International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor :6.887
Volume 5 Issue XII December 2017- Available at www.ijraset.com

Efficient Hybrid Usage-Based Ranking Model

Mrs. Kanchanmala D. Talekar[1], Prof. Mayank Bhatt[2]
[1, 2]*R.I.T. Indore*

Abstract: ***There are billions of site pages accessible on the Internet. Search engines have an issue for the best rated list to the user's query from those huge amounts of pages. A lot of search results that equivalent to a user's query aren't relevant to an individual need. The majority of the page rank algorithms use Link-based ranking (web composition) or Content-based ranking to estimate the relevancy of the info to the user's need, but those rank algorithms might be insufficient to give a good rated list. In this way, in this paper we proposed an Efficient Hybrid Utilization based Ranking Algorithm called EHURA. EHURA was applied to 1033 English Corpus to measure its performance. The result shows improvement of the accuracy for using EHURA over the Content-based ranking algorithm rendering while realizing approximately the same recall percentage.***
Keywords: **Information Retrieval (IR), Tokenization, Usage-based Ranking, Content-based Ranking, Link based Ranking.**

I. INTRODUCTION

Browsing becomes a normal habit in our life. Millions of users connect to search engines daily. They are following some of the hyperlinks in the final results, click on ads, spend time on pages, reformulate their queries, and perform other activities. These interactions may problem some a valuable source of information for tuning and enhancing search outcome ranking and can complement more costly explicit decision. On the other hand, others choose the traditionally information retrieval (IR) scenario, a user formulates research online query and triggers a retrieval process which results in a set of ranked documents in decreasing order of relevance. Most of the existing Information Retrieval Systems still relies on various approaches of ranking algorithms, like Content-based ranking algorithms that apply the words in each document to determine its ranking; Link-based ranking algorithms assign scores to web pages based on the number and quality of hyperlinks between web pages. Links that point to a particular page or recommend a page can help to improve link-based rankings; Usage-based ranking algorithms rating documents by how often they are viewed by Internet users. Regarding Usage-based ranking, there are limited works to utilize the usage data in the web information retrieval systems, especially in the ranking algorithm. For few systems [2] and [3] that use the utilization data in ranking, they determine the value of a web page by their selection frequency. This measurement is not that accurate to show the real relevance. The time spent on reading the page, the procedure of saving, printing the web page or adding the web page to the bookmark, and the action of following the links in the web page, are all good indicators, perhaps better than the simple selection frequency. Therefore it is worth further exploration how to apply this kind of real user behaviour to the ranking mechanism. The objective of the paper is to provide a hybrid ranking algorithm to utilize the usage data called EHURA (Efficient Hybrid Usage-based Ranking Algorithm). This ranking algorithm to improve the ranked list provided from search engines that based only on content base ranking. The improvement is important to study, because it will effect on the effectiveness of Information retrieval systems and web search engines.

II. RELATED WORK

Ranking search results is a fundamental problem in information retrieval. Most common approaches mostly give attention to similarity of query and a web page, as well as the overall page quality. On the other hand, with increasing popularity of search engines, the recording of user behaviours demands to appear on the surface more. Much information such as links user’s click how long users spend on a webpage and the user's satisfaction degree from the relevance of the page could be approximated. 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 enhance the rankings [4] [5] [6] [7] [8].A great deal of work has been done on the understood measures of user preference in the field of IR (i.e. implicit feedback in IR), One of the soonest assessments of time perspectives was introduced by Morita et al. in 1994. Their examinations demonstrated a positive relationship between user interest and the perusing time of articles. Likewise, Usage-based ranking Algorithm was presented by Ding et al. in 2002 for web Information Retrieval systems that applies time pent on page against standard selection- frequency based ranking, i. e. the basic idea of rank score is calculated on the time users spend on reading the page and browsing the connected pages, the high- ranked pages may have a negative adjustment value if their positions couldn't match their actual usage, and the low-ranked pages may have a positive adjustment value if uses tend to dig them out from low positions [2].

Based on the study of Kellar et al. 2004 [10] focused on the relation between web search tasks and the time spent on reading results. Their results support the correlation and show that it can be even more powerful as the complexity of a given activity increases. Agichten et al. (2006) researched user behavior data to improve ordering of results real web search settings. Their report involved over 3000 queries and 12 million user interactions with a popular web search engine, the results of this study show the accuracy of entering end user feedback term was increased in comparing with the original ones [4].Tuteja's study in [11] was structured on user behaviours in order to improve the measured Page Rank Algorithm by considering a term Visits of Links (VOL) created by the end of 2013. This kind of research idea presented as modifying the standard Measured Page Rank algorithm by including Visits of Links. A few consumption behavior 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 put in by you on the online web page corresponding to that particular link. Many of the times, enough time spent on the junk pages is very less when compared with relevant pages. As a result this factor will help in lowering the rank of junk pages. Most recent use of link: The link which can be used most recently by users should have more priority than the hyperlink which has been not used so far. Therefore latest use of website link can be used to compute the page rank.

