Query-Optimized PageRank: A Novel Approach

Rajendra Kumar Roul, Jajati Keshari Sahoo and Kushagr Arora

Abstract This paper addresses a ranking model which uses the content of the documents along with their link structures to obtain an efficient ranking scheme. The proposed model combines the advantages of *TF-IDF* and PageRank algorithm. *TF-IDF* is a term-weighting scheme that is widely used to evaluate the importance of a term in a document by converting textual representation of information into a vector space model. The PageRank algorithm uses hyperlink (links between documents) to determine the importance of a Web document in the corpus. Combining the relevance of documents with their PageRanks will refine the retrieval results. The idea is to update the link structure based on the document similarity score with the user query. Results obtained from the experiment indicate that the performance of the proposed ranking technique is promising and thus can be considered as a new direction in ranking the documents.

Keywords Cosine-similarity · Pagerank · Query-optimized · Spearman's footrule · TF-IDF

1 Introduction

The dynamic Web which contains a huge volume of digital documents is growing in an exponential manner due to the popularity of the Internet. This makes difficult for the search engine to arrange the retrieved results from the Web according to the

R. K. Roul (✉) · K. Arora
Department of Computer Science, BITS-Pilani, K. K. Birla Goa Campus,
Zuarinagar 403726, Goa, India
e-mail: rkroul@goa.bits-pilani.ac.in

K. Arora
e-mail: kushagrarora786@gmail.com

J. K. Sahoo
Department of Mathematics, BITS-Pilani, K. K. Birla Goa Campus,
Zuarinagar 403726, Goa, India
e-mail: jksahoo@goa.bits-pilani.ac.in

H. S. Behera et al. (eds.), *Computational Intelligence in Data Mining*,
Advances in Intelligent Systems and Computing 711,
https://doi.org/10.1007/978-981-10-8055-5_60

user's choice. *Ranking* which is one of the powerful technique in machine learning can shed light in this direction by arranging the retrieved results in a better manner. Many research works have been done in this field [1–7].

The existing PageRank algorithm [8, 9] has many short comings. The current paper tries to improve the existing PageRank algorithm. The problem with the traditional PageRank algorithm is that the link structure used in the calculation of PageRank assumes that a surfer will jump from a page to the other uniformly which may lead to topic drifting, i.e., suppose a user is looking for the documents on *computer science* and some documents may have outgoing links to *biological documents* (because many biological documents also related to computer science) then those documents also incorporate in the PageRank calculation. The proposed approach overcomes this by biasing the next jump of the user to only those documents which are relevant to the particular query he is searching for. The modified link structure which is input to the PageRank algorithm will contain those output links which are connected to a document having some standard relevance (which in our case is nonzero cosine-similarity with the user query). When these resulted documents from a normal ranking function (cosine-similarity in the proposed approach) with the updated link structure are supplied to the existing PageRank algorithm, the new ranks are obtained. These new ranks are query dependent and give a new direction to the existing PageRank algorithm that incorporates the needs of the user as well. Modifying the existing PageRank algorithm and considering it for restructuring the links based on query would be more beneficial while implementing it in search engine with datasets which have good link structure such as research journal databases that have a lot of citations and may not be linked to the topics as well. Hence, if their structure is modified, one may be able to get better results. Empirical results on different datasets show the effectiveness of the proposed approach.

The rest of the paper is organized on the following lines: Sect. 2 discusses the background of those techniques which are used in the proposed approach. In Sect. 3, we have discussed the proposed technique used for query-optimized PageRank. Section 4 discusses the experimental results of the proposed work. Finally, in Sect. 5, we concluded the work with some future enhancement.

2 Basic Preliminaries

2.1 Vector Space Model

An algebraic model called Vector Space Model (VSM) [10] aimed to facilitate information retrieval by modeling the documents as a set of terms. VSM transforms a full text version to a vector which has various patterns of occurrence. It represents document (D) as a vector of words, $D = (w_{1j}, w_{2j}, w_{3j}, w_{4j}, \ldots, w_{nj})$, where w_{ij} is weight of ith word in jth document.

