Ranking Web Pages Based On Searching, Keywords and Incoming Links

Syed Muhammad Khalid Jamal 1, Babar Iqbal 2

**1 Department of Computer Science, UBIT, University of Karachi
Karachi, Sind, 75270 , Pakistan**

**2 N.P.I Computer and Emerging Sciences,
Karachi, Sind, Pakistan**

Abstract

In this article, a new Page-Rank strategy is proposed based on the searching keywords and incoming links. Page-Rank is an analysis algorithm that search engine uses to determine pages relevance and importance. For reaching this goal, we extract the meta data from the hypertext documents then we analyze it with the searching keywords and the links that pointing to that hypertext document.

Keywords: *Page-Rank, performance analysis, meta-data, search relevancy*

1. Introduction

Every day, we use different search engines like Google, Yahoo and Bing for searching of our required information. As World Wide Web (WWW) is growing very rapidly and millions of pages are added daily, finding our required information has now become a very difficult task. Over WWW, there are millions of pages about a specific term but most of the time we did not get the relevant result on searching. Google is the most widely used search engine that crawl the pages over the WWW and rank it using the technique of Page-Rank and HITS algorithms. It firstly find related hypertext documents based on the search keyword and secondly count the incoming link pointing to that document and rank that page. However the main drawback is that some of the website use link building software that creates links in bulk on that documents that are not relevant. Therefore, most of the time users get irrelevant result on their searches.

In this article, we propose an algorithm to re-rank the hypertext document so that the user will get the more relevant and informative results on their searching. For this, we check that if the meta data and the content of the document contain the keyword that is being search and the links that are pointing to it also contain that keyword or relevant to it then the document is more relevant and informative for the user.

2. Literature Review

Larry Page and Sergey Brin proposed a link analysis algorithm called Page-Rank that is used to determine the importance of a hypertext document. Over the WWW, the hypertext document that has more incoming link is important as Page-Rank consider every link as a vote and its Page-Rank will be higher. It's also considering the importance of the incoming link so that that has higher Page-Rank will be counted as higher vote.

Page-Rank value can be calculated by considering all the incoming links to a specific page. Let

assume a page *u* and B_U is the set of all pages that point to page *u*, so that the Page-Rank in general case can be defined as:

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

But in many of the cases it's not necessary that every visitor will come from any incoming link or through any other page, it can be visited directly so we add a damping factor *d* which is assumed that it would be around 0.85. This damping factor is subtracted from 1 and then divided by N i.e number of documents in a collection and then this value is added to the above formula, so that it will become:

$$PR(p_i) = \frac{1-d}{N} + d \sum_{p_j \in M(p_i)} \frac{PR(p_j)}{L(p_j)}$$

Web pages that have higher Page-Rank are important because they are more viewed by the users. Therefore it is important that these pages must be fetched and indexed by the crawler for better search results.

3. Methodology

A. Collecting Information

The web crawler has been used to collect all new hypertext documents from the web all the time and save it to database. Every search query will also be saved in a database so that we can collect and categorize keywords from it and assign score to each of them.

B. Preprocessing documents

The keywords from the meta tag from the HTML of the page will be collected. These keywords may contain unnecessary text like preposition, conjunction, punctuations and etc, which are considered as stop word. A list of all these words will be generated so that these words will not consider as a keywords.

C. Keyword Categorization

The categorization of each keyword will be done by simple querying the database. The database contains two types of list of keywords i.e general and specialized. Each keyword of general list will have a score between 0.1 and 0.5 and each keyword in specialize list will have a score between 0.6 and 1. If keyword is not found in both lists then their score will be 0 and will not be considered as keywords.

K.R = Score of the keyword

A data dictionary will be used that will hold all the keywords and phrases to be search and this data dictionary will continuously be updated as per different keywords search by the user that are not already defined in the data dictionary. Keyword score will be assigned after calculating the percentage of the part that is matched in keyword or phrase.

D. Process Incoming Links

All the incoming links that are pointing to a page that is being processed. All keywords from these links will be fetched and compared with the keywords of the page to compare the relevancy between them. If they are not relevant then their keyword relevancy score will be 0 and will not be considered as an incoming link. However if they are relevant, a keyword relevancy score between 0.1 and 1 will be assigned according to the frequency count.

