International Journal of Advanced Research in Engineering and Applied Sciences
ISSN: 2278-6252

SEARCH ENGINE ALGORITHMS

Archana Verma*

Prof(Dr) Ajay Rana**

Abstract: *The algorithm of a search engine determines how efficiently we can retrieve what we desire. Here we compare the algorithms used in top two search engines and draw comparison between the two.*

Keywords: *Yahoo, Google, Page rank, Optimization*

*Head, Information Technology Department, IEC Group of Institutions, Greater Noida, U.P.

**Program Director, Amity School of Engineering and Technology, Amity University, Noida, U.P.

1. INTRODUCTION

Search engines index tens to hundreds of millions of web pages. They answer tens of millions of queries every day. Automated search engines that rely on keyword matching usually return too many low quality matches. Half of the searches that are done in USA are done on Google and then comes Yahoo with 23.8%.Data Structures, Databases and Algorithms are designed to so that pages can be indexed per second to handle a no of queries per second. Google uses a page rank algorithm to rank its pages and return the top 10 results. Yahoo uses its web directory to return the results.

Here we first look at both the algorithms that Google and Yahoo use and draw a comparative conclusion between the two.

2. GOOGLE'S ALGORITHM

Google uses a very sophisticated and meticulous algorithm to get a page with text which has inbound links, internal links, appears in the title, is at the beginning of first page header. Google pretty well manages to differentiate between spams and real documents. In crawling they look at both the inbound and outbound link quality.

Google is heavily based towards informational websites and web pages. Google first develops a trust among sites before ranking them. New pages on a new site do not develop trust easily.

Google's indexing system needs to handle hundred of gigabytes of data and handle queries at a rate of hundreds to thousands per second. The search results may contain many relevant documents but the users are interested in only looking at a few first 10 results.

While calculating the page rank algorithm Google takes care of the following:

1) Anchor Text: Anchor text is the text of links. In Google the page containing the link and the page to which the link points is both taken into consideration.

2) Meta Data Information : This is the information about the document such as the

   a. Reputation of the source
   b. Update Frequency
   c. Quality
   d. Popularity or Usage
   e. Citations

2.1 Procedure Involved

1) The crawler is used to download the web pages. Usually 3 distributed crawlers are used. The lists of URL's to be fetched is sent to the crawler by a URL server.

2) The URL's fetched are then sent to the store server which compresses the documents and stores them in the repository, every web page now is given a unique ID called the DocID.

3) The Indexer and the Sorter are used together to perform indexing of web pages.

4) The Indexer reads from the repository, uncompresses the documents and parses them. It also converts the set of word occurrences called hits. It also records the word's position, font size, capitalization etc and distributes these words into barrels and also creates a forward index.

The other function of the indexer is to find out the no of links in every page and store it into the anchor file. The anchor file is used to record the no of inbound, outbound links and the text of the link.

5) The URL resolver takes the help of anchor file and converts relative URLs to absolute URLs or in other words the DocIds. The anchor text is used to provide information to forward index associated with the DocId and a database of links are created.

6) The sorter takes information from the barrels which are sorted by the DocId and resorts them by wordId to create the inverted index.

After each document is parsed, it is encoded into a number of barrels. Every word is converted into a wordId by using an in memory hash table – Lexicon. Once the words are converted into wordIds, their occurrences in the current document are translated into hit lists and are written into forward barrels.

The sorter takes each of the barrels and sorts it by WordId to produce an inverted barrel. This happens one barrel at a time and it can be done in parallel for a number of barrels at the same time to save time.

2.2 How to Rank

The hit lists include position, font, capitalization, information. Also the anchor list I information and the page rank of the document is there.

Page rank of a page is a sum of all the values of the links that point to it.

A simplified equation is

$$PR(u) = c \sum_{v \in B_u} PR(v) / N_v$$

where PR(u) is the page rank of a page u

B_u is a set of pages pointing to u

N_v is the number of links on page v

c is the normalization factor.