2.2 Term Frequency and Inverse Document Frequency

Term Frequency (*TF*) measures how often a term *t* occurs in a document *D* whereas inverse document frequency (*IDF*) measures the importance of *t* in the entire corpus *P*. *TF-IDF* [11] is a technique which finds the importance of terms in a document based on how they appear in the corpus. If *t* appears in many documents, its importance goes down. Therefore, the common terms need to be filtered out. *TF-IDF* is calculated using Eq. 1.

$$TF\text{-}IDF_{t,D} = TF_{t,D} \times IDF_t \tag{1}$$

where

$$TF_{t,D} = \frac{number\ of\ occurance\ of\ t\ in\ D}{total\ length\ of\ D}$$

$$IDF_t = log\left(\frac{number\ of\ documents\ in\ P}{number\ of\ documents\ contain\ the\ term\ t}\right)$$

2.3 Cosine-Similarity

Cosine-similarity[1] is one among the standard similarity techniques which measures the similarity between two document vectors, say $\overrightarrow{D_1}$ and $\overrightarrow{D_2}$ and can be represented using Eq. 2:

$$cosine\text{-}similarity(\overrightarrow{D_1}, \overrightarrow{D_2}) = \frac{\overrightarrow{D_1} \cdot \overrightarrow{D_2}}{\overrightarrow{|D_1|} * \overrightarrow{|D_2|}} \tag{2}$$

If the angle between $\overrightarrow{D_1}$ and $\overrightarrow{D_2}$ is near to zero, then they share most of the common terms between them, and if the angle approaches toward 90°, then the dissimilarity between them increases. To retrieve the document relevant to a query, it is necessary to have a measure that computes the degree of similarity between the query and each of the document in the corpus. As in the vector space model we deal with vectors, the cosine-similarity between the vectors is used to obtain the similarity measure.

[1] https://radimrehurek.com/gensim/tutorial.html.

The cosine-similarity between the document $\overrightarrow{D_1}$ and the query $\vec{Q}$ is represented in Eq. 3.

$$cos\text{-}sim(\overrightarrow{D_1}, \vec{Q}) = \frac{\overrightarrow{D_1} \cdot \vec{Q}}{|\overrightarrow{D_1}| * |\overrightarrow{Q}|} \tag{3}$$

3 Proposed Approach

This section discussed the query-optimized PageRank algorithm starting with a detail description of the PagaRank algorithm. The PageRank of a document combined with its *TF-IDF* vector will generate the desire ranking. Details are discussed in Sect. 3.2.

3.1 PageRank Algorithm

PageRank determines how important the Web site is by counting the number and quality of links to a page where the links are dynamic. It analyzes each link and thus assigns a weight to each element of a hyperlink set of documents. It basically measures the importance of a page compared to all other pages on the Web. A link to a page leads to an increase in the weight, while the page with no link can be neglected as it is less significant. The following steps are used to compute the PageRank:

1. Add an edge directed from node i to node j in the graph when Web site i references j. Since only the connections between the Web sites are important, any navigational links such as next and back buttons are ignored while computing their PageRanks.
2. In the proposed approach, all the pages will get equal importance that is linked by a single page. Thus, the importance of each page will be $\frac{1}{n}$, iff a node has n outgoing edges. Let us denote the transition matrix of the graph G by A and represented as

$$A = \begin{bmatrix} x_{11} & x_{12} & x_{13} & \dots & x_{1n} \\ x_{21} & x_{22} & x_{23} & \dots & x_{2n} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ x_{d1} & x_{d2} & x_{d3} & \dots & x_{dn} \end{bmatrix}$$

3. Supposing that initially the importance is uniformly distributed among all the nodes then the initial rank vector v will have all the entries as $\frac{1}{k}$, if G has k nodes. Since the importance of a Web page is increased by each incoming link using step 1, the rank of each page will be updated by adding the importance of the incoming links to the current value. It is equivalent to multiplying A with v. Thus, after first iteration, the new importance vector is $v_1 = Av$. We can keep

iterating the above process to get the sequence $v, Av, A^2v, ..., A^kv$. This is called the PageRank of the Web graph G.

