RANKING SEARCH ENGINE RESULTS

Rajesh Kumar Goutam [1], Pankaj Nigam [2]and Somya Tyagi [3]

[1] Department of Computer Science, University of Lucknow, Lucknow
[2] Department of Computer Engineering & Application Mangalayatan University Aligarh
[3] Department of Computer Science Engineering, Institute of Engineering and Technology, Dr. Bhimrao Ambedkar University Agra

ABSTRACT

To retrieve the relevant results from World Wide Web repository, we need a search engine that prioritize the results gradually as per their relevant scores. The paper describes how prioritization of documents is done with the help of algorithm and organized on landing page to satisfy searchers needs in minimum span of time. We analyze the two dimensional user's traversing approach and established the relationship between the parallel links and series links with relevant scores and on the basis of this approach a metric have been derived for the search engine evaluations. The main attraction of this metric is its incorporation of implicit feedback and explicit feedback.

KEYWORDS

Information retrieval; Search engine performance; Search engine evaluation; Correlation based Ranking.

1. INTRODUCTION

This Search engines are the software, which are used to retrieve the information from the World Wide Web. We need to pose queries to the interface of search engines. As per the internal architecture of algorithm, search engines cleans up the irrelevance words from the posed queries and matches the remaining words with the contents of web documents [1,2]. Search engines prioritize the documents as per frequency of words matching and indexed these documents with snippets for end users [3].

Globalization promotes world wide access of information and World Wide Web is rich repository of almost every kind of information and freely available to each one. We need search engines to fetch the desired documents from the repository. Thousands search engines exist to facilitate the searchers but not all are popular equally. Few search engines like Google and Yahoo are leading in market and tremendously attracted large portion of population with their satisfaction. In this situation, we need to know why only few search engines are leading in search market. What makes them better than others. The answer is hidden in the algorithm that is used to prioritize the documents. The algorithm decides which documents are relevant for the users and under what context

2. SEARCH ENGINE EVALUATION

Search engine evaluation refers to the analysis of working of the algorithm used in the background of search engine. It is the study of the parameters that make a search engine better than other [4,5]. It is performed to create best search engine with ability to reduce user efforts in searching and to enhance the users‟ satisfaction as well.

 DOI:10.5121/ijci.2026.150129

How search engine decides which documents are relevant and which are not relevant for users is the major issue. How search engines fix the relevancy level of documents and assign them priority in ranked list is central concern.

Documents are often evaluated with two types of feedback named explicit feedback and implicit feedback [6,7]. Explicit feedbacks are the experts‟ judgments about the relevancy of documents [8]. They fix the relevancy level of documents. This type of relevancy collection requires huge time. This type of feedback are not real type as feedback are not coming directly from the real users as a result the same information is relevant for someone and it is irrelevant for some others under same query. Explicit judgments require large number of judges of various fields that is another serious problem. The size of World Wide Web is continuously expanding and assessing of all uploaded documents in synchronized way is not possible. In this way, we can say explicit judgments are not adequate to evaluate and prioritize the web documents.

To resolve the aforesaid issues with explicit feedback, implicit judgments came in existence. In implicit judgments, none experts are required for relevance grades computation [9,10,11]. These are automatic signals that tell about relevancy about the web page. With implicit feedback we do not need to evaluate the documents manually instead it is achieved automatically [12,13]. Suppose a page is frequently being visited by large number of people across the globe means user finds something relevant in that page. If a page is opened for minutes to hours having mouse clicks indicates that page contains relevant material for searchers. If page is visited till depth with the help of scroll bar and the action like clicking on save/print button in tool bar signals that page is relevant for users. With the help of implicit users‟ feedback, evaluations of documents get starts when it is uploaded to data repository.