3. YAHOO'S ALGORITHM

When a user clicks on a high ranked page Yahoo profits, therefore yahoo pays more attention to what the content is on the page. They are commercially oriented so their search results are biased towards commercial websites than information sites.

In crawling Yahoo does not create a multi variable URL string. It has contents of its own that they want to promote for shopping etc. In query processing it puts quite a bit of weight on the commonly occurring words.

In Yahoo the page acquire a high rank which have poorly mixed anchor text on low quality links. Yahoo looks at both links to a page and links to a site while determining the relevancy of a page. So pages on newer sites also rank well even if their domain is not so trusted. Yahoo has a feature where users can ask or answer questions.

Yahoo is commercially biased in their search results for two reasons:

1) Offering paid inclusion.
2) Having so much internal content.

Yahoo gives much interest in taking its web directory as part of its ranking algorithm. If you are not in Yahoo's web directory you have to wait long enough. To have a business website you have to pay money annually to remain inside Yahoo's web directory.

Yahoo is a directory not a search engine.

3.1 Factors Involved

Yahoo takes care of the following factors while ranking its pages. The factors are given in the order of priority.

1) **Title**: This is the biggest ranking factor. The title of the website must contain your major keywords.

2) **Description**: The description you give when submitting to yahoo web directory should include major keywords however do not repeat them too much.

3) **Popularity**: The more the website is visited, the more chances it has of getting to top rank positions.

4) **Category**: The category you are listed in Yahoo's web directory should (if possible) contain some keywords. This plays a slight role in the ranking process.

5) **Site Wide Linking**: In Google one link per domain is enough to rank well but in Yahoo you can get a llot of site wide linking from other sites.

6) **Rules**: Sometimes Yahoo uses page rank algorithm of Google and sometimes not. Yahoo mainly uses its own ranking rules.

7) **Submit Homepage**: You have to submit your website over Yahoo's web directory and search engines, you do not have to submit all your web pages, just need to submit your home page.

3.2 How to Optimize

To get your website into the web directory of yahoo you need to do the following:

1) **Relevant Content**: Keep updating your website with content addition on a regular basis.

2) **Keyword Density**: Google prefers a keyword density of 2%. Yahoo prefers keyword density of 8%. This means your webpage should consist of synonyms and plurals of a word.

3) **Website Structure**: The content that occurs higher up in the code of your web pages is given higher priority.

4) **Inbound Links**: A link near the top of the web page is more valuable than a link nearer to the bottom. The quantity of the links will not get you high rankings on yahoo while the quality of those links is more important. Yahoo considers a link is of high quality if it is from a site that is ranking well on Yahoo.

5) **Aging**: When a new website is launched it should not have any inbound links. The inbound links are subject to a delay of upto 3 to 4 months before they hold their full weight.

4. CONCLUSION

The conclusion is given by comparing the two search engines.

1) Yahoo Does not have page ranking feature, whereas Google has.

2) Meta Data description is given more importance in Yahoo than in Google.

3) Yahoo gives higher priority for on page optimization, this helps Yahoo to achieve higher ranking position. Yahoo is volatile in the sense that on page optimization always searches for fresh content and therefore it is not stable, if you search for the same content after 2-3 days it will not give you the same result.

Whereas Google has better off page optimization. This is giving priority for inbound links. Google gives high weightage for quality incoming links.

4) Yahoo has paid and unpaid sites and takes more time to index unpaid sites whereas Google is absolutely free.

5) Yahoo gives higher priority to keyword density. Keyword density is the ratio between total no of keywords as measured against non keywords used in the body, text title, link anchor text and meta tags.

6) New websites are ranked well inn Yahoo before they are ranked on Google.

7) Yahoo matches text whereas Google matches concepts.

REFRENCES

[1] Sergey Brin , Lawrence Page, "The Anatomy of a Large –Scale Hypertextual Web Search Engine", Computer Science Department , Stanford University, Stanford, CA 94305.

[2] Petteri Huuhka, "Google: Data Structures and Algorithms", Department of Computer Science, University of Helsinki.

[3] Sources from the Internet.