ISSN:0975-9646

Impacts of Keyword Frequency and Ranking Scope on Relevance of Web Search Results

Preeti Sharma
Shri Ram College of Engineering & Management
70 KM Stone, Delhi-Mathura Road (NH-2), Palwal
Haryana - 121105, India

***Abstract*- In the past few decades, The World Wide Web has expanded vigorously and so has the information accessible on The Internet. In today's world, it would be impossible to navigate through this gigantic network without the help of a modern Web Search Engine to find the relevant information. Since introduction of PageRank, many variations of it have been proposed which mainly focus on link structure of web and query context. In this work, a generalized PageRank is proposed to incorporate two important factors for quality search results; keyword frequency and ranking scope. Different variations of the proposed ranking algorithm are analyzed to establish importance of proposed factors in influencing relevance of search results with respect to keywords of interest. The experimental results show that the proposed algorithm significantly outperforms PageRank algorithm.**

Keywords: Information retrieval, PageRank, Ranking Algorithm, Keyword Frequency, Web Search

1. INTRODUCTION

The World Wide Web and the Internet have grown tremendously over past few decades. Internet's most visible component, the World Wide Web, is a huge network with billions of pages hosting information on an amazing variety of topics. Using search engines for looking-up information has become an essential part of Internet browsing today. The importance of the link structure of web in webpages' relevance for Web search results is well known. PageRank [1] is one such algorithm which works by assigning higher importance to pages with more number of incoming links from important pages than pages with fewer in-links.

A significant body of research has been dedicated to improving PageRank in terms of relevance of search results by including page content and query context in rank computation process. Bharat and Henzinger [2] exploited content analysis to improve relevance of results of documents retrieval related to a query topic. Chakrabarti et al. [3] proposed differential link weighing to automatically compile resource lists on broad topics. Rafiei and Mendelzon's [4] algorithm biases PageRank computation by preferentially ranking a page on topics for which the page has a high reputation. Richardson and Domingos [5] introduced a Directed Surfer model with the surfer probabilistically jumping from page to page, depending on the content of the pages and the query terms the surfer is looking for. Haveliwala [6] proposed Topic-Sensitive PageRank, which biases page ranks by using a small number of representative basis topics, taken from the Open Directory Project (ODP) [7]. In a conventional search process (a user with a specific information need issues a query to web search engine by providing keywords), we propose to use Keyword Frequency, the number of times a query keyword appears on a page, as an influencing factor in page rank computation, in addition to link structure of the Web.

When computing page ranks corresponding to a query, another influencing factor is whether to include all pages used for building search index or just the pages containing query keywords. According to our literature survey, this aspect has not attracted much research attention apart from Haveliwala's [8] study which proposed to apply PageRank to only the set of pages containing the query terms for faster computation of ranks with less machine resources. We consider this factor along with keyword frequency in our ranking algorithm and analyze its impact on quality of search results.

In this work, we propose a page ranking algorithm which generalizes PageRank to take into account the impact of Keyword Frequency on relevance of results. We consider variations of our algorithm around ranking scope i.e. whether to include all webpages or just the ones containing query terms. A comparison of experimental results of different variations of our algorithm and PageRank is also presented for demonstrating the improvements in search results' quality when the ideas we propose are incorporated in page ranking. The results are obtained using a light-weight search engine that we developed by crawling the web with the Open Directory Project [7] and Wikipedia [9] as seed pages.

2. ALGORITHMS

Our ranking algorithm has two important constituents and the motivation behind each of them can be understood from the following discussion:

Webpage Popularity: In a manner similar to PageRank, the link structure of the web is used to arrive at web page popularities. This procedure can have two variations; compute popularities over set of all web pages included in index or the set of only the pages containing query keywords. In further discussion, we will refer to the first variation of the algorithm as Global or G, because the scope of popularity calculation spans across all web pages of index. The second variation will be referred to as Local or L, since the popularities are computed over only the pages containing the queried keywords.

Keyword Frequency: The second constituent of our ranking algorithm has been introduced to acknowledge the contribution of Keyword Frequency in determining relevance of pages for the query keywords. Introduction of keyword factor in rank calculation is aligned with the intuition that a page would be more relevant for a given keyword if it appears on the page more frequently compared to other pages.