E. Ranking the hypertext document

In the next step we will calculate the importance of a page and assign a Page-Rank according to the keyword score and relevancy score of the links. Let P_i be the page of which we are going to calculate Page-Rank value. First we will gather all the incoming links that has relevancy score greater than 0. Each incoming link Page-Rank will be

divided by number of the link to the page and sum all of them i.e:

$$\sum_{p_j \in M(p_i)} \frac{PR(p_j)}{L(p_j)}$$

Where Pj is the incoming link to page Pi and M(Pi) is the set of all pages that links to Pi and have relevancy score greater than 0. Since there may be some visitor that reaches to the page Pi without any link, therefore we will add a damping factor that has predefined value of 0.85. This damping factor will be minus from 1 and then divided by total number of pages N and again value of damping factor will be added to it, i.e:

$$\frac{1-d}{N} + d$$

After calculating this value it will be multiplied with the sum of PR(u) that has been calculated previously. The result will be the value of Page-Rank that is currently being used but for keyword relevancy we will multiply the score of keyword (K.R) which is calculate previously.

$$PR(p_i) = \left(\frac{1-d}{N} + d \sum_{p_j \in M(p_i)} \frac{PR(p_j)}{L(p_j)} \right) K.R$$

Where K.R is the value of keyword relevancy. The result will be our final Page-Rank value that search engines will use on searching. Pages having higher Page-Rank will have more precedence and will be more relevant to the user query and they have indexed at the top of the search result.

4. Conclusion

The goal of this research is to apply a new Page-Rank technique to retrieve more relevant data from the web on searching. The importance of any hypertext document greatly depends upon the users need i.e how much page is relevant to the user query. Since any hypertext document is enriched with both text and images and currently we are only analyzing the text information. In future some image recognition technique and relevance detection method would be pursued and will try to rank the pages on both the information presented in textual and graphical formats.

5. References

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

[2]. Kale, M, “DYNA-RANK: Efficient Calculation and Updation of Page-Rank”, Computer Science and Information Technology, 2008. ICCSIT '08. International Conference, 2008.

[3]. Fuyong Yuan, “Improvement of Page-Rank for Focused Crawler”, Software Engineering, Artificial Intelligence, Networking, and Parallel/Distributed Computing, 2007. SNPD 2007. Eighth ACIS International Conference, 2007.

[4]. Chia-Chen Yen, “Page-Rank algorithm improvement by page relevance measurement”, Fuzzy Systems, 2009. FUZZ-IEEE 2009. IEEE International Conference, 2009.

[5]. Sunil, A.N.V, “Page-Rank based detection technique for phishing web sites”, Computers & Informatics (ISCI), 2012 IEEE Symposium, 2012.

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

[7]. Wei Huang, “An improved method for the computation of Page-Rank”, Mechatronic Science, Electric Engineering and Computer (MEC), 2011 International Conference, 2011.

[8]. Li Cun-he, “Hyperlink Classification: A New Approach to Improve Page-Rank”, Database and Expert Systems Applications, 2007. DEXA '07. 18th International Workshop, 2007.

[9]. Zhou Cailan, "Improved Page-Rank algorithm based on feedback of user clicks", Computer Science and Service System (CSSS), 2011 International Conference, 2011.

[10]. Xing, W, "Weighted Page-Rank algorithm", Communication Networks and Services Research, 2004. Proceedings. Second Annual Conference, 2004.

[11]. Nora Yahia, Sahar A. Mokhtar and Abdelwahab Ahmed, "Automatic Generation of OWL Ontology from XML Data Source", International Journal of Computer Science Issues (IJCSI), Publication of Volume 9, issue 2, in 2012.

[12]. Kanagaraj.S and Sunitha Abburu, "Converting Relational Database Into Xml Document", International Journal of Computer Science Issues (IJCSI), Publication of Volume 9, issue 2, in 2012.

[13]. E.J.Thomson Fredrick and G. Radhamani , "Fuzzy Integrity Constraints for Native XML Database", International Journal of Computer Science Issues (IJCSI), Publication of Volume 9, issue 2, in 2012.

[14]. Pushpa Suri and Divyesh Sharma , "A Model Mapping Approach for storing XML documents in Relational databases", International Journal of Computer Science Issues (IJCSI), Publication of Volume 9, issue 3, in 2012.

Syed Muhammad Khalid Jamal is an Academic Scholar, Researcher and Faculty Member and is associated as Assistant Professor with Department of Computer Science, Karachi University (the largest and primeval Higher Education institution of Pakistan) since 2001. He is currently author of various research publications and Books at National & international level.

Babar Iqbal is an independent pollster and is associated with N.P.I.C.E.S since for the last 3 years.