AN EFFECTIVE RANKING METHOD OF WEBPAGE THROUGH TFIDF AND HYPERLINK CLASSIFIED PAGERANK

Md. Mahbubur Rahman[1], Samsuddin Ahmed[2], Md. Syful Islam[3], Md.Moshiur Rahman[4]

[1]Department of CSE, Bangladesh University of Business and Technology, Bangladesh.
mahabub.cse.buet@gmail.com
[2]Department of CSE, Bangladesh University of Business and Technology, Bangladesh.
sambd86@gmail.com
[3]Department of CSE, Bangladesh University of Business and Technology, Bangladesh.
mahfuzisl@pstu.ac.bd
[4]Department of CSE, Islamic University of Technology, Bangladesh.
moshi.cse408@gmail.com

ABSTRACT

Web pages represent various information that is easy and efficient way to meet the user requirement. A large type of data contains in various web page from different website and domain. Now a day's these huge amounts of data create a large data crowd and from these crowds it is not so easy for a search engine to retrieve the perfect information that the user actually wants. The information that will meet the user requirement are contain in different website so the search engine needs to rank these web pages according to the significance of user query to the search engine. Different search engine use different techniques for ranking purpose. But all these technique does not fulfil the user requirements fully. The common using techniques are relevance ranking, Term frequency, Inverse document frequency, page rank. In this paper we propose a new formula for ranking the retrieved information that is a combination of TFIDF with hyper link classified page rank.

KEYWORDS

Page Rank, Hyperlink classification, Hyper Link Classified PageRank, HITS, TFIDF.

1. INTRODUCTION

With the rapid growth of internet, the number of document in World Wide Web is increasing. For the documents there are almost eight billion addresses according to the index of Google. People want to get useful and effective data with a shortest time in case of web query. So different techniques are invented to meet the user requirements. From the various searching method almost two factors that distinguish the high quality page and low quality page are relevance factor and the ranking factor. The relevance factors give the attention of the contents of the web whereas the ranking factor gives the attention on the web structure not the, contents. TFIDF is the ranking technique for the relevance rank and page rank. HITS algorithm is very well known ranking technique which hyperlink n based is ranking. But the later hyper link base ranking there are a common problem is topic drift problem. Many techniques are given to eliminate topic drift problem's PR weighted PR, Hyperlink classification but they cannot totally eliminate the topic drift problem. The paper is structured as follows: Section 2 introduces existing system. Section 3 presents our new approach. In section 4. We give conclusions and discuss ongoing work.

DOI : 10.5121/ijdkp.2013.3411

2. EXISTING TECHNOLOGY

2.1 TFIDF

Relevance ranking is not an exact science, but there are some well-accepted approaches. Given a particular term t, how relevant is a particular document d to the term. One approach is to use the number of occurrences of the term in the document as a measure of its relevance, on the assumption that relevant terms are likely to be mentioned many times in a document. Just counting the number of occurrences of a term is usually not a good indicator: First, the number of occurrences depends on the length of the document, and second, a document containing 10 occurrences of a term may not be 10 times as relevant as a document containing one occurrence.
One way of measuring $TF\ (d,t)$, the relevance of a document d to a term t, is

$$TF(d,t) = \log\ (1 + n(d,t)/n(d))$$

Where n (d) denotes the number of terms in the document and n(d, t) denotes the number of occurrences of term t in the document d.

2.2 HITS

HITS (Hypertext Induced Topic Search) is a ranking algorithm introduced by Jon Kleinberg in 1998 [4] that utilizes web graph's hyperlink structure to create two metrics associated with every page. The first is authority and the second is hub. Given a user query, the algorithm first iteratively computes a hub score and an authority score for each node in the neighbourhood graph. The documents are then ranked by hub and authority scores, respectively.
Documents that have high authority scores are expected to have relevant content, whereas documents with high hub scores are expected to contain hyperlinks to relevant content. The intuition is as follows.