III. PROBLEM STATEMENT

The problem is on combining different ranking algorithms in order to design an effective hybrid ranking approach that uses a combination of content based, link based and usage based ranking algorithms such that it meets user specific needs and goals. In order to narrow down the ranked list even further to meet the user specific search goals the reordering of the set of top n ranked pages is imperative using re ranking algorithm.

IV. PROPOSED SYSTEM

A. Input

User Query

B. Output

Relevant documents.

1) *Step 1:* A repository (database) of web pages is created
2) *Step 2*: After resulting in the database a link structure will be created that will describe how web pages are linked to each other. On the most basic of links, page rank will be calculated for each and every page at the beginning.
3) *Step 3*: User will add a query and database will be searched for the pages related to user query.
4) *Step 4:* Pages will be search user query. Web pages will be selected on the basis of their similarity content and that similar user search will be selected for user. Website similarity will be calculated using modified Sim-Rank technique i. e. content based rank
5) *Step 5:* After getting the web pages those are matched with user query, their web 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. According to Figure 1, our system consists of several modules; we divided them into two Phases:
6) *Phase I*: Document Pre-processing Phase consist the following Modules:
7) Module 1: Tokenization: this stage for breaking a stream of text up into words, and keeping the words in a list called Word's List.
8) *Module 2*: Data Cleaning: it removes useless words from the Word's List; These useless words are stored in a stop words database as appear in the figure. The database has 311 stop words with a size 3KB.

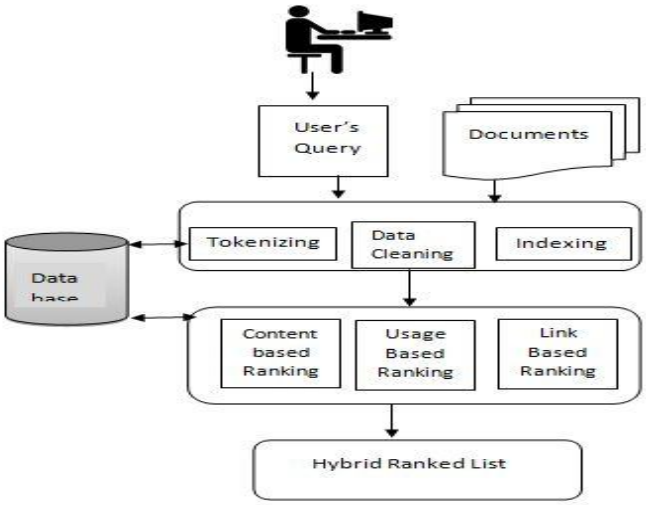

Fig1: System Architecture

1) *Module 3:* Log Files Analysis: it removes irrelevant records from Log file. In order to enhance the efficiency of usage based retrieval algorithm by a useful records only. Log file Analysis consist a series of process like data cleaning, user identification, session identification as appear in Figure 1, To clear those process see section A.
2) *Module 4:* Indexing: Indexing is a process for describing or classifying a document by index terms; index terms are the keywords that have meaning of its own (i.e. which usually has the semantics of the noun). This index terms are grouped in an indexer and stemmer is service this stage by improving the group of these keywords in the indexer.
3) *Phase II:* Ranking Phase (EHURA) consist the following Modules:
4) *Module 5:* 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 using ranking algorithms according to the most relevant to the user's query.
5) *Module 6*: Usage-Based Parameters: In this stage we calculate several parameters which are the inputs to our algorithm. These parameters are presents in detail in section B Number 2.
6) *Module 7:* Usage-Based re-ranking: it's the combination of the pervious modules to provide a new weight called usage based weight for the pages, then ranking those pages according to their new weight.

