IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 5, Issue 3, March 2016

Hybrid Usage-Based Ranking Model for Information Retrieval (Web Search Engine)

Nikita Patil[1], Neha Raverkar[2], Tanvi Sardare[3], Manisha Upadhyay[4], Prof. Shweta Barshe[5]

Bharati Vidyapeeth College of Engineering Navi Mumbai, India[1,2,3,4,5]

Abstract: There are plenty of web pages available on the Internet. It is very challenging for Web Search Engines to find precise result up to user's expectations for his query from those huge numbers of pages. Most of the search results that correspond to a user's query are not relevant to the user browsing behaviour. Lots of the page ranking algorithms are available to calculate the relevancy of the web page such as Weighted PageRank algorithm or Content Based Ranking to provide a good ranked list. But these ranking algorithms might be not enough to provide a good ranked list, so in this paper, we have come up with a Hybrid Usage-Based Ranking Algorithm where page ranking is generated as a combination of user usage (user hits per link) and vector space result for users query. The result helps the users to get the relevant information quickly.

Keywords: Search engine, Information Retrieval (IR), Usage based, Vector space, Visits of Links (VOL)

I. INTRODUCTION

A web search engine is said to be a software system which is devised to search for information on the internet. This information can be a combination of images, web pages and various other files. Plenty of users connect with search engines daily which has become an integral part of their lives. They reformulate their queries, visit some of the links in the results, spend time on web pages, and perform other actions. These interactions may result into some a important information source to tune and improve search result ranking. On contradictory, others prefer the traditionally information retrieval (lR) scenario, a user formulates a search query and triggers a retrieval process which results in a list of ranked documents in decreasing order of relevancy .

Most of the current Information Retrieval Systems still depends on various approach of ranking algorithms like Content-based ranking algorithms that use the words belonging to each document to predict its ranking ;Link-based ranking algorithms designate scores documents depending upon the number and quality of hyperlinks between pages ;Usage-based ranking algorithms assigns score to documents by how often they are viewed by Internet users.

In some systems that do use the usage data in ranking can determine the relevance of a web page by its selection frequency. This measurement is not that proper to indicate the real relevance , so it is worth further exploration on how to apply this kind of actual user behavior to the ranking mechanism. The aim of the paper is to provide a hybrid ranking algorithm to utilize the usage data called Hybrid Usage-based Ranking Algorithm. This ranking algorithm is to improve the ranked list provided from search engines that based on **combination of user usage (user hits per link) and vector space result for users query**. The improvement is crucial to study, because it will reflect on the effectiveness of Information retrieval systems and web search engines.

The organization of the paper is as follows: Section II presents brief idea about research background**.** Section III discusses the proposed work. Finally, the conclusion and future work appears in Section IV.

II. RESEARCH BACKGROUND

Ranking search results is a fundamental problem in information retrieval. Most common approaches primarily focus on similarity of query and a page, as well as the overall page quality. However, with increasing popularity of search engines, the capturing of user behaviours insists to appear on the surface more. Much information such as links user's click how long users spend on a page and the user's satisfaction degree from the relevance of the page could be estimated. It is actually kind of implicit feedback (i.e., the actions users take when interacting with the search engine), such kind of usage data could be used to improve the rankings.

A lot of work has been done on the implicit measures of user preference in the field of IR (i.e. implicit feedback in IR), One of the earliest evaluations of time aspects was presented by Morita et al. in 1994. Their experiments showed a positive correlation between user interest and the reading time of articles. In addition, they found a low correlation between reading time and the length and readability of an article [1].

Kritikopoulos et al. was studied method in [2] for evaluating the quality of ranking algorithms. Success Index considers a user's click-through data, the result implies their method is better compared to explicit judgment.

Another study was presented by Rekha et al. in 2011. This study was provided a new model to find a user's preferences from click-through behaviour and using the exposed preferences to adapt the search engine's ranking function for improving search service.

DOI 10.17148/IJARCCE.2016.5364

IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 5, Issue 3, March 2016

In this proposed model, the combination of viewed and stored document summaries is used. The results show that this combining improved the reliability of ranked-list than ever was [3].