3. USER FEEDBACK DEPENDENT RANKING PARAMETERS

Search engines are compared and evaluated on the basis of some parameters. Cleverdon suggested six ranking parameters that have been largely cited in last few years and still are being considered to measure the performance of web search engines. These parameters are coverage, time lag, recall, precession, presentation and user efforts [12]. The coverage, time lag, recall, precession are the statistical measures to judge the relevancy of results as well as the quality of search engines [13]. Coverage refers to total number of documents available in database to which searching process is performed [14]. This is an important measure as the database size is continuously expanding and if this is not updated, the retrieval of new relevant documents will not be possible. Time lag denotes the total time taken by end user to get its relevant and desired result from the query posing to search engines. The ratio of relevant documents with in the collection of given number of documents to the number of retrieved documents is called precision [13,14]. Recall refers the ratio of the number of relevant documents retrieved to the total number of relevant documents available in the database [13,14]. Both the measures precision and recall have some problem. The documents relevancy must be known in advance. Obviously few can exist which are marginally relevant or somewhat relevant. Few others may be closest relevant and completely irrelevant in the web. This problem is very complicated and completely depends upon individual perception: what is relevant to one person may not be relevant to others”. There are few parameters on the basis of which the search engines are evaluated and ranking is performed (1) dwell time (2) Session duration (3) query formulation (4) Users Satisfaction (5) Search Length [7,9].

4. REAL TIME USERS SEARCHING APPROACH

This article examines users searching process and presents a two dimensional traversing approach. Most of search engines return the web links with the index page address that contains only the introductory information, the main information becomes hidden in the sub links of index page, as a result users are required to go through some unnecessary sub links to achieve the relevant, desired and satisfactory information [7, 9]. Suri [8] et al. tried to explain the search length successfully and proposed a metric which deals with links existing on landing page only.
Parallel links or „Root links" are the links which are retrieved in the form of list, obtained after a search. These are links that are visible to the users directly with a short explanation of the text involved inside the links. Series links are the sub links that are embedded in the all the pages under same domain name. In most of the cases these are the links from the single website. The concept behind the series links is that a page becomes relevant if it contains only the links of the pages that hold the desired information. These links act as bridge through which users can achieve their desired information.

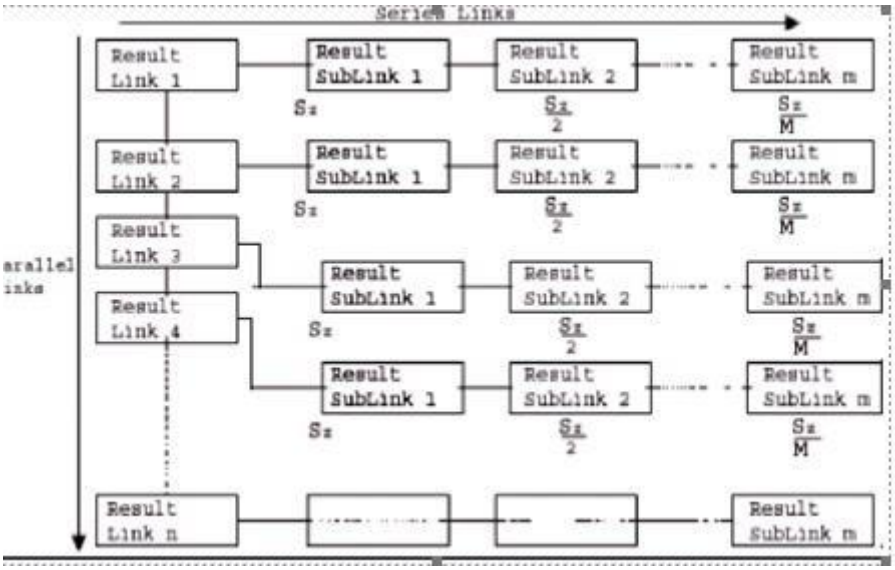

Fig. 1. Two Dimensions Traversing Approach.

The basic concept behind proposed metric is originated from the two dimension traversing apporach of searchers during information retrieval process. Our evaluation method is expected to differentiate the search engines and facilitate the users to be capable of choosing best one among various leading information systems.

