Webology, Volume 16, Number 2, December, 2019

Improvement of Page Ranking Algorithm by Negative Score of Spam Pages

Ateye Zeraatkar

Department of Computer Engineering, Bahonar University, Kerman, Iran. ORCID: 0000-0003-1049-1651. E-mail: a.zeraatkar88@gmail.com

Hamid Mirvaziri

Department of Computer Engineering, Bahonar University, Kerman, Iran. ORCID: 0000-0003-2863-3750. E-mail: hmirvaziri@uk.ac.ir

Mostafa Ghazizadeh Ahsaee

Department of Computer Engineering, Bahonar University, Kerman, Iran. ORCID: 0000-0002-3482-7642. E-mail: mghazizadeh@uk.ac.ir

Received June 12, 2019; Accepted December 20, 2019

Abstract

There are billions of web pages in the web and the most significant point is how to search these pages regarding their usefulness. Usually a user enters a query into a search engine and looking for the best responses. Search engines work based on ranking algorithms. Ranking algorithms employ web mining methods. Web mining is divided into structure, content and web usage mining. In search engines, related pages with user's query must be listed as a result and to obtain a better results, content mining algorithms may be included. Features such as input and output weight, content weight, spam score, length of URL, the number of related pages and href tags values are considered with a negative score. Proposed algorithms are implemented and compared with well-known algorithm. In this algorithm, Spam score feature is used which is combined with the age of domain and content weight of pages. In PRS (PageRank with Spam score) and PRST (PageRank with Spam score and Time factor) algorithms a better response is achieved than PR (PageRank) algorithm. Obtained results of the other algorithm, WPCRST (Weighted Page Content Rank with Spam score and Time factor) indicates that all measures are improved comparing with PR algorithm and provide better responses. First proposed algorithm combines

extracted spam features from “moz.com” with structural features of Web mining. In the second proposed algorithm age of pages is used in addition to the neighborhood matrix and spam score. This feature is used to fix the rich getting richer. Logically this algorithm offers better scores than the previous algorithms. Third proposed algorithm uses weighting methods. In fact, in this algorithm, weight of inbound links and weight of outbound links and content weight of pages are used simultaneously. This algorithm includes web content mining by using content weight of pages. Values of TFIDF and BM25 algorithms are used to obtain the content weight. Obtained results specify that this algorithm has a better ranking than PRST algorithm except for the measure of precision. About 41 percent improvement can be seen in the measure of NDCG and 48 percent improvement for precision while AP has improved 0.8 percent and finally 2.5 percent improvement can be seen in Mean NDCG. In this algorithm, value of TFIDF algorithm is considered as the content weight.

Keywords

PageRank; Content weight; Spam score; Age of domain

Introduction

In the past few years the web space is growing dramatically. In addition to these enhancements, new search engines are made. Now there are more than 52 search engines, like Google, Yahoo, Bing, Salam, Parsijoo and the others (Lewandowski, 2015). Cyberspace contains about 11 specialized search engines and 13 directories which may not satisfy millions of users so there may be necessities to some worthy search engines. Web mining algorithms are developed and new methods are provided in recent years. In addition, the number and size of documents stored on web sites are growing and web sites have to be updated and information should be uploaded on them every day. Therefore the main problem of web space is dedicated search engines. These huge available resources are only understandable and usable for humans in terms of understanding the meaning. Because machines have some limitations to share and integrate vast quantities of information in a search action and Internet search engines should be able to decide about the usefulness of a page like human.

Web mining

Web mining is applying data mining and statistical and artificial intelligence techniques to discover and extract useful and structural information from documents and web services. Web mining works on the Web such as web page content, users' access information and hyperlinks through pages. More ever, the Web is not only consists of a huge collection of information, but also includes dynamic collection of links to access web pages which provides a rich dataset for data mining. Web mining includes the following steps (Kumar, 2016):

- Finding the source: This step involves retrieving of desired web documents.
- Selection and preprocessing of information: Specific information of documents automatically are retrieved, selected and preprocessed
- Generalization: General patterns are discovered automatically in one or several web sites.
- Analyzing: Obtained patterns from previous stage will be validated and interpreted.

