Evaluation of Web Search Engines Based on Ranking of Results and Features

Rakesh Chandra Balabantaray rakeshbray@gmail.com
Department of Computer Science and Engineering,
IIIT Bhubaneswar
Bhubaneswar-751003, India

Monalisa Swain monalisaswain1988@gmail.com
CLIA-II Lab,
Department of Computer Science and Engineering,
IIIT Bhubaneswar
Bhubaneswar-751003, India

Bibhuprasad Sahoo s.bibhuprasad@gmail.com
CLIA-II Lab,
Department of Computer Science and Engineering,
IIIT Bhubaneswar
Bhubaneswar-751003, India

Abstract

Search engines help the user to surf the web. Due to the vast number of web pages it is highly impossible for the user to retrieve the appropriate web page he needs. Thus, Web search ranking algorithms play an important role in ranking web pages so that the user could retrieve the page which is most relevant to the user's query. This paper presents a study of the applicability of two user-effort-sensitive evaluation measures on five Web search engines (Google, Ask, Yahoo, AOL and Bing). Twenty queries were collected from the list of most hit queries in the last year from various search engines and based upon that search engines are evaluated.

Keywords: Web Search Engine, Ranking Algorithm, Evaluation of Web Search Engine, Result Relevancy.

1. INTRODUCTION

Now a day's most of the people prefer to use the World Wide Web for getting information. It is worth saying that web searching is the most popular online activity; behind email [1]. Nobody has time to enter the URLs of different websites to get particular information, so the alternative is web search engine. By using the web search engine user need not required to remember the large collection of URL names. The web search engine lets the user to enter the query on which user requires information and then gives all possible result related to that query. Formally "A web search engine is software code that is designed to search for information on the World Wide Web". Typically, web search engines work by fetching as many documents as possible and indexing these document by using a program indexer which creates an index based on the words contained in each document. During indexing, stop words(that, the, this, etc.) elimination may be performed since such words usually have no information value and many statistical information about words (e.g., number of occurrences in the individual pages or in all of the indexed Web pages, position of the word) are usually stored in an inverted file structure. When an user enters a query for finding the related information this index is used for matching the query term and then returns the URLs of the pages to the user .But before showing the pages to the user, a ranking

algorithms is used by the search engine to rank the document and on the basis of ranking value, shows the most relevant pages at the top and less relevant ones at the bottom .The different search engine returns different results which mostly depends upon the different ranking algorithms used.

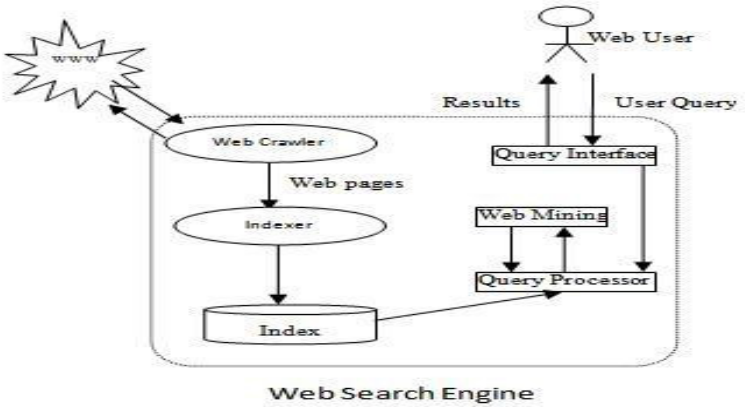

FIGURE 1: Architecture of Web Search Engine.

With the number of search engines available, it is important for the user to know the different search engine in terms of their result relevancy and effectiveness. In this paper, the survey of different web page ranking algorithms and evaluation of result returned by the different search engines is carried out.

2. THE ALGORITHM

The size of the World Wide Web is growing rapidly and the number of queries submitted to the search engines by the increasing number of users on the web are also growing exponentially. Therefore these queries must be efficiently processed by the search engines.

To present the documents in an ordered manner, web page ranking methods are applied which can arrange the documents in order of their relevance, importance and content score and use web mining techniques to order them [2].