A document which points to many others might be a good hub, and a document that many documents point to might be a good authority. Here, the figure shows the hub and authority of web page

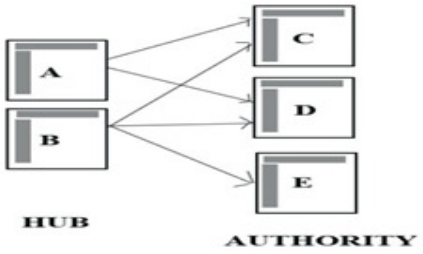

Figure 1.Hub and authority of web page

Recursively, a document that points to many good authorities might be an even better hub, and similarly a document pointed to by many good hubs might be an even better authority. This leads to the following algorithm: [6]

(1) Let, N be the set of nodes in the neighbourhood graph.
(2) For every node n in N, let H[n] be its hub score and A[n] its authority score.
(3) Initialize H[n] to 1 for all n in N.
(4) While the vectors H and A has not converged:

(5) For all n in N, A[n]:=∑ H[n].
(6) For all n in N, H[n]:=∑ A[n].
(7) Normalize the H and A vectors.

The main problem of the HITS algorithm is the Topic Drift problem. Topic drift is the major problem, because all the links in a web page is not important. But all the links contain the same value.

2.3 Page Rank

The Page Rank algorithm, one of the most widely used page ranking algorithms, states that if a page has important links to it, its links to other pages also become important. Brin and Larry Page first introduce Page Rank and a well-known method used by GOOGLE.
A slightly simplified version of Page Rank is defined as-[3]

$$\sum_{n}^{1}(PR(v)/N(v))$$

Page rank model is based on the Random surfer model. The original Page Rank is published as [3],

$$PR(I) = (1-d) + d\sum_{n}^{i=1}(PR(i)/N(i))$$

Where d is a dampening factor that is usually set to 0.85, (1-d) denotes the probability of hopping the next URL and d denotes the probability of stay on the same URL. Take the following 4 node page link to one another,

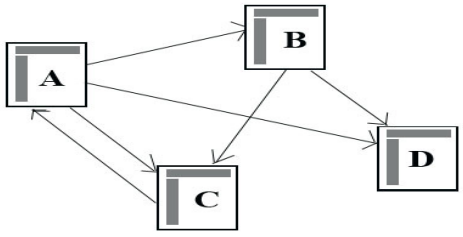

Figure 2.Link from one node to another.

In PR every link consider 1 on the matrix format,

	A	B	C	D
A	0	1	1	1
B	0	0	1	1
C	1	0	0	0
D	0	0	0	0

Figure 3.Link in the matrix form.

If a page has more than one link, then total summation is 1 for each link, that is,

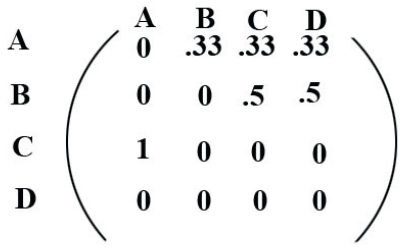

Figure 4.Setting value in the matrix.

The page rank value of one page is equally divided to its outgoing link as,

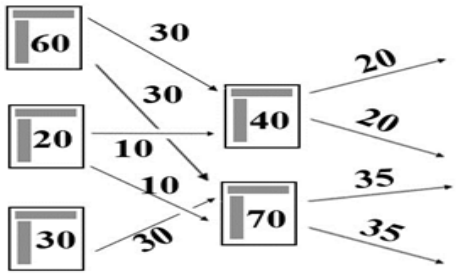

Figure 5. Rank is distributed in all pages.

The main drawback of Page Rank is that all the hyperlink count as a vote, so topic drift problem is also exists like HITS.

2.4 TS-Page Rank

Taher H. Havliwala first introduces the Topic Sensitive Page Rank to improve efficiency of Page Rank. In topic-sensitive Page Rank [2], precompute the importance scores offline, as with ordinary Page Rank. However, in TS-Page Rank compute multiple importance scores for each page; compute a set of scores of the importance of a page with respect to various topics. At query time, these importance scores are combined based on the topics of the query to form a composite Page Rank score for those pages matching the query.