Current search engines Problems are as follows:

- Returning a lot of responses and their responses have low quality.
- Hidden data is not usually searched.
- Solving the search engine problems requires a good page ranking algorithm which is the main discussion in current article (Radfar, 2010).

1. Web Content Mining

Web content mining is exploiting of useful information in web content, data and documents. Web contains a lot of information in the form of HTML, news, magazines, e-books, text, image, audio and video. Working with these documents and extracting information from them, is the task of web content mining.

Web content mining may have looks over two aspects: information retrieval and database points of view. Information retrieval deals with facilitating or improving the search or filtering the information presented to the users while the target of the database is presenting a model and integrating of web data so that complex queries can be concluded from keyword based queries (Ein Abadi, 2001).

One of the problems in web content mining is ranking spammers. Spammers effect on the content and structural aspect of pages. One type of spammers, try to increase their pages similarity with queries on that area by changing the content of the documents and add keywords into hidden places while another type try to increase their ranking by pointing of many links from other sites to their own site (Camastra, 2015).

2. Web structure mining

Web has a heterogeneous structure where documents are connected to each other and form a huge graph. Nodes of this graph are web pages and hyperlinks constructs edges of the graph (Aggarwal & Zhao 2013). Web links have valuable information, therefore new ranking algorithms were proposed based on them. This method can be used to categorize web pages or to generate information such as the web sites' similarity.

Web links have valuable information, therefore new ranking algorithms were proposed based on them. These methods can be used to categorize web pages or to generate information such as web sites' similarities. These methods can solve the problem of ranking of spammers' content

methods and helps to achieve more accurate responses. If a query is related to an image, audio and video files, the structural mining is helpful because the connections of a web page have more detailed description of its content which is more suitable than content mining.

Algorithms based on web structure mining are divided into query dependent and query independent. In query independent methods, rankings is performed offline by using the web graph; thus the page's rank for each query is constant, while query dependent ranking method in graph consists of a set of pages related to the user's query.

The rich getting richer is the problem of this method. Popular pages are always on the top of the user's recommended list, causes the users to observe only certain pages so newly born high quality pages could not be in front of them. This problem will be continued for popular pages to become more and more popular and the number of their pointed links increases.

Previous web PageRank algorithms

In this section, more than 10 different famous PageRank algorithms in the literature are introduced. Advantageous and disadvantageous of the mentioned algorithms are also reported.

1. PageRank

This algorithm (PR) is presented in (Brin & Page, 1998) and it is an independent query method. Ranking of each page is obtained by assigning weights to the link of the given page. Its weight depends on the quality of the link which is located in. The more important the page links, the higher weight will be assigned.

$$r(i)=c*\sum\_(j\in B(i))\,r(j)/N(j) \qquad (1)$$

Where B(i) is the set of input pages to page i, N(i) is the set of output pages from page j and r(i), r(j) are the ranks of pages i and j, respectively. C is a constant that is normally set to 1.

Since the web is not absolutely connected it has the problems of dropout and web holes (a fully connected cluster of the internal web graph that has no connection to the outside of the cluster is referred as dropout). The first step is clearing web holes by deleting all nodes with the zero out-degree. In continue, a damping factor d is employed to solve the dropout problem.

$$r(i)=((1-d)/n)+d*\sum\_(j\in B(i))\,r(j)/N(j) \qquad (2)$$

Where n is the number of all web pages.

Here a page might get a higher rank if a large number of pages are pointing to it or the pages pointing have a high rank (Brin & Page, 1998).

2. Weighted PageRank

This algorithm (WPR) is actually an extended version of the PR algorithm. In this algorithm (WPR), importance of a page depends on both outbound links and inbound links.

$$WPR(i) = (1 - d) + d \sum_{j \in B(i)} WPR(j) W^{in}_{(j,i)} \; W^{out}_{(j,i)} \qquad (3)$$

$W^{in}_{(j,i)}$ ،$W^{out}_{(j,i)}$ parameters are input link weight and output link weight, respectively (Xing & Ghorbani, 2004).