2.1 Web Page Ranking Algorithms

Among the different ranking algorithm available Link analysis algorithm (LAR) is used by most of the search engine. Some of the algorithms based on link analysis are:

- InDegree Algorithm
- PageRank Algorithm
- HITS Algorithm
- SALSA

2.1.1 InDegree Algorithm

InDegree algorithm is a graph-based link analysis algorithm. It ranks the web pages, according to the popularity of the web pages. The popularity of a page on the web is measured by the number of pages that point to it. Let G = (V, E) be a directed graph with the set of vertices V and set of edges E, where E is a subset of V × V. For a given vertex Vi, let In (Vi) be the set of vertices that point to it (predecessors).If we take the web pages as the nodes of a graph G, then we refer to this algorithm as the InDegree algorithm, since it ranks pages according to their in-degree in the graph G. That is, for every node Vi, a_i= |In (Vi)|. This simple heuristic was applied by several search engines like Altavista, HotBot etc in the early days of web Search. In his study [3] Kleinberg makes a convincing argument that this algorithm is not sophisticated enough to capture the authoritativeness of a node, even when restricted to a query dependent subset of the web.

2.1.2 PageRank Algorithm

PageRank Algorithm was proposed by SurgeyBrin and L.Page [4] and this algorithm is used by Google. PageRank works on a mathematical algorithm which is based on the web graph. In this all worldwide pages are denoted by nodes and hyperlinks by edges. Authority hubs e.g. usa.gov.in and cnn.Com are taken into consideration. Importance of a particular page is indicated by its rank value. A hyperlink to the page adds to the importance of that page. That means if a page is linked to by number of pages, its importance increases and thus increases its rank value. On the contrary if a page has no links to it, its rank value remains low. This concept can be better illustrated by below cartoon. The size and happiness of each face is proportional to the total size and happiness of the other faces which are pointing to it.

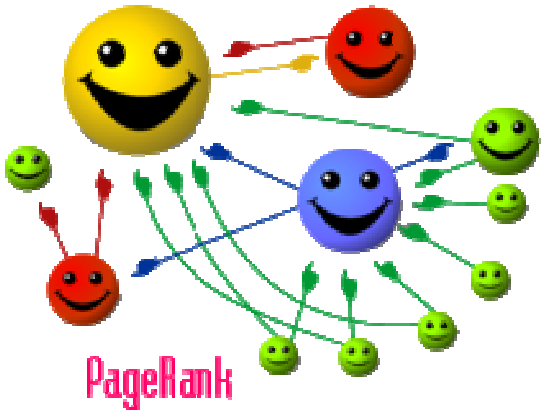

FIGURE 2: Basic Principle of PageRank.

Page rank is nothing but probability distribution i.e. probability of arriving at a particular page while randomly clicking on the links. PageRank can be calculated for collections of documents of any size. A page rank of 0.5 means there is a 50% chance that a person will be directed to the document with page rank 0.5 while randomly clicking the link.

PageRank value for any page **u** can be expressed as:

$$PR(u) = \sum_{v \in B_u} \frac{PR(v)}{L(v)}$$

This signifies that page rank value of **u** depends on PageRank values of each page **v** contained in the set $\mathbf{B_{u,}}$ divided by the number of links $L(v)$ from page **v**. Here $\mathbf{B_u}$ is the set containing all pages linking to page **u**.

2.1.3 HITS Algorithm

Hyperlink-Induced Topic Search (HITS) algorithm is developed by Jon Kleinberg [3] and is also known as hubs and authorities that rates Web pages. HITS algorithm ranks the web pages by taking into consideration in links and out links of the web pages. In this algorithm web page pointing to many hyperlinks are called as hub and the web pages pointed by many hyperlinks are called as authority.

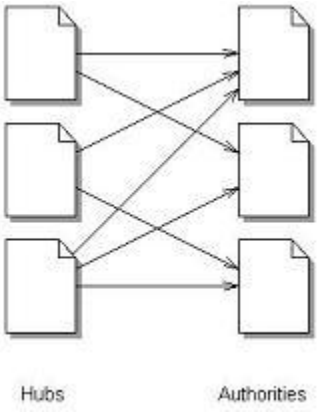

FIGURE 3: Hubs and Authorities.