SimRank [4] is a new page rank algorithm which is based on similarity measure from the vector space model and is used to assign the more relevant score to the web pages.

Tuteja's study in [5] was based on user behaviours which enhanced the weighted PageRank Algorithm by taking into account Visits of Links (VOL) done by the end of 2013. This research idea presented as customizing the standard Weighted PageRank algorithm by integrating Visits of Links. Some usage behaviour factors included in this research to VOL like:

- Time spent on web page corresponding to a link: The algorithm must assign more weight to the link if more time is spent by the users on the web page corresponding to that link. Most of the times, the time spent on the junk pages is very less as compared to relevant pages. So this factor will help in lowering the rank of junk pages.

- Most recent use of link: The link which is used most recently by users should have more priority than the link which has been not used so far. So most recent use of link can also be used to calculate the page rank. The result shows that adding number of visits of links (VOL) to calculate the values of page rank holds to be more relevant results are retrieved fist. In this way, it may help users to get the relevant information much quicker.

III. PROPOSED WORK

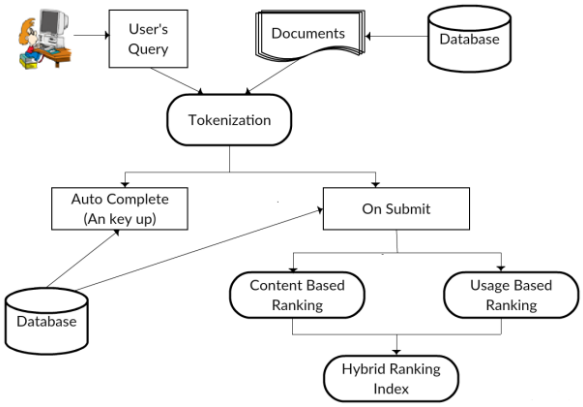
Section A : Outline of Algorithm

Input : User Query
Output : Relevant documents.

Step 1 : A repository (database) of web pages is created
Step 2 : After creating the database a link structure will be created that will explain how pages are linked to each other.On the basis of links, page rank will be calculated for each page at the beginning.
Step 3 : User will add a query and database will be searched for the pages related to user query.
Step 4 : Pages will be searched for user query. Web Pages will be selected on the basis of their similarity content and those are similar to user search will be selected for user. Web Page similarity will be calculated using modified Sim-Rank technique i.e. content based rank.
Step 5 : After having the web pages those are matched with user query, their page ranks will be compared. Pages with high page and content rank will be placed on top of the search result list. To build our final search list we will consider both web pages content and links.

Section B : Module Description

According to Figure, our system consists of several modules :



Dig : System Architecture

Module 1 : Tokenization : This stage is for breaking a stream of text up into words, and keeping the words in a list called Word's List.

Module 2 : Auto Complete : It provides the suggestions while the user types into the field and shows it in a dropdown format.

Module 3 : On Submit : When the user submits his query, the database will be searched for the pages related to user query.

Module 4 : Content-Based Ranking : the user's query is matched with the index terms to get the relevant documents to the query. Documents are then ranked as per Vector Space Algorithm according to the most relevant to the user's query.

Vector Space Algorithm
This algorithm focuses on ranking the documents based on user query. Tf-idf denotes the term frequency-inverse document frequency, and the tf-idf weight is a weight mainly used in information retrieval and text analytics. This weight is a arithmetical measure used to calculate how significant a word is to a web pages in a corpus. The significance increases reciprocally to the no. of times a word occurs in the document but is balanced by the frequency of the term in the collection. Variety of the tf-idf weighting schemes are used by search engines as a crucial tool in scoring and ranking a web page's utilization given a user query.

The ranking function is computed by taking summation of the tf-idf for each query; many more refined ranking functions are alternatives of this model.