2.1. Input Link Weight

In the weight of the link (j, i) is calculated based on the number of incoming links to the page i, relative to the number of incoming links to all pages related to the page j as the following formula.

$$W^{in}_{(j,i)} = \frac{Ii}{\sum_{p \in B(j)} Ip} \qquad (4)$$

where Ii and Ip are the number of incoming links to the page i and p, respectively.

2.2. Output Link Weight

Weight of link (j, i) is calculated based on the number of outgoing links to page i, relative to the number of outgoing links to all pages related to page j within the following formula.

$$W^{out}_{(j,i)} = \frac{Oi}{\sum_{p \in B(j)} Op} \qquad (5)$$

Where O_i and O_p are the number of outbound links to the page i and p, respectively.

WPR algorithm ranks only based on the links. The pages which get the top position of the rank list may be irrelevant to the user's query. To solve the problem, four types of pages are defined:

Very Relevant pages (VR), which contain very important information about the given query, Relevant pages (R), which have relevant but not important information about the given query, Weak-Relevant pages (WR), which do not have relevant information about the given query even though they contain the keywords of the given query, and Irrelevant pages (IR), which not only include the keywords of the given query but also does not have relevant information about it.

The values of VR, R, WR and IR are 1, 0.5, 0.1; and 0, respectively which will be added to the values obtained by the PR and WPR and leads to reach a better result (Kadry & Kalakech 2013).

3. TFIDF

TFIDF algorithm is introduced by Salton (1971). This algorithm is employed statistical document properties and queries to determine the weight of used words in that document as follows:

$$D = TF(w,d).IDF(w) \quad (6)$$

$$IDF(w_i) = \log(\frac{|D|}{DF(w_i)}) \quad (7)$$

TF (w,d): The repeated number of word w in document d. If a word appears in a document several times, that word describes better that document. It is usually normalized with the length of document and the word frequency which is divided by the highest frequencies.

DF(w): The number of documents that the word w exists in them.

|D|: The total number of all documents

IDF indicates the inverse of repetition of a word in all documents. Therefore those words appears in fewer documents has more information than words appears in more documents (like prepositions).

4. BM25

This algorithm is introduced by as an unsupervised algorithm [Robertson & Zaragoza 2009]. It is based on the probability of similarity between query Q and document D.

This algorithm is one of the best ranking methods and has a high accuracy. Total weight of all words in the query Q is obtained from the following formula.

$$S(D,Q) = \sum_{i=1}^{n} \frac{IDF(q_i)\big(f(q_i,D)(k_1+1)\big)}{f(q_i,D) + k_1\left(1 - b + b\left(\frac{|D|}{avgdl}\right)\right)} \quad (8)$$

$$IDF(q_i) = \log\frac{(N - n(q_i) + 0.5)}{(n(q_i) + 0.5} \quad (9)$$

Where $k_1 = 1.2$ and b=0.75.

Smaller values of b often lead to have a better ranking. N is the number of documents and n is the number of documents that involve qi. K3 is equal to 7 or 1000 for big length queries and the value of K2 is often zero (Robertson & Zaragoza, 2009).

The proposed algorithms

1. PageRank with Spam score

This algorithm combines extracted spam features from "moz.com" with structural features of Web mining. PRS is calculated by the following formula.

$$PRS(i) = \left(\frac{1-d}{n}\right) + d * \sum_{j \in B(i)} \frac{PRS(j)}{N(j)} * S(j) \qquad (10)$$

In the above formula S is Spam score. D is damping factor and n is the total number of pages.