5. FORMULIZATION OF METRIC

In information retrieval area of research, various measures and metrics have been suggested but we are still away from satisfactory achievement. To evaluate the search engines, we have proposed an HTML independent evaluation metric which holds the capability to filter the combined relevancy (Relevancy Thickness) of whole or top ranked relevant documents. It is noted that we also involved the relevancy of partial relevant results as well as somewhat relevant results. We considered that junk results are irrelevant having 0 as relevancy score.

We utilized the two dimension users traversing approach [7,9] to derive the metric which uses the relevant thickness. Relevant thickness is the quantity of the similar types of results having different relevancy score. Relevant thickness is the coverage area of web for a particular query.

We derive the metric in five steps.

Step-1:

$$A_1 = \{(n+1) - r_j\} \quad (1)$$

Where n = first n documents displayed under consideration

- r_j= rank of the jth document.
- RP= Ranked Precession
- S_z = Sub Link‟s Relevance Level
- b_j = Broken link
- A1 works as head that points the parallel links.

Step-2:

$$A_2 = \left(\sum_{z=1}^{z=m} \frac{S_z}{z} + w_j\right) \quad (2)$$

Muh Chyun Tang [10] declared that a result which holds the links for satisfactory and desired results, becomes partially relevant. Equation 2 is based on the same concept, which utilizes the variable S_z to assign the relevance score for the sub links found to be exists in way to achieve the most relevant results (m^{th} result). As search length is conversely proportional to search engine performance i.e. the increment in the search length results decrement in search engine performance. To perform this task we divide the S_z by z. S_z is not necessarily to be the same for all the middleware results. The value of S_z depends upon the quality of middleware results. In equation 2

w_j is the relevancy score which is provided by us to the m^{th} result. m^{th} is actually most relevant result originated from the root link.

Step-3:

Multiplying both equation and including bj, we found

$$A_3 = (A_1 * A_2) * b_j \quad (3)$$

b_j is the variable whose value becomes 0 if j^{th} root result is found broken (404 error). This variable helps to minimize the calculation because if system initially finds that root link is broken then, it will stop the calculation for later stages.

To calculate the total relevant thickness, we extend the proposed metric for all n results to be judged for relevancy level

Step-4:

$$A_4 = \sum_{j=1}^{j=n} (A_1 * A_2) * b_j$$

Or

$$A_4 = \sum_{j=1}^{j=n} \{[(n+1) - r_j)] * (\sum_{z=1}^{z=m} \frac{s_z}{z} + w_j) * b_j\} \quad (4)$$

(4)

To find the total relevant thickness for particular search engine. We have divided the equation (4) by $\frac{n(n+1)}{2}$ it is noted that we are interested only in calculating the relevant thickness of search engine instead the average relevance score.

Step-5:

$$RP = \frac{\sum_{j=1}^{j=n} [\{(n+1) - r_j)\} * (\sum_{z=1}^{z=m} \frac{s_z}{z} + w_j) * b_j]}{\frac{n(n+1)}{2}} \quad (5)$$

(5)

Equation (5) is metric that can be utilized to calculate the relevant thickness of search engine on the basis of retrieved links. The metric is less calculative and is capable to differentiate the search engines on the basis of relevant thickness. This metric can also be used for search engine evaluation but the main problem with this metric is that it wholly depends on the experts judgments and becomes fail in the absence of human editorial grades. The metric discussed above is very simple and totally based on user traversing approach between links but there are several modification are possible with this metric. First, it is not following the basics of cascade model. Second, it becomes fail when user judgments are not available.

6. CONCLUSION

The paper identifies six parameters that can be used to evaluate the information retrieval software. The real time searching approach is modeled to know the users efforts in searching and the effect of search length and clicks in ranking of documents. Two dimensional users traversing approach has been used to derive the ranked precession metric. The novelty of this paper is its

algorithm that uses implicit and explicit feedback to prioritize the documents in repository. The paper explains that how ranked precession metric is capable to reduce the possibility of HTML based META Tagging based forge ranking.