In the first step of HITS algorithm top n pages returned by a text-based search algorithm is taken. After the retrieval of pages the HITS algorithm focuses on the structure of the web pages without concentrating on their textual contents. Original HITS algorithm has some problems [5] which are:

- High rank value is given to some popular website that is not highly relevant to the given query.
- Drift of the topic occurs when the hub has multiple topics as equivalent weights are given to all of the out links of a hub page.

In order to minimize the problem of the original HITS algorithm, a clever algorithm is proposed which is the modification of standard original HITS algorithm [6]. In this algorithm every link is provided a weight value depending on the terms of queries and endpoints of the link. Weights to the link are decided by combining an anchor tag and a large hub is divided into smaller parts so that every hub page is concentrated only on one topic.

2.1.4 SALSA (*Stochastic Approach for Link Structure Analysis*)

SALSA algorithm proposed by Lempel and Moran [7], based upon the theory of Markov Chains, and relies on the stochastic properties of random walks performs on their collection of pages. This algorithm combines ideas from both HITS and PAGE RANK algorithm. On the bipartite hubs-and-authorities graph this algorithm performs a random walk alternating between the hubs and authority sides.

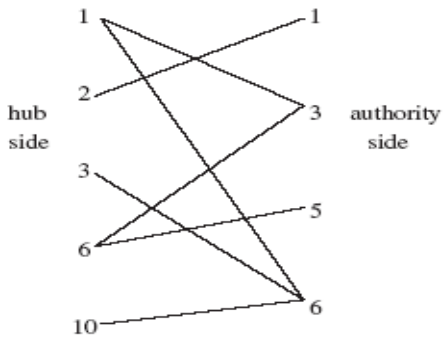

FIGURE 4: Hub Side and Authority Side.

The random walk starts from an authority node selected at random and then proceeds by taking backward and forward steps alternatively. When on the authority side node, the algorithm selects one of the incoming links randomly and moves to a hub node. When on the hub side node, the algorithm selects one of the outgoing links randomly and moves to a authority node. In the SALSA algorithm each hub instead of broadcasting its weight, divides its weight equally among the authorities to which it points and each authority divides its weight equally among the hubs that point to it.

3. FEATURES OF SEARCH ENGINES

Here we have considered five search engines for comparing the different features given by search engine. Table 1 shows the options available for searching on these five search engines. From the above table, it is clear that Google provides more options as compared to other search engines which make it having better search capabilities than others. The description of each feature is described in Table 2.

Features	Google	Yahoo	Aol	Ask	Bing
Website	www.google.com	www.search.yahoo.com	www.aol.com	www.ask.com	www.bing.com
Search Operator	**AND, OR, NOT**	**AND, OR, NOT**	**AND, OR , NOT**	**AND, OR, NOT**	**AND, OR, NOT**
Search Web	Yes	Yes	Yes	Yes	Yes
Search Images	Yes	Yes	Yes	Yes	Yes
Search Videos	Yes	Yes	Yes	Yes	Yes
Search News	Yes	Yes	Yes	Yes	Yes
Search Maps	Yes	NO	Yes	NO	Yes
Search Books	Yes	NO	NO	NO	NO
Advance Search	Yes	Yes	Yes	Yes	Yes
Advertising Programs	Yes	Yes	Yes	Yes	Yes
Case Sensitive	NO	NO	NO	NO	NO
Change Search Settings	Yes	Yes	Yes	Yes	Yes
Display No. of Results	Yes	Yes	Yes	NO	Yes
Multi-Language Support	Yes	NO	NO	NO	NO
Preferences	Yes	Yes	Yes	Yes	Yes
Questions/ Answers	NO	Yes	No	Yes	NO

Safe Search	Yes	Yes	Yes	Yes	Yes
Shopping	Yes	Yes	Yes	Yes	NO
Themes	NO	NO	NO	NO	Yes
Translation Services	Yes	NO	NO	NO	Yes

TABLE 1: Features Available in Five Search Engines.