2. PageRank with Spam score and Time factor

In this algorithm age of pages is used in addition to the neighborhood matrix and spam score. This feature is used to fix the rich getting richer. Logically this algorithm offers better scores than the previous algorithms (Bhardwaj et al., 2015).

$$PRST(\mathrm{i}) = T(i)\left[\left(\frac{1-\mathrm{d}}{\mathrm{n}}\right) + \mathrm{d} * \sum_{\mathrm{j} \in \mathrm{B(i)}} \frac{\mathrm{PRST(j)}}{\mathrm{N(j)}} * S(j)\right] \qquad (11)$$

In the above formula T is age of a page. Age of domain is calculated based on the number of days in the implementation. It is possible to use normalized method such as logarithmic mean and etc to have a better results.

3. Weighted Page Content Rank with Spam score and Time factor

This algorithm uses weighting methods. In fact, in this algorithm, weight of inbound links and weight of outbound links and content weight of pages are used simultaneously. This algorithm includes web content mining by using content weight of pages. Values of TFIDF and BM25 algorithms are used to obtain the content weight. BM25 algorithm gives a better ranking so the average of these two methods can be used for the content weight.

$$WPRST(\mathrm{i}) = T(i)[\left(\frac{1-\mathrm{d}}{\mathrm{n}}\right) + \mathrm{d} * Cw * \sum_{\mathrm{j} \in \mathrm{B(i)}} \frac{\mathrm{WPRST(j)}}{\mathrm{N(j)}} * W_{in}(j,i) * W_{out}(j,i) * S(j)] \qquad (12)$$

In the above formula Cw is content weight. But the Win and Wout are inbound weight of links and outbound weight of links. These values are achieved by formulas 9 and 10 in the previous section.

Dataset

We need a dataset which have the relationship between links (Bhardwaj et al., 2015). Therefore, a dataset named DOTIR is selected (Doroudi et al., 2008). This dataset is the collection of Iranian sites and have .ir extension. It has the feature of the relevant pages with queries that is employed during algorithm evaluation. This dataset has examined collection of a million of

pages and provides XML format files Including tags, body, docid, links, anchor text title, URL and HTML. These tags are used to obtain the desired data. To achieve the value of age of pages, seomastering.com web site and to get spam score MOZ.com web site is used.

Compare the purposed algorithms with PageRank algorithm

In this section, obtained results are evaluated and compared with the results of other methods in the previous section. The results are shown in the form of diagrams of implementation of the algorithm in MATLAB.

1. Comparison between PR and PRS algorithms

Four criteria are used to compare these algorithms including: precision, NDCG, AP and average of NDCG.

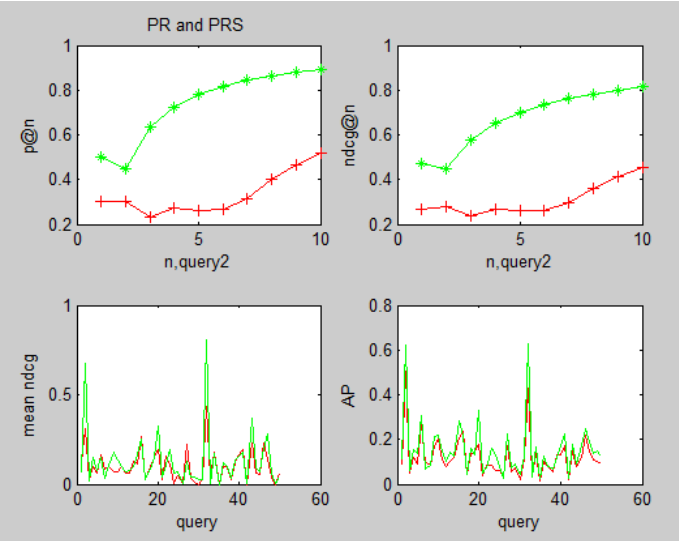

Figure 1. Comparing PR with PRS algorithm in precision, NDCG, AP and mean NDCG

Measure of NDCG in query 2 is improved about 36 percent, precision is improved 40 percent, AP is improved about 3 percent and finally Mean NDCG has 2.75 percent improvement. Obtained results are shown in Table 1 and Figure 1.

