want, you can include it by putting a '+' sign.
6. '-'Search : sometimes what you are searching for has more than one meaning, for example 'bass' can refer to fishing or music. You can exclude a word from your search by

putting a minus sing ('-') immediately in front of the term you want to avoid.

7. 'OR' Search: Google supports the logical 'OR' operator. To retrieve pages that include either word A or word B, use an uppercase OR between terms.
8. Domain Search: If you know the website you want to search but aren't sure where the information is located within that site, you can use Google to search only that domain. Do this by entering what you're looking for followed by the word 'site' and a colon followed by the domain name.
9. Phrase Searches: Search for complete phrases by enclosing them in quotation marks. Words enclosed in double quotes ("like this") will appear together in all result exactly as you have entered them.
10. Advance Search: Google has advance search feature, where one can restrict its search by "Language" "Date" "Occurrences" and "Domains" etc.

Conclusion

In the age of information explosion, latest and right information is key for personal and professional development, with advent of Internet, users have more opportunities to access required information from the World Wide Web. As we know that the web is a dynamic entity and it continues to grow in all dimensions. Because of this more and more information will appear on the net. To find information we continue to rely on Search Engines, in such case it is important to understand the concept of Search Engines especially Google, the number one search engine at present and remains so as its competitors are nowhere near to it.

Google is offering newer search features every day to its user, by understanding it's working and using its features wisely one can easily find pin-pointed Information without wasting time and energy.

Selected Readings

Muller (Jeanne Froidevaux). A Librarian's guide to the Internet, 2005 Chandos Publishing, Oxford

The Anatomy of a Large-Scale Hypertextual Web Search Engine Sergey Brin and Lawrence Page{sergey, page}@cs.stanford.edu

Microsoft ® Encarta ® 2009 retrieved on 26/03/2010

www.searchenginewatch.com retrieved on 27/03/2010.

www.internetworldstats.com/stats retrieved on 27/03/2010.