To illustrate the method take the query: Bangla music download (3 words). By the TS-Page Rank method it creates set at 2^ (word), excluding empty (Power set).

e.g. {bangla, music, download, (bangla, music), (music, download), (bangla,download), (Bangla, music, download)}.

The main problem of the TS-Page Rank is that if the number of word is high, then the combination is also high and is time consuming. Another problem is that, it checks the pair of query word, unwanted result may come, e.g. after some rank Hindi music download page may come because it checks the pair music download.

2.5 Hyperlink Classification

Li Cun He et al. propose a Link Classification method to improve Page Rank [4] and carry out in the Chinese Internet environment. They categorize the hyperlinks into 4 classes as follows: (1) recommending Inner Link, the weight factor $\beta 1$ of Class, $\beta 1$ = 5; (2) other Inner Link including the commercial ads, $\beta 2$=1.0; (3) recommending outer Link, $\beta 3$=6.0 ;(4) other outer Link, $\beta 3$=1.0.

There is a big question is that how we understood the recommended and other link? They propose three steps [2], as follows:

Step1: Classify all the hyperlink inner links and outer links. If the source web URL of a hyperlink and the destination web URL belong to the same domain name, it is an inner link, or else it is an outer link.

Step2: For each inner link *i*, if PR score *(i)*>average inner link score, link *i* belongs to recommending Inner Link, or else link *i* belongs to other Inner Link.

Step3: For each outer link *i*, if PR score *(i)*>average outer link score, link *i* belongs to recommending outer Link, or else link *i* belongs to other outer Link.

We try to give a theatrical view of the Link classification method by taking three page A,B and C as bellow:

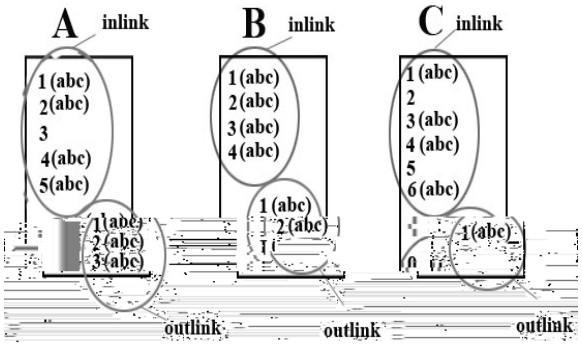

Figure 6: .Hyperlink Classification.

PR	Inner Link	Outer Link	PR(Inner Link)	Average	Value	PR(Outer Link)	Average	Value	Total
A	5	3	.125 .014 .212 .104 .085	.108	5 1 5 1 1	.212 .184 .185	.192	6 1 1	21 (1st)
B	4	2	.201 .189 .012 .125	.131	5 5 1 1	.205 .198	.205	6 1	19 (3rd)
C	6	1	.019 .102 .212 .189 .028 .104	.109	1 1 5 5 1 1	.184	.184	6	20 (2nd)

Table 1.Theoretical value of hyperlink classification

3. PROPOSED METHOD

3.1 Multiply the Hyperlink Classified PR with TFIDF

Our first propose is modified the hyperlink classified page rank multiply with the term frequency inverse document frequency (TFIDF). We think it can totally eliminate the topic drift problem because a page can contain various links but if we check every page TFIDF, then it may helpful for eliminate the topic drift.

Rank Position= (TFIDF*Hyperlink classified PR).