Table 1. Mean values of evaluation criteria for PR and PRS

Test	PR	PRS	Difference
Mean NDCG	0.1019	0.1294	0.0275
Precision in query 2	0.3336	0.7378	0.4042
NDCG in query 2	0.3103	0.6739	0.3636
AP	0.1229	0.1538	0.0309

2. Comparison between PR algorithm and PRST algorithm

This algorithm uses the age of pages and spam score.

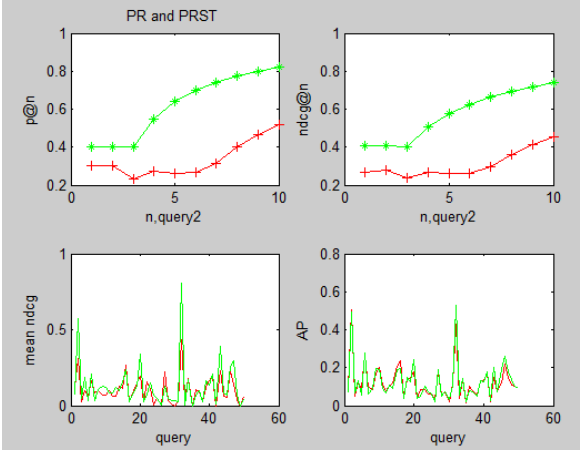

Figure 2. Comparing PR with PRST algorithm in precision, NDCG, AP and mean NDCG

This algorithm is better than the PR in all criteria. Improvement of NDCG is about 26 percent and for precision this value is 39 percent. Improvement of AP is about 0.5 percent and finally average of NDCG is improved about 2.6 percent. Overall this algorithm is better than PRS algorithm. Obtained results are shown in Table 2 and Figure 2.

Table 2. Mean values of evaluation criteria for PR and PRST

Test	PR	PRST	Difference
Mean NDCG	0.1019	0.1287	0.0268
Precision in query 2	0.3336	0.7285	0.3949
NDCG in query 2	0.3103	0.5737	0.2634
AP	0.1229	0.1282	0.0053

3. Comparison between PR and WPCRST (BM25) algorithms

The algorithm combines content weight and age of domain and spam score.

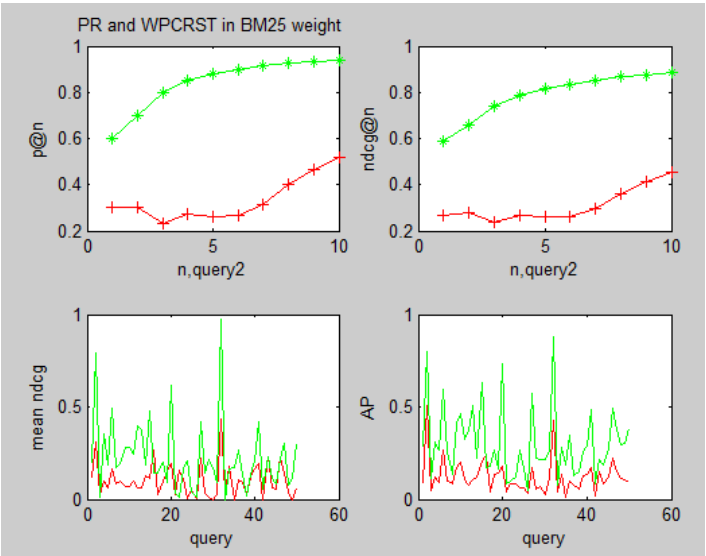

Figure 3. Comparing PR with WPCRST (BM25) algorithm in precision, NDCG, AP and mean NDCG

Table 3. Mean values of evaluation criteria for PR and WPCRST (BM25)

Test	PR	WPCRST (BM25)	Difference
Mean NDCG	0.1019	0.2261	0.1242
Precision in query 2	0.3336	0.8443	0.5107
NDCG in query 2	0.3103	105.2919	0.4797
AP	0.1229	0.3035	0.1821

According to Table 3 and Figure 3 this algorithm is working better than PRST algorithm (except in the precision), so it provides a better ranking. In this algorithm, value of bm25 is considered as content weight. This algorithm ranks better than the PR in all criteria, i.e., 47 percent improvement in the measure of NDCG and 51 percent improvement for precision and 18 percent improvement for AP measure finally in Mean NDCG 12 percent improvement is achieved.