4. Since the dataset is so large, the graph G is not expected to be connected. Likewise, our dataset may have plain descriptive pages that do not have any outgoing links. Thus, for any directed Web graph, one require a non-ambiguous meaning of the rank of a page. To overcome such problems, damping factor (p) is used which is a positive constant ranging between 0 and 1. The typical value of damping factor is 0.15. The PageRank of G is defined in Eq. 4.

$$\text{PageRank(G)} = \text{(1-p)A} + \text{pB} \tag{4}$$

where

$$B = \frac{1}{n}\begin{bmatrix} 1 & 1 & 1 & \dots & 1 \\ 1 & 1 & 1 & \dots & 1 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & 1 & 1 & \dots & 1 \end{bmatrix}$$

3.2 *Query-Optimized PageRank*

It is a general thought to develop a technique which can combine the PageRank of a Web page (i.e., document) with its *TF-IDF* vector to get advantages of both these techniques. In this paper, a framework is drawn by using the content of Web page and the outlink information synchronously though primarily we have modified the popular "PageRank" algorithm. For a given user query Q, rank and return the documents which have nonzero cosine-similarity score with the query Q. In those return documents, we modify the link structure matrix to compute the new ranks. All the links which have a zero cosine-similarity with the query are removed as an outlink; therefore, for every page, one gets a smaller link structure. This idea basically emphasizes that a surfer who has started searching for a query will only jump to those pages which are related to the query in a certain way. Once the link structure is modified, adjustments are made by including the damping factor and then the new ranks are calculated using the steps mentioned below:

1. *Preprocessing of documents*:
   Consider a corpus P having C number of classes (where $C = \{C_1, C_2, ..., C_n\}$) and each class has m number of documents $D = \{D_1, D_2, ..., D_m\}$. Documents are preprocessed by removing the stop-words, and stemming is done for the entire corpus using porter stemming algorithm[2]. The stemmed vocabulary forms a dictionary of all the features (i.e., terms). Assume that the corpus dimension is $p \times r$ (p is the number of documents and r is the number of terms of the corpus P).

[2] http://tartarus.org/martin/PorterStemmer/.

Table 1 Term-document table

	D_1	D_2	D_3	...	D_p
t_1	t_{11}	t_{12}	t_{13}	...	t_{1p}
t_2	t_{21}	t_{22}	t_{23}	...	t_{2p}
t_3	t_{31}	t_{32}	t_{33}	...	t_{3p}
.	.	.	.	...	.
.	.	.	.	...	.
.	.	.	.	...	.
t_r	t_{r1}	t_{r2}	t_{r3}	...	t_{rp}

2. *Finding TF-IDF vectors*:
   From all the preprocessed documents and terms, the text data are later represented in numerical values. The formal VSM has been chosen for this purpose which converts every document into vector using *TF-IDF* values called *document-term* vectors (i.e., documents in the term space) as shown in Table 1.

3. *Finding cosine-similarity with the query vector:*
   Similar to the documents, the *TF-IDF* vector of the query is obtained. The cosine-similarity is calculated between the query vector and each document vector. The documents are arranged in the decreasing order of their cosine-similarity values, and all the documents that have zero cosine-similarity are removed from the corpus.

4. *Computing the ranks of the documents:*
   The ranks of the documents are calculated using the PageRank algorithm on the remaining documents in the corpus. Initially, same importance is given to all the documents. After that, the rank is updated by adding the importance of the incoming links to the current value of the document. The process is repeated for all the documents. The ranks are returned after incorporating the damping factor to the obtained rank matrix.

4 Experimental Work

To implement the PageRank algorithm with *TF-IDF*, a specific research dataset has been chosen called dbpedia[3] which includes both link structure and content. This dataset does not have a predefined set of relevant documents for any given query, so the measures such as accuracy, precision, recall, and F-measure could not be com-

[3] http://wiki.dbpedia.org/Datasets.

puted directly. This led to use a method called Spearman's footrule [12] to check accuracy of our ranking approach as compared to standard cosine-similarity ranks.