Mathematically, our complete algorithm is represented by equation (3) below:

$$P(0,w) = \frac{1}{N} \quad (1)$$

$$P(s,w) = \frac{1-d}{N} + d \sum_{l \in Inlinks[w]} \frac{P(s-1,l)}{C_{o,l}} \quad (2)$$

$$R(w) = (1-f)P(w) + f\left(\frac{n(k,w)}{n(w)} + \frac{n(k,w)}{n(k,all)}\right) \quad (3)$$

P(s, w) denotes popularity of webpage w at step s during the iterative computation of popularities using equations (1) and (2) until P(s, w) converges for each webpage in the set over which popularities are being evaluated. Here, N is the number of such webpages for which we will consider two variations around popularity scope as discussed earlier. Inlinks[w] is the set of pages from which there are incoming links to page w. In equation (2), the contribution of an incoming link l to popularity of page w is offset by number of outgoing links from l ($C_{o,l}$). The damping factor d (= 0.8) is used to model cases where the random web surfer chooses to start over instead of following links in a sequence. P(w) is the final popularity of page w after convergence is achieved.

The rank of a webpage which is used to sort search results of a query is denoted by R(w). For calculation of webpage ranks, we introduce another factor of keyword frequency as described by the term on the right in equation (3). This idea is aligned with the intuition that for the result webpages to be relevant for a query, the pages should contain query keywords. However, a better criterion would have been that the pages should have semantic relevance to query keywords. But analysis of semantics is complex and requires use of machine learning and natural language processing techniques [10, 11]. Our intention is to study impact of ranking scope and keyword frequency in the context of a simple informational search query. Therefore, it is justified to use the criterion that the pages should contain keywords of interest. But still, it is important to weigh down a keyword's frequency on a page by total number of words on the page and total occurrences of the keyword on all pages. The term on the right in equation (3) summarizes all these ideas mathematically. f signifies importance of keyword frequency term in rank calculation. We have used a value of 0.6 for f to give more importance to contribution of keyword factor over the contribution of webpage popularity in determining relevance of result pages. n(k, w) is the number of times keyword k appears on page w, n(w) is the total number of words on page w and n(k, all) is the total number of occurrences of keyword k on all pages. The justification for using two separate fractions in weighing down n(k, w) is that n(k, w) is typically much smaller compared to n(w) and n(k, all).

Let us point out that in equation (3), if keyword frequency factor is eliminated (i.e. f = 0) and global scope is used (i.e. popularity calculation over all pages in index) then our algorithm boils down to PageRank. We are going to compare results for following 4 variations of our algorithm:

Algorithm 1: Use local ranking scope i.e. compute popularities over the set of only the pages containing the keywords of interest and do not include keyword factor i.e. f = 0. We will refer to this version of our algorithm as **L**.

Algorithm 2: For this version, ranking scope is local like Algorithm 1 but the keyword factor is included (f = 0.6). Let us refer to this version as **LK**.

Algorithm 3: Here, ranking scope is global i.e. popularities are computed over the set of all the pages included in index and keyword factor is not included i.e. f = 0. This variation of our algorithm is essentially PageRank, but for consistency we will refer to it as **G**.

Algorithm 4: In this variation, global ranking scope is used along with keyword factor and it will be referred to as **GK**.

3. RESULTS AND DISCUSSION

We developed a light-weight search engine for obtaining relevance results of search queries. The Open Directory Project and Wikipedia have been used as seed pages for web crawling.

TABLE 1: LIST OF QUERY KEYWORDS

facebook	reddit	youtube
twitter	vimeo	myspace
linkedin	netflix	digg
tumblr	itunes	zuckerberg

TABLE 2: COMPARISON OF RELEVANCE RESULTS

Keyword	Rank of First Relevant Page			
	L	LK	G	GK
facebook	1	1	1	1
twitter	1	1	1	1
linkedin	1	1	2	1
tumblr	1	1	2	1
reddit	1	1	1	1
vimeo	8	1	2	1
netflix	2	2	2	2
itunes	3	1	1	1
youtube	1	1	1	1
myspace	2	2	1	1
digg	1	1	1	1
zuckerberg	1	1	1	1
Mean Rank:	1.917	1.167	1.333	1.083

As hinted earlier, semantic analysis would have unnecessarily complicated this discussion, therefore our search index consists of single word keywords and the words consist of only the symbols from sets {a, b, c, …, z} and {0, 1, 2, …, 9}. As our goal is to compare the relevance of results for the four algorithms defined in section 2, we have not focused on performance aspects of these algorithms in terms of query time or resource requirements.

How do we compare the relevance of results of the four algorithms? First, we need a good sample of keywords to feed to the search engine implementations of the algorithms. We generated search results for the keywords listed in Table 1. These keywords belong to the broad category of "Social Media and Online Media Streaming" and most of these have been collected from among the frequently searched keywords of popular web search engines [12, 13]. Second, a set of criteria is required for assessing relevance of result pages to query keywords. The criteria for qualifying a result page as being relevant to the keyword of interest is defined below:

Relevance Criteria: A webpage w, in list of query results for a keyword k is considered relevant to k either if w describes k such that direct significance of k can be clearly understood from this description or if w has an outgoing link to an official page related to k (which has authoritative information on k).