4. Comparison between PR and WPCRST (TFIDF) algorithms

This algorithm combines the content weight and age of domain and spam score. Obtained results specify that this algorithm has a better ranking than PRST algorithm except for the measure of precision. It has worse response compare to the proposed algorithms, but all of benchmarks show that it is working better than the PR algorithm.

About 41 percent improvement can be seen in the measure of NDCG and 48 percent improvement for precision while AP has improved 0.8 percent and finally 2.5 percent improvement can be seen in Mean NDCG. In this algorithm, value of TFIDF algorithm is considered as the content weight. Obtained results are shown in Figure 4 and Table 4.

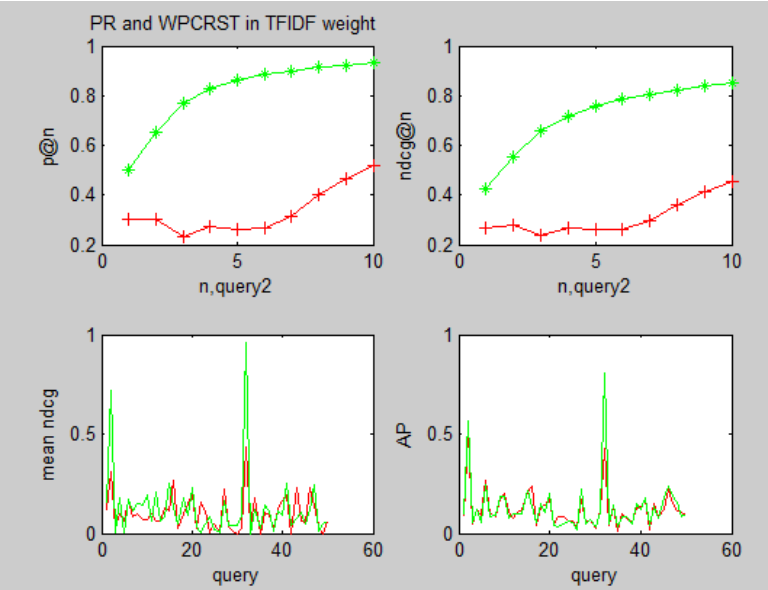

Figure 4. Comparing PR with WPCRST (TFIDF) algorithm in precision, NDCG, AP and mean NDCG

Table 4. Mean values of evaluation criteria for PR and WPCRST (TFIDF)

Test	PR	WPCRST (TFIDF)	Difference
Mean NDCG	0.1019	0.1270	0.0251
Precision in query 2	0.3336	0.8150	0.4814
NDCG in query 2	0.3103	0.7204	0.4104
AP	0.1229	0.1314	0.0085

Comparing purposed algorithms with WPR, PR, ECPR and WPCR

Finally, the proposed algorithms are compared with PR, WPR, WPCR and ECPR algorithms that was explained in the second part of this article. WPR algorithm is working based on the weight of incoming and outgoing links. WPCR algorithm not only uses the weight of inbound links and outbound but also content weight is used in this algorithm. But in ECPR algorithm only weight of inbound links and content weight are used.

Obtained results in Table 5 show that proposed algorithms, PRST and PRS, works much better than PR algorithm and proposed algorithm WPCRST which uses values of BM25 as the content weight and weight of incoming and outgoing links works even better than PR, WPR, WPCR and ECPR and provides better results in the ranking mechanism.