Spearman's footrule is applied to both the techniques (query-optimized and cosine-similarity) for obtaining the rankings. Assuming the size of the dataset to be N, this in turn implies that the rankings should range between 1 and N. As the rankings given by each of the two techniques being compared is basically a permutation of the other, hence there are no ties allowed. Let us say that the result of the rankings is permutations σ_1 for the proposed approach and σ_2 for the ranking based on the cosine-similarity score. This permutation is over S, the set of overlapping results when the top 'k' rankings of each model are considered. Spearman's footrule is computed using Eq. 5.

$$Fr^{|S|}(\sigma_1, \sigma_2) = \sum_{i=1}^{|S|} |(\sigma_1(i) - \sigma_2(i)| \tag{5}$$

$Fr^{|S|}$ is zero when the two lists are identical. When $|S|$ is even, $\frac{1}{2}S^2$ achieved its maximum value of $\frac{1}{2}S^2$ and when $|S|$ is odd, it achieved the maximum value of $\frac{1}{2}(|S|+1)(|S|-1)$. $Fr^{|S|}$ value will lie between 0 and 1, and this can be achieved by dividing the obtained result with its maximum value which is independent on the size of the overlap. The normalized Spearman's footrule (NFr) for $|S| > 1$ is computed using the Eq. 6.

$$NFr = \frac{Fr^{|S|}}{max\ Fr^{|S|}} \tag{6}$$

Thus, NFr will range between 0 and 1. Tables 2, 3, 4, 5, 6, and 7 show the rankings for different queries (both unigram and bi-gram) using the cosine-similarity, and the proposed query-optimized PageRank algorithm (NFr shows the difference between the NFr value obtained on cosine-similarity and NFr value obtained on the proposed approach). Only one query "Massachusetts" (Table 4) had three results, and Spearman coefficient turned out to be 0 for that since the retrieved documents are very less

Table 2 Query: agriculture (NFr = 0.333)

Cosine-similarity ranking	Query-optimized PageRanking
Agriculture	Agriculture
Algeria	Africa
Albania	Algeria
Almond	African_union
Accountancy	Albania
Africa	Almond
2005_Atlantichurricane_season	2005_Atlantichurricane_season
Aberdeen	Aberdeen

Table 3 Query: wikipedia (NFr = 0.821)

Cosine-similarity ranking	Query-optimized PageRanking
2005_Lake_TAnganyika_earthquake	African_Great_Lakes
African_Darter	2005_Lake_TAnganyika_earthquake
African_Black_Oystercatcher	African_Grey_Hornbill
African_Jacana	Abstract_art
African_dwarf_frog	06-02-2006
African_Great_Lakes	African_Brush_tailed_Porcupine
African_Grey_Hornbill	02-08-2000
Abstract_art	16_Cygni_Bb
06-02-2006	African_Buffalo
African_Brush_tailed_Porcupine	Almaty
02-08-2000	African_Darter
16_Cygni_Bb	African_Black_Oystercatcher
African_Buffalo	African_Jacana
Almaty	African_dwarf_frog

Table 4 Query: massachusetts (NFr = 0)

Cosine-similarity ranking	Query-optimized PageRanking
Abu_dhabi	Abu_dhabi
2004_Atlantichurricane_season	2004_Atlantichurricane_season
Alternative_rock	Alternative_rock

Table 5 Query: nobel (NFr = 0.635)

Cosine-similarity ranking	Query-optimized PageRanking
Alfred_Nobel	Albert_Einstein
Albert_Einstein	Alan_Turing
ABO_blood_group_system	Aberdeen
Alan_Turing	Alfred_Nobel
Action_potential	ABO_blood_group_system
Arican_Amrican_literature	Action_potential
Aberdeen	Arican_Amrican_literature

Table 6 Query: Roman Empire (NFr = 0.661)

Cosine-similarity ranking	Query-optimized PageRanking
1st_century_BC	14th_century
13th_century	13th_century
10th_century	Abacus
5th_century	9th_century
3rd_century	11th_century
11th_century	10th_century
6th_century	5th_century
Abacus	6th_century
14th_century	1st_century
1st_century	Akkadian_Empire
Akkadian_Empire	16th_century
9th_century	Aachen
16th_century	1st_century_BC
Aachen	3rd_century

Table 7 Query: general history (NFr = 0.714)