3.2 Theoretical Implementation

The given formula is theoretically implemented. We first select 100 different keyword and run a experimental searching through 3 top search engine that are Google, Yahoo and Bing.The retrieve list of information is manually checked by our group work and we found in many case that the expected document rank in lower position. But using these keyword we and the expected page we apply our techniques and most of the case we see that the expected page ranking in the top position. In experimental evolution of techniques we calculate three values of any web page, Hyperlink number, PageRank, TFIDF Value. We use the link extraction java program and run in online to extract inner and outer link of a given web page or website. We also use PageRank calculator and TFIDF calculator to find the PageRank and TFIDF value. From the beginning of our experimental evolution Suppose ‗A' is an important page but it has 5 links and ‗B' is another page which contain 20 links .Also we search the term ‗abc' page ‗A' contain ‗abc' all of its 5 links, but only 2 links of B contain ‗abc' .So in normal hyperlink classified PageRank, page ‗B' show first before page ‗A' .But in our propose method, all 18 links of page B except the 2 links will not consider for any significance they must be ignored and consider link value is 0. So, the total PageRank value of B is lower than the total PageRank value of A. So, the important page A show first in the rank window of a search engine. Now considering an example for three pages A, B and C, we give a theoretical value of our proposed method in table 2.

Launching experiment with theoretical data with pages A,B,C we first consider that page A have total 8 link in which it have 5 inner link and 3 outer link. Accordingly from the inner and outer link the recommended inner link or other inner link and the recommended outer link or other outer link can be calculate by PageRank score as describe in upper section [2.5]. Within 5 inner links there are 2 recommended inner link and 3 other inner links. Therefore for each recommended inner link considering the weight factor $\beta 1$ of Class, $\beta 1 = 5$, and assigning the value to the recommended inner links. For three other inner links considering the weight factor $\beta 2$ of Class, $\beta 2 = 1$ and assigning the value to the other inner links. Now considering outer links, I which there have 1 recommended outer link and 2 other outer links. Therefore for each recommended outer link considering the weight factor $\beta 3$ of Class, $\beta 3 = 6$, and assigning the value to the recommended outer link. For 2 other outer links considering the weight factor $\beta 4$ of Class, $\beta 4 = 1$ and assigning the value to the other outer links. Now finding the TFIDF of that page according to the respective links as for inner and outer links (including all recommended links and other links) as described in the formula in [2.1].

Now multiplying the value of term frequency inverse document frequency (TFIDF) to the value of hyper link classified PageRank value $\beta 1$, $\beta 2$, $\beta 3$, $\beta 4$ for each respective page. Now summing up both the values for inner links and outer links and finally adding the inner and outer PageRank value which will be the final value of PageRank for page A. Now follow the same procedure for page B and C and we get two final value for the page B and C. Now comparing these PageRank value of page A, B, C and which have largest value that page will

display in first position in the rank window of search engine. Accordingly other pages are display according to this method. We have shown in our theoretical experiment that there page B contain 4 inner link and 2 out links, but the entire page contain our required terms, so it's (TFIDF*Hyperlink classified PR) is higher than Page A and C.

Table 2.Theoretical value of Combined TFIDF and hyperlink classified Pagerank

Page	Inner Link (IL)	Outer Link (OL)	PR(IL)	Value	TFIDF	TFIDF* Value	Sum (IL)	PR(OL)	Value	TFIDF	TFIDF* Value	Sum (OL)	Sum(IL)+ Sum(OL)
A	5	3	.125 .014 .212 .104 .085	5 1 5 1 1	.027 .045 .000 .005 .012	.135 .045 .000 .005 .012	.197	.212 .184 .185	6 1 1	.011 .049 .002	.066 .049 .02	.117	.314 (2nd)
B	4	2	.201 .189 .012 .125	5 5 1 1	.021 .009 .015 .044	.105 .045 .015 .044	.209	.205 .198	6 1	.018 .002	.108 .002	.110	.319 (1st)
C	6	1	.019 .102 .212 .189 .028 .104	1 1 5 5 1 1	.021 .000 .019 .013 .00 .019	.021 .000 .095 .065 .000 .019	.200	.184	6	.013	.078	.078	.278 (3rd)

4. FUTURE WORK

In this paper we theoretically implement our formula, and create a java program that can finds the number of hyperlink of a web page, by which we can determine the number of inner link and the number of outer link. We want to practically implement our proposed method through an open source search engine and continuously search better techniques to find for an effective result for ranking.

REFFERENCE

[1] Abraham Silberschatz et al. "Database System Concept",seventh edition.