Features	Explanation
Search Operator	The operators used internally by the search engine for retrieving the results.
Search Web	Search the information from the web.
Search Images	Search for images on the web.
Search Videos	Search online videos, T.V. shows from the web.
Search News	Search for news, top stories.
Search Maps	Search engine enables users to search for different location and directions from one location to another and more.
Search Books	Search engine enables users for search and preview millions of books from libraries and publishers worldwide.
Advance Search	Allows Users with advanced options to write specific query and return more precise results.
Advertising Programs	Search engines provide the facility to the users to advertise their business and products.
Answers	Good facility where people ask and answer questions on any topic and can share facts, opinions and personal experiences on communities.
Case Sensitive	Search engine is case sensitive or not.
Change Search Settings	User can change the search settings.
Display No. of Results	Search engines display the number of results fetched per page.
Multi-Language Support	Search engine supports multiple languages?
Safe Search	Allow the user to filter out explicit, adult-oriented content from results.
Shopping	Facility of buying online products.
Themes	User can change theme according to his own choice.

Translation Services	Search engine translates text and web pages to another language.

TABLE 2: Description of Each Feature.

Besides these features, many other features which are unique to that particular search engine are provided by these search engines just like Google provides summary of the retrieved pages.

4. RELEVANCE EVALUATIONS

This test evaluates five search engines i.e. Google, Yahoo, Bing, Ask, Aol in the english language. We are selecting the above 5 search engine for evaluation considering the popularity of the search engines. We are selecting 20 queries randomly from the mostly searched top queries in various search engines in the year 2012. The list of queries that we have taken for evaluation are given below.

Query No.	Query	Query No.	Query
1	President Pranab Mukharjee	11	Amarnath yatra
2	Rajesh Khanna Death	12	Samsung dual phone
3	Kingfisher Airlines	13	Puri jagannath temple
4	IBPS	14	Adarsh housing society scam
5	Anna Hazare	15	Sachin tendulkar
6	London Olympics 2012	16	Aseem Trivedi Cartoons
7	2G scam	17	Ajmal kasab
8	Online shopping store	18	Vaishno Devi
9	Taj Mahal	19	Bikaner Rajasthan
10	Gate way of India	20	Coalgate Scam

TABLE 3: Query List.

As results are retrieved from the index created during indexing, in order to avoid index change problem, we are processing all the queries on the same day and at the same time a single query is processed on all the search engines. The change of index may create problem because the latest index may contain additional information that others don't have or may be the coverage is wider than other.

From the set of result returned by each search engine we are taking top 10 results for performance measure [8][9][10]. Here the results returned by the search engines are categorized as Relevant, Partially Relevant, Irrelevant and Only link. The result which gives the complete information related to the query is categorized as relevant result, and which gives some of the query related information is categorized as partially relevant, and the result page which gives a link, by visiting which we can get the information about searched query is categorized as link and the result which does not provide any information related to query is taken as irrelevant. Again we are categorizing the description (snippet) of each result as seems relevant, seems partial relevant, seems irrelevant and not available. For calculating the average value of result we are assigning different value to different category given below.

Content of Retrieved URL	
Relevant	1
Partially relevant	0.5
Only link	0.25
Not relevant	0

Description(snippet)	
Seems Relevant	1
Seems Partially relevant	0.5
Seems Irrelevant	0
Not Available	0

TABLE 4: Result and Description Category.

According to the category of result as explained above each result is given a numeric value. We are calculating the average value of the top 5 results for all the queries for each search engine separately and same calculation is also done for the top 10 results separately. Average for top 5 and top 10 is calculated separately because most of the users look only top 3 to 4 results rather than navigating the entire page and also user satisfaction label is high if the required information is found within top 5 results. From below figures, figure-5 represents average value of top 5 results and figure-6 represents average value of top 10 results for each search engine.