Cosine-similarity ranking	Query-optimized PageRanking
9th_century.txt	12th_century
1st_century_BC.txt	11th_century
12th_century.txt	9th_century
6th_century.txt	13th_century
African_slave_trade.txt	10th_century
Acceleration.txt	20th_century
13th_century.txt	4th_century
Adriaen_van_der_Donck.txt	1st_century_BC
Alfred_the_Great.txt	6th_century
10th_century.txt	African_slave_trade
Acts_of_Union_1707.txt	Acceleration
20th_century.txt	Adriaen_van_der_Donck
4th_century.txt	Alfred_the_Grea
11th_century.txt	Acts_of_Union_1707

and the link structure could not refine the ranks much based on the cosine-similarity of the documents with the query. A nonzero spearman's score says that the ranking given by the proposed approach that incorporates the cosine-similarity with PageRank is different, and puts forward a new direction of research for a modified PageRank which is query dependent.

5 Conclusion

This paper developed a ranking model that utilized the link structure and the cosine-similarity of the documents with the query, so as to improve the set of retrieved pages and the ranking. It has brought together the merits of ranking using *TF-IDF* weights and the PageRank algorithm. Unlike the *TF-IDF* weighting scheme, it does not just concentrate on the content of the documents nor like the PageRank algorithm rely solely on the link structure rather it ranks based on the relevance of the documents and their outlinks to the query. Our approach puts forward new unexplored ideas on PageRank algorithm which take cares the user's preferences. This work can be further extended as follows:

- To compare the ranking model of the proposed approach with other standard and recognized ranking models to understand the true efficiency of our approach.
- Modifications on the PageRanking based on query and subsetting the link structure accordingly may improve existing ranking systems such as on research journal databases, Wikipedia database (only on those links which are in Wikipedia), etc.
- Further, each query may be viewed as a mixture of various topics where each document is considered to have a set of topics that are assigned to it via. latent dirichlet allocation. This way PageRank matrix will receive more relevant and important outlinks.

References

1. A. J. Roa-Valverde and M.-A. Sicilia, "A survey of approaches for ranking on the web of data," *Information Retrieval*, vol. 17, no. 4, pp. 295–325, 2014.
2. H. Wu, Y. Hu, H. Li, and E. Chen, "A new approach to query segmentation for relevance ranking in web search," *Information Retrieval Journal*, vol. 18, no. 1, pp. 26–50, 2015.
3. J. B. Vuurens and A. P. de Vries, "Distance matters! cumulative proximity expansions for ranking documents," *Information Retrieval*, vol. 17, no. 4, pp. 380–406, 2014.
4. S. Gugnani and R. K. Roul, "Article: Triple indexing: An efficient technique for fast phrase query evaluation," *International Journal of Computer Applications*, vol. 87, no. 13, pp. 9–13, 2014.
5. Y. Wang, J. Lu, J. Chen, and Y. Li, "Crawling ranked deep web data sources," *World Wide Web*, vol. 20, no. 1, pp. 89–110, 2017.
6. P. Chahal, M. Singh, and S. Kumar, "An efficient web page ranking for semantic web," *Journal of The Institution of Engineers (India): Series B*, vol. 95, no. 1, pp. 15–21, 2014.

7. R. K. Roul and S. Sanjay, "Cluster labelling using chi-square-based keyword ranking and mutual information score: a hybrid approach," *International Journal of Intelligent Systems Design and Computing*, vol. 1, no. 2, pp. 145–167, 2017.
8. L. Page, S. Brin, R. Motwani, and T. Winograd, "The pagerank citation ranking: Bringing order to the web." Stanford InfoLab, Tech. Rep., 1999.
9. S. Gugnani, T. Bihany, and R. K. Roul, "A complete survey on web document ranking," *International Journal of Computer Applications*, vol. ICACEA, no. 2, pp. 1–7, 2014.
10. G. Salton, A. Wong, and C.-S. Yang, "A vector space model for automatic indexing," *Communications of the ACM*, vol. 18, no. 11, pp. 613–620, 1975.
11. K. Sparck Jones, "A statistical interpretation of term specificity and its application in retrieval," *Journal of documentation*, vol. 28, no. 1, pp. 11–21, 1972.
12. P. Diaconis and R. L. Graham, "Spearman's footrule as a measure of disarray," *Journal of the Royal Statistical Society. Series B (Methodological)*, pp. 262–268, 1977.