REFERENCES

[1] B. Carterette and R. Jones. "Evaluating Search Engines by Modeling the relationship between relevance and clicks". Proceeding of the 18th ACM conference on Information and knowledge management. 2009, pp.217-219, New York, USA.

[2] E. Agichtein, E. Brill, S. Dumals. "Improving web Search Ranking by Incorporating User Behavior Information" Proceedings of SIGIR. 2006, pp. 19-26.

[3] Joachims T. "Optimizing search engines using clickthrough datProceedings of the ACM Conference on Knowledge Discovery and Data Mining (KDD). 2002, pp.11 3-117.

[4] Chapelle, Metzler, Zang and Grinspan., "Expected Reciprocal Rank for Graded Relevance", Proceeding of the 18th ACM conference on Information and knowledge management. 2009, pp.217-219, New York, USA.

[5] Olivier Chapelle and Ya Zhang. "A Dynamic Bayesian Network Click Model for Web Search Ranking" Proceedings of the 18th international conference on World wide web. 2009, pp.217-219.

[6] Cooper, W.S. "Excepted search Length: A single measure of retrieval effectiveness based on the weak ordering action of retrieval systems", Information Journal of American Society of Information Science, 1968, pp. 30-41

[7] Sanjay K. Dwivedi and Rajesh Kumar Goutam "Factors Affecting Information Retrieval", Proceedings of the Advances in Computer Science Conference. 2010, pp.217-219. Trivandurm, India.

[8] Suri, Rakesh kumar, R.K Chauhan "Search Engine Evalution", DESIDOC Bulletin of Information Technology, 2005, pp. 3-10.

[9] Sanjay K. Dwivedi and Rajesh Kumar Goutam "Evaluation of Search Engines using Search Length", Proceedings of the International Conference of computer Modeling and Simulatione. 2011, pp.502-505. Mumbai, India.

[10] Muh-Chyun Tang and Ying Sun., "Evaluation of Web-Based Search Engines Using User-Effort Measures", Journal of the American Society for Information Science, 2000, pp. 493-503.

[11] N. Craswell, O. Zoeter, M. Taylor and B. Ramsey "An Experimental Comparison of Click Position-Bias Models." Proceedings of the international conference on Web search and web data mining. 2008, pp.217-219. New York, USA.

[12] Cleverdon, C.W., Mills, J., and Keen, E.M. (1966), "An inquiry in testing of information retrieval systems", Cranfileld, U.K. Aslib Cranfield Research Project, College of Aeronautics.

[13] Gwizdka, J. & Chignell, M. (1999), "Towards Information Retrieval Measures for Evaluation of Web Search Engines", IML Technical Report. [14] Chu, H., Rosenthal M. (1996), "Search engines for the world wide web: a comparative study and evaluation methodology", In Proceeding of the Annual Conference for the American Society for Information Science, pp.127-135.

AUTHOR'S

Dr. Rajesh Goutam is currently serving as an Assistant Professor in the Department of Computer Science at the University of Lucknow, Lucknow. He earned his Ph.D. from Babasaheb Bhimrao Ambedkar (A Central) University, Lucknow. With extensive experience in teaching and academic research, he has authored numerous research papers published in prestigious international journals and has presented his work at several high-impact academic conferences.

Pankaj Nigam is an Assistant Professor in the Department of Computer Engineering & Applications at Mangalayatan University, Aligarh (U.P.), India. He received his MTech in Computer Science from Babasaheb Bhimrao Ambedkar University (A Central University), Lucknow, India. He pursued his B.Tech in Information Technology from Krishna Institute of Engineering & Technology, Ghaziabad (U.P.), India. His research interests include AI-driven threat detection and automated defense systems.

Somya Tyagi is associated with the Department of Computer Science and Engineering, with a strong technical background in Web technologies. Her research interests focus on ranking search engine results, emphasizing methods to improve relevance and efficiency in information retrieval systems.