V. IMPLEMENTATION DETAILS

A. *Content-Based Ranking*

Cosine measure calculates the angel between two documents (between document and user's Query which is treated as a document). Each document represented as vector. Thus a cosine value of zero meant that the query and document vector were orthogonal to each other and meant that there was no match or the term simply did not exist in the document being considered. To know cosine relation between two documents (document D and query Q) see Equation below.

$$Con\sin e(D,Q) = \frac{\|D \cap Q|}{\sqrt{|D| * |Q|}}$$

B. *Where*

Cosine (0, Q): the Cosine Similarity relationship between document 0 and user's query Q. 0: refer to the document in the collection. Q: refer to user's query Usage-Based Parameters In this stage the system calculates two Usage-Based Parameters as in the following: 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:

$$FM = \frac{number\ on\ visit\ on\ page(u)}{Total\ number\ of\ visit\ on\ all\ pages} \times PR(u)$$

FW: Frequency Weight. PR(u): The Page rank of a page u. Time Spent that shows how long the users spend on a page after removing the download time of the page

$$TW = \frac{time\ spent\ on\ a\ page(u) - download\ time(u)}{\max(time\ spent\ on\ page(u) - download\ time(u))}$$

Where: TW: Time Spent Weight.

$$download\ time(u) = \frac{size\ of\ \ page(u)}{transfer\ rate\ \ for\ page(u)}$$

Usage-Based re-ranking This is the final stage in our EHURA algorithm, it's basically used the two 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 EHURA depends on usage parameters and the ranking from Content-based algorithm results, their combination provided a new weight for a pages in order to reranking them as a new ranked list appear to the user.

VI. CONCLUSION

The World Wide Web (WWW) is rapidly and exponentially growing on all aspects and is a massive, explosive, diverse, dynamic and mostly unstructured data repository. As on today WWW is the huge information repository for knowledge reference. There are a lot of challenges in the Web: Web is large, Web pages are highly semi structured, and Web information tends to have diversity in terms of meaning, degree of quality of the information extracted and the conclusion of the knowledge which is extracted from the information. So it is important to understand and analyse the underlying data structure of the Web for efficient information retrieval. Thus web search ranking algorithms play a vital role in ranking of the web pages so that we could retrieve the web pages that are relevant. With the rapid growth of the information sources we are drowning in data but starving for knowledge therefore it has become necessary for the user to use information retrieval techniques and combination of different ranking algorithms to find and extract and filter the desired information. Many of the existing Information Retrieval Systems still relies on various approach of ranking algorithms, like Content-based ranking algorithms, Link-based ranking, or a few of them based on utilize user behaviours via usage-based ranking algorithm. Unfortunately, those ranking algorithms still have some drawbacks to a ranked list provided from some search engines. A combination of these algorithms enables to filter out the redundant results to gather the useful information.

REFERENCES

[1] Safaa i. Hajeer, rasha m. Ismail, nagwa l. Badr, m. F. Taiba," an efficient hybrid usage-based ranking model for information
[2] Retrieval systems & web search engine", 20156th international conference on information and communication systems (icics).
[3] Ding, c., chi, c.-h., and luo, t., "an improved usage-based ranking", waim '02: proceedings of the 3rd international conference on advances in web-age information management, london, uk: springer-verlag, p.p. 346- 353, 2002. .
[4] Rodriguez-mula g., garcia-molina h. And paepcke a, "collaborative value filtering on the web", computer networks, vol. 30 no. 8, pp. 736-738, 1998
[5] Agichtein e., brill e. And dumais s., "improving web search ranking by incorporating user behavior", in proceedings of the acm conference on research and development on information retrieval (sigir),2006
[6] Weiler a., "infonnation-seeking behavior in generation y students:motivation, critical thinking, and learning theory", journal of academic librarianship, vol. 31 no. 1 p.p.46-53, 2005.
[7] Taherizadeh s. And moghadam n., "integrated web content mining into web usage mining for finding patterns and predicting user's behaviors", international journal of information science and management, vo1.7, no.1 pp. 51-65,2009.
[8] Sanderson, m., paramita, m., clough, p. And kanoulas, e., "do user preferences and evaluation measures line up?", in proceedings of the 33rd annual acm sigir conference, geneva, switzerland, pp. 555-562,2010
[9] Konstan, 1. A, miller, b. N., maltz, d., herlocker, 1. L., gordon, l. R., and riedl, j., "grouplens: applying collaborative filtering to usenet news", communications of the acm, vol. 40, no. 3, p.p. 77- 87,1997
[10] Hofgesang, p., "relevance of time spent on web pages, in 'workshop on web mining and web usage analysis", the 12th acm sigkdd international conference on knowledge discovery and data mining (kdd 2006), 2006.
[11] Kellar, m., watters, c., duffy, 1., and shepherd, m., "eflect of task on time spent reading as an implicit measure of interest". Asist 2004 annual meeting, p.p. 168-175,2004
[12] Tuteja s., "enhancement in weighted pagerank algorithm using vol", journal of computer engineering, vol. 14, issue 5, pp. 135-141, 2013.