Comparing the proposed algorithm with previous algorithms in standard mean NDCG

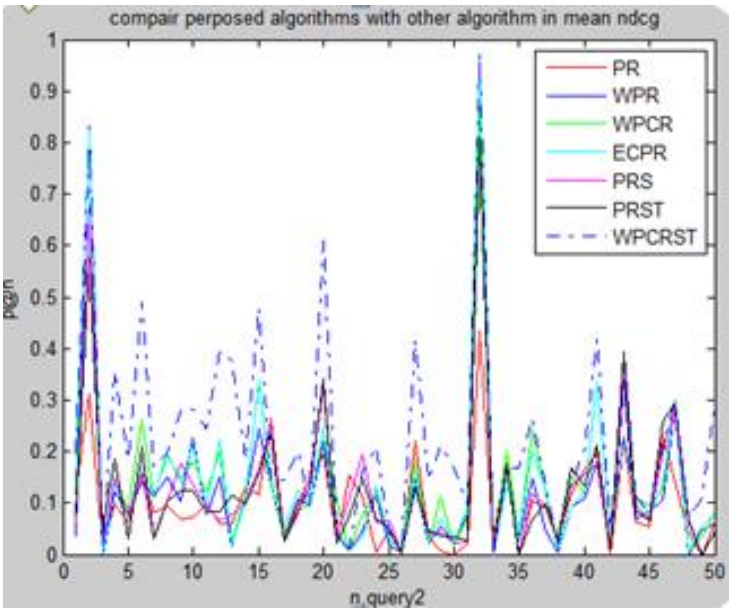

Figure 5. Comparing proposed algorithm with PR, WPR, WPCR and ECPR in standard mean NDCG

In the above diagram WPCRST algorithm that is marked with a blue dotted line in NDCG parameter is better than the other algorithms. This algorithm is 7.8 percent better than the next best algorithm WPCR. It is also works 12 percent better than the PR algorithm.

Comparing the proposed algorithm with previous algorithms in standard AP

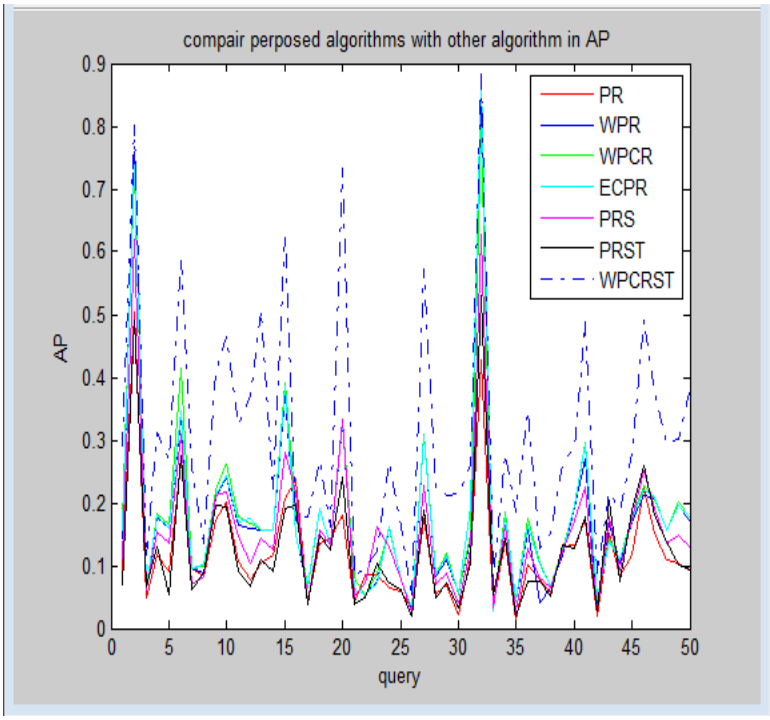

Figure 6. Comparing the proposed algorithm with PR, WPR, WPCR and ECPR the criteria AP

This comparison is done for 50 queries and obtained results show that proposed algorithm WPCRST is much better than other algorithms. Mean of average precision values are shown in Table 5.