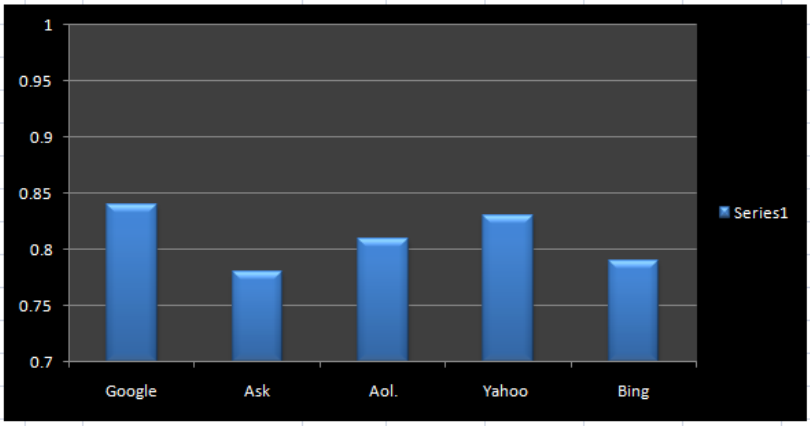

FIGURE 5: Average Value of Top 5 Retrieved Results.

This study shows that the average of top 5 result retrieved by Google is high followed by Yahoo and Aol whereas Bing and Ask results lower value. This result gives an idea about the relevancy of retrieved result by the above search engine.

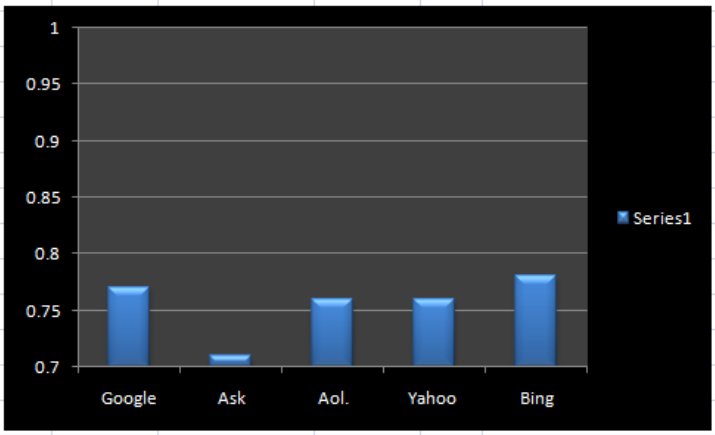

FIGURE 6: Average Value of Top 10 Retrieved Results.

But the average of top 10 result retrieved by Bing is high followed by Google, Yahoo and Aol whereas Ask results lower value. This study shows that in case of Bing search, more number of relevant results is found between top 5 to top 10 results while comparing with the other four search engine.

Similar to the average of result calculation we are here calculating the average of description by reading the snippet of the result. The figure-7 and figure-8 shows the average of description for top 5 results and top 10 results separately.

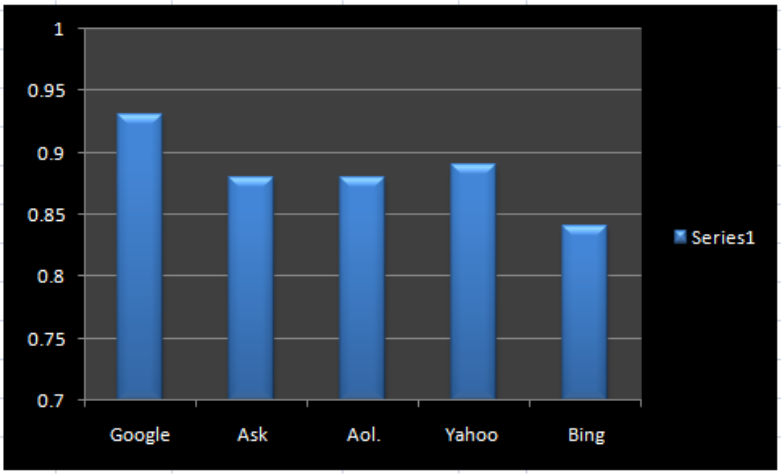

FIGURE 7: Average Value of Top 5 Result's Snippet.