As an example, consider a webpage, w1 in results of query for the keyword "facebook". If w1 describes "facebook" as "a social networking website hosting user content", then w1 is a relevant page for this keyword. According, to the second part of relevance criteria, if w1 does not describe what "facebook" is, but directs to official facebook website: www.facebook.com, even then w1 is considered relevant.

Using the relevance criteria defined above, ranks of first relevant pages are determined for the keywords listed in Table 1 with all four ranking algorithms and the results are summarized in Table 2. From the mean of ranks of first relevant pages for all keywords (as well as from first relevant page ranks for individual keywords), it is clear that GK offers best quality of results while L is the worst from point of view of relevance of results to query keyword. Collectively, the ranking algorithms with global scope (G and GK) produce better results compared to the ones with local scope (L and LK). However, the local scope algorithm with keyword factor (LK) significantly outperforms the global scope algorithm without keyword factor (G). Furthermore, GK delivers outstanding quality of results compared to G. Thus, these initial experimental results establish that both ranking scope and keyword frequency have important influence on relevance of search results to queried keywords, but the impact of keyword frequency is higher.

For the results described above, we limited web crawling depth to two hops from seed pages. Impacts of variation of crawl depth on relevance results will be studied going forward. Apart from this, the results will be extended to more categories of keywords e.g. Education, Research, Entertainment, Gadgets, Technology, News, Events etc.

4. CONCLUSIONS

A generalized PageRank algorithm is proposed to study impacts of ranking scope and keyword frequency on relevance of search query results. Results are obtained for queries of popular keywords from the variations of proposed algorithm using corresponding web search engine implementations. The results establish importance of ranking scope and keyword frequency in influencing quality of search engine results. It is further established that keyword frequency has greater influence on relevance of results than scope of ranking.

REFERENCES

[1] L. Page, S. Brin, R. Motwani, T. Winograd, The PageRank Citation Ranking: Bringing Order to the Web. Technical Report. Stanford InfoLab, 1999.

[2] K. Bharat, M. R. Henzinger, Improved algorithms for topic distillation in a hyperlinked environment, In Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval (SIGIR), 1998, ACM, New York, NY, USA, 104-111. DOI=10.1145/290941.290972.

[3] S. Chakrabarti, B. Dom, P. Raghavan, S. Rajagopalan, D. Gibson, J. Kleinberg, Automatic resource compilation by analyzing hyperlink structure and associated text, Computer Networks and ISDN Systems, Volume 30, Issues 1–7, April 1998, Pages 65-74, ISSN 0169-7552.

[4] D. Rafiei, A. Mendelzon, "What is this Page Known for? Computing Web Page Reputations", 9th International World Wide Web Conference, Amsterdam, 2000.

[5] M. Richardson, P. Domingos. The Intelligent Surfer: Probabilistic Combination of Link and Content Information in PageRank, volume 14. MIT Press, Cambridge, MA, 2002.

[6] T. H. Haveliwala, "Topic-sensitive PageRank: a context-sensitive ranking algorithm for Web search", Knowledge and Data Engineering, IEEE Transactions on , vol.15, no.4, pp.784,796, July-Aug, 2003, doi: 10.1109/TKDE.2003.1208999.

[7] The Open Directory Project: Web directory for over 4 million URLs, Retrieved 15 May, 2014, http://www.dmoz.org/.

[8] T. H. Haveliwala, Efficient computation of PageRank, Stanford University Technical Report, 1999.

[9] Wikipedia: The Free Encyclopedia, Wikimedia Foundation Inc., Retrieved 15 May, 2014, http://www.wikipedia.org/

[10] S. Lawrence, Context in Web Search, IEEE Data Engineering Bulletin, Vol. 23 (2000), pp. 25-32.

[11] L. Finkelstein, E. Gabrilovich, Y. Matias, E. Rivlin, Z. Solan, G. Wolfman, E. Ruppin, Placing search in context: the concept revisited, In Proceedings of the 10th international conference on World Wide Web, 2001, ACM, New York, NY, USA, 406-414, DOI=10.1145/371920.372094.

[12] Google Trends, Retrieved 15 May, 2014, http://www.google.com/trends

[13] Top Bing Searches, Retrieved 15 May, 2014, http://www.bing.com/blogs/site_blogs/b/search/archive/2013/12/01/eoy.aspx