How to Compute :-
More often, the tf-idf weight is evaluated by two terms :

- **TF: Term Frequency**, which measures how seldom a term occurs in a document. As every document is different in length, it may happen that a term would appear more frequently in long documents compared to shorter ones. Thus, the term frequency is usually divided by the the total number of terms in the document as a way of standardization:

DOI 10.17148/IJARCCE.2016.5364

IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 5, Issue 3, March 2016

- TF(t) = (No. of times term t appears in a document) / (Total no. of terms in the document).
- **IDF: Inverse Document Frequency**, which evaluates the importance of the term. While calculating TF, all terms are considered equally significant. However it is obvious that certain terms, such as "of", "is" and "that", may appear frequently but are not that important. It is required to scale up the rare terms while weigh down the frequent ones, by calculating the following:

IDF(t) = log_e(Total no. of documents / No. of documents with term t in it).

See below for a simple example.

Example :
Consider a document which contains 200 words where the word bank appears 8 times. The term frequency for bank is then (8 / 200) = 0.04. Now, assume there are 20 million documents and the word bank appears in two thousand of these. If the inverse document frequency is evaluated as log(20,000,000 / 2,000) = 4 then the Tf-idf weight is the product of these quantities: 0.04 * 4 = 0.16.

Module 5: Usage-Based Ranking : In this stage we calculate frequency weight which are the inputs to our algorithm. The calculations of frequency weight are discussed in detail below.

Frequency of visit that determine the relevance of a web page by its selection frequency, in order to find the frequency weight, which is The admittance frequency of a page u, is the number of times the page is visited and the page rank which is appear in the ranked list from the previous stage. The frequency weight formula is:

$$FW = \frac{Number\ of\ Visit\ on\ a\ Page(u)}{Total\ Number\ of\ visit\ on\ all\ Page} \times PR(u)$$

Where:

FW: Frequency Weight.

PR(u): The Page rank of a page u.

Module 6: Hybrid Ranking : In this section we introduced the Hybrid Usage Based Ranking Algorithm, explains how it is works and shows the equations used. This is the final stage in our Hybrid Usage Based Ranking Algorithm, it basically used the parameters that calculated in the previous stage to find the usage-Based weight which is equal the new weight for each Page, this weight used to re-rank the pages and the effective reflects on the pervious Rank list to get a new ranked list. So the result is a new search engine appears to the user.

As a result Hybrid Usage Based Ranking Algorithm depends on usage parameters and the ranking from Content-based algorithm results, their combination provided a new weight for a pages in order to re-ranking them as a new ranked list appear to the user.

IV. CONCLUSION AND FUTURE WORK

Due to massive amount of information available on the web, the users end up spending lot of time to get the pages relevant to them. So the proposed algorithm makes use of user usage (user hits per link) and Vector Space result to calculate the page rank so that more relevant results are retrieved first. The result helps the users to get the relevant information quickly. The Hybrid Usage Ranking Algorithm improves the efficiency of web searching by overcoming the drawbacks of the ranking algorithms.

The system can be further enhanced and several other functionalities can be added by implementing online method. This framework can be further updated by using the query log data, which provides valuable co-existence information of keywords.

REFERENCES

[1]. Hofgesang,P.,"Relevance of Time Spent on Web Pages,in 'Workshop on Web Mining an Web Usage Analysis", the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD 2006),2006.

[2]. Kritikopoulos A., Sideri M. an Varlamis I., "Success index: measuring the efficiency of search engines using implicit user feedback", in: Proceedings of the 11th Pan-Hellenic Conference on Informatics,Special Session on Web Search an Mining,2007.

[3]. Rekha c., Usharani J. and Iyakutt K., "Improving the Information Retrieval System through Efleive Evaluation of Web Page in Client Side Analysis", International Journal of Computer Applications, Vol. 15,No. 6,p.p. 35- 39,2011.

[4]. Qiao, S., Li, T., Li, H., Zhu, Y., Peng, J., Qin, J., "SimRank : A Page Rank Approach based on Similarity Measure", Published in IEEE, Print ISBN No: 978-1 - 4244 -6793-8, 2010, pp. 390-395.

[5]. Tuteja S., "Enhancement in Weighted PageRank Algorithm Using VOL",Journal of Computer Engineering,Vol. 14,Issue 5,pp. 135-141,2013.