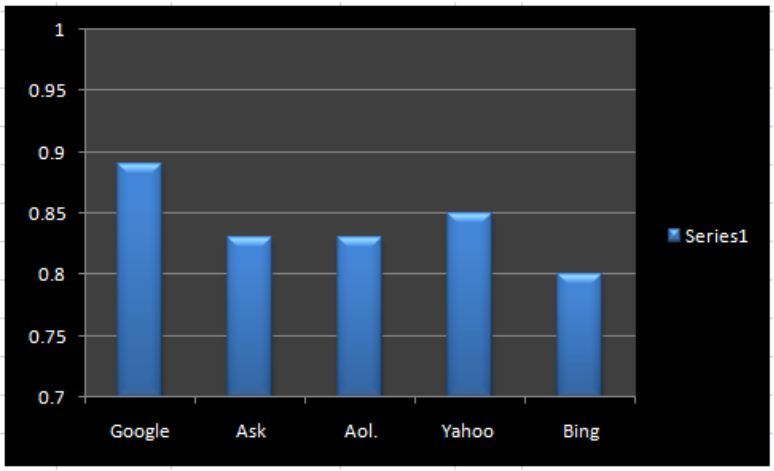

FIGURE 8: Average Value of Top 10 Result's Snippet.

The above two figures show that the average value of description in case of Google is more followed by Yahoo, Aol, Ask and Bing results low.

5. CONCLUSION

In this study we compare the features of five web search engines and also measure the performance of these web search engines examining top 10 result returned by these search engines. Considering the features provided by the web search engines we can conclude that Google provides more number of features and better interface than others. One of the features provided by Bing i.e. Theme which enables users to change the theme of home page according to their own choice is not provided by Google. If Google provides this feature, it can draw attention of more users towards it. Considering the evaluation of top 5 results we conclude that Google is the best search engine and provides more relevant result than others; but for top 10 results, the result shows that Bing is better than others whereas In case of snippet of result provided by Web search engines, Google is the best.

6. REFERENCE

[1] Spink, A., Jansen, B. J., Blakely, C., and Koshman, S. "A study of results overlap and uniqueness among major web search engines". Information Processing and Management,

[2] N. Duhan, A.K. Sharma, K.K. Bhatia. "Page Ranking Algorithms: A Survey" *Advance Computing Conference, 2009. IACC 2009 IEEE International.*

[3] J. Kleinberg. "Authoritative sources in a hyperlinked environment", *Journal of ACM (JASM), 1999*.

[4] L. Page, S. Brin, R. Motwani, and T. Winograd. "The PageRank Citation Ranking: Bringing order to the Web", *Technical report, Stanford Digital Libraries SIDL-WP-1999-0120, 1999.*

[5] D.K. Sharma, A.K. Sharma. "A Comparative Analysis of Web Page Ranking Algorithms", *International Journal on Computer Science and Engineering, 2010*.

[6] S. Chakrabarti, B. E. Dom, S.R. Kumar, P. Raghavan, S. Rajagopalan, A. Tomkins, D. Gibson, and J. Kleinberg. "Mining the Web's Link Structure", *Computer, 1999*.

[7] R.Lempel, S.Moran, "The stochastic approach for link-structure analysis (SALSA) and the TKC effect", Proceedings of the 9th International World Wide web Conference, 2000.

[8] Chu, H., and Rosenthal, M. 1996. "Search engines for the World Wide Web: A Comparative study and evaluation methodology", *In Proceedings of the 59th annual meeting of the American Society for Information Science, 1996.*

[9] Tomaiuolo, Nicholas G. , Packer, Joan G. "An analysis of Internet search engines: Assessment of over 200 search queries", *Computers in Libraries, 1996.*

[10] L. Vaughan. "New measurements for search engine evaluation proposed and Tested", *Information Processing and Management, 2003*.

[11] M. Kaur, N. Bhatia and S. Singh. "WEB SEARCH ENGINES EVALUATION BASED ON FEATURES AND END-USER EXPERIENCE", *International Journal of Enterprise Computing and Business Systems, 2011*.

[12] M.P. Selvan, A.C. Sekar and A.P. Dharshini. "Survey on Web Page Ranking Algorithms", *International Journal of Computer Applications, 2012.*