[2] Haveliwala T H. "Topic-sensitive Page Rank". IEEE Transactions on Knowledge and Data Engineering, Volume 15. August, 2003, pp.784-796.

[3] J. M. Kleinberg. Authoritative sources in a hyperlinkedenvironment. Journal of the ACM, 46(5):604–632, September 1999.

[4] Li Cun He, Lv K Quing."Hyperlink Classification: A New Approach to Improve Page Rank" School of Computer &Communication Engineering, China University of Petroleum, Dongying 257061, China.

[5] Michael Chau, Patrick Y. K. Chau, Paul J. Hu,"Incorporating Hyperlink Analysis in Web Page Clustering".

[6] Monika Henzinger, Google Incorporated, Mountain View,California," Link Analysis in Web Information Retrieval"

[7] S. Brin and L. Page. The anatomy of a large-scal hypertextual Web search engine. In Proceedings of the Seventh International World Wide Web Conference 1998, pages 107–117.

[8] Wenpu Xing and Ali Ghorbani," Weighted PageRank Algorithm".

[9] S. Chakrabarti, B. Dom, D. Gibson, J. Kleinberg, P. Raghavan, and S. Rajagopalan. "Automatic Resource compilation by analyzing hyperlink structure and associated text". Computer Networks and ISDN Systems, Volume 30,April 1998,pp.65-74

[10] J Furnkranz. "Hyperlink ensembles: A case study in hypertext classification". Technical Report OEFAITR- 2001-30, Austrian Research Institute for Artificial Intelligence, Wien, Austria, 2002

[11] MIZUUCHI Y, TAJIMA K. "Finding context paths for Web pages". Proceedings of the Tenth ACM Conference on Hypertext and Hypermedia. ACM Press New York, USA, 1999 , pp.13-22.

[12] Matthew Richardson, Pedro Domingos. "The Intelligent Surfer: Probabilistic Combination of Link and Content Information in PageRank", volume 14. MIT Press, Cambridge, MA, 2002.

[13] Decai Huang, "TC-Page-Rank Algorithm Based on Topic Correlation", Intelligent Control and Automation, 2006. WCICA 2006. The Sixth World Congress, 2006.

[14] Zhang Ji-Lin, "Webs ranking model based on Page- Rank algorithm", Information Science and Engineering (ICISE), 2010 2nd International Conference, 2010.

Authors

Md.Mahbubur Rahman, received his B.Sc.Engineering degree in computer science and engineering from Patuakhali Science and Technology University in 2011. He is now working toward his M.Sc. Engineering degree in computer science and engineering at Bangladesh University of Engineering and Technology (BUET), Bangladesh and working as a Lecturer in computer science and engineering at one of the top most private Universities in Bangladesh named Bangladesh University of Business and Technology (BUBT). His research interests are Data mining, Search Engine optimization, Biometric system, Network Security.

Samsuddin Ahmed has been lecturing in CSE since mid of 2010. He is in Computer Science and Engineering from University of Chittagong with highest CGPA till date. His under-grade Thesis was on "Handling Uncertainties in Spatial Feature Extraction". His hobbies include thinking about underlying mathematical formulations in natural phenomena. His research interests include data and image mining, Semantic Web, Business Intelligence, Spatial Feature Extraction etc. He is now serving one of the top most private Universities in Bangladesh named Bangladesh University of Business and Technology (BUBT).

Md. Syful Islam, received the B.Sc.Engineering degree in computer science and engineering from Patuakhali Science and Technology University in 2012. He is now working as a Lecturer in computer science and engineering at Bangladesh University of Business and Technology. His research interests are Data mining, Search Engine optimization, Biometric system, Semantic Web, Business Intelligence.

Md Moshiur Rahman, has completed his B.Sc Engg. in Computer Science and Engineering from Patuakhali Science and Technology University. Now studying in M.Sc Engg. in Computer Science and Engineering in Islamic University of Technology (IUT), Gazipur, Dhaka and working as an Assistant Hardware Maintenance Engineer at Bangladesh Open University. His research interest is Semantic web, Routing algorithm, High dimensional data mining.