Table 5. Values proposed algorithms and algorithms PR, WPR, WPCR & ECPR the criteria AP for better comparison

Algorithm	Mean AP
PageRank	0.1229
Weighted PageRank	0.1729
Enhanced Content PageRank	0.1805
Weighted Page Content Rank	0.1785
PageRank with Spam score	0.1538
PageRank with Spam score and Time factor	0.1275
Weighted Page Content Rank with Spam score and Time factor	0.3035

In the above table WPCRST, ECPR, WPCR, WPR, PRS, PRST and PR algorithms can be evaluated based on mean AP value. PRS algorithm is better than PRST. Age of domain as a criteria have been added to WPCR, ECPR and WPR algorithms and obtained results show that by adding this criteria better results are achieved. The best results are obtained when age criteria is added to these algorithms. These results prove that purposed algorithms is overwhelm other algorithms in the literature.

Conclusions and recommendations

Spam score feature is used in proposed algorithms in this article which is combined with the age of domain and content weight of pages. Obtained results show that the proposed algorithm is much better than PageRank algorithm and rank the pages better than the others.

It is suggested that web usage mining aspects is added to WPRST algorithm because in the previous studies it was demonstrated that the combination of these three aspects leads to a better response. Also for more improvement other positive and negative features can be added to the algorithm such as domain registration length, expires date, trust rank, relevance title with contents of the site, up to date contents of the site, having site map, no errors in the HTML code and etc.

References

Aggarwal, C. C., & Zhao, P. (2013). Towards graphical models for text processing. *Knowledge and Information Systems*, 36(1), 1–21.

Bhardwaj, E Kumar, S. Tomar, K.(2015). Enhancing PageRank Algorithm. *International Journal on Recent and Innovation Trends in Computing and Communication*, 3(5), 3381-3385.

Brin, S., & Lawrence, P. (1998). The anatomy of a large-scale hypertextual web search engine. *Proceedings of 7th World Wide Web Conference*,

Camastra, F., Ciaramella, A., Placitelli, A., & Staiano, A. (2015). Machine learning-based web documents categorization by semantic graphs. *Advances in Neural Networks: Computational and Theoretical Issues*, Volume 37, pp. 75-82, Springer.

Doroudi, E., Baradaran Hashemi, H., Al-Ahmad, A., Zare Bidaki, A. M., Habibian, A.M., Mahdikhani, F., Shakeri, A., & Rahgozar, M. (2008). *A standard benchmark for web information retrieval research Persian*. Technical reports, Research Group Database of the University of Tehran, No.: DBRG-TR-138702.

Ein Abadi, H. (2001). Various methods and capabilities of their web mining. *Fifth Conference on Intelligent Systems*, Jan 2004, Mashhad, Iran.

Kadry, S., & Kalakech, A. (2013). On the improvement of weighted page content rank. *Journal of Advances in Computer Network*, 1(2), 110-114.

Kumar, S. (2016). Web usage mining techniques and applications across industries. *Advances in Data Mining and Database Management*, 1st Edition IGI Global; August 12, 2016.

Lewandowski, D. (2015). Evaluating the retrieval effectiveness of web search engines using a representative query sample. *Journal of the Association for Information Science and Technology*, 66(9), 1763–1775.

Radfar, H. R. (2010). *Web mining: Methods and applications*. PhD Dissertation, Library and Information Science, Islamic Azad University, Science and Research Branch of Tehran.

Robertson, S., & Zaragoza, H. (2009). The probabilistic relevance framework: BM25 and beyond. *Journal of Foundations and Trends in Information Retrieval*, 3(4), 333-389.

Salton, G. (1971). The SMART retrieval system. *Experiments in Automatic Document Processing*, Prentice-Hall, USA.

Xing, W., & Ghorbani, A. (2004). Weighted PageRank algorithm. *Proceedings of the Second Annual Conference on Communication Networks and Services Research (CNSR '04)*, IEEE, 2004.

Bibliographic information of this paper for citing:

Zeraatkar, Ateye & Mirvaziri, Hamid, & Ghazizadeh Ahsaee, Mostafa (2019). "Improvement of page ranking algorithm by negative score of spam pages." *Webology*, 16(2), Article 187. Available at: http://www.webology.org/2019/v16n2/a187.pdf