Improved PageRank algorithm using structural web mining techniques and bee colony algorithm

Helen Namaki[†], Ali Harounabadi[††] and Seyed Javad Mirabedini[†††]

[†]Department of Computer Engineering, Islamic Azad University, Electronic Branch, Tehran, Iran
[††,†††]Department of Computer Engineering, Islamic Azad University, Central Tehran Branch

Summary
The current explosion of information has made use of search engines inevitable. Search engines are also trying to provide better responses to users. In this paper, an extended version of PageRank algorithm will be presented. In the proposed version, first the pre-processing operation is performed on the log file that includes data cleaning, separation of users and identification of sessions. To delete inappropriate inputs from the log file, data clearing will be used. Time-based methods were used to identify sessions and consider the pages as a session. Interests of users in pages are derived based on two criteria: page visiting frequency and duration of stay on page. Then, inspired by the bee colony algorithm, accuracy was improved. The results indicate that the proposed method has less errors compared to rival methods.

Key words:
Search engine, Page ranking, Web mining, PageRank, Bee colony.

1. Introduction

Upon creation of databases and a plethora of information available on the web, for faster access to information, the need for search engines seeker has become inevitable. Search through information became much faster, but the problem was that all data should be stored on a computer and with such high volumes of data, it was virtually impossible. It was then that decentralized searches and search engine emerged and caused a dramatic change in the world. These kind of innovations have a significant effect on the individual's daily life and change the direction of the searching dramatically [1]. On the other hand, these engines are so important that various companies seeking to achieve better financial position are trying to be in the top rankings in terms of such sites and it has been argued that only those industries or firms which introduce an efficient plan for themselves can be successful in this path [2].World Wide Web is a rich source of data that has been continually expanding and increasing its complexity. Thus, effective and efficient retrieval of web pages is a major challenge. Therefore, methods and techniques for efficient data access, data sharing, data mining and use of this information are required.

Manuscript received October 5, 2017
Manuscript revised October 20, 2017

Web can be considered as a tagged graph, whose nodes are documents or pages and whose edges are hyperlinks between pages. Directed graph structure is known as the web graph. There are three link analysis algorithms including PageRank, weighted PageRank and Hits, which is one of the factors used by Google in the calculation of the relative importance of web pages, which will be addressed in this study. The PageRank value of a web page is dependent on PageRank values of pages and the number of outgoing links from those pages. The performance of this algorithm is as follows: pages with further references are more important. Advantage of PageRank algorithm lies in the fact that in determining the importance, it not only uses the number of references to the page but also considers the importance of the referring page.

Accordingly, in this paper, the problems raised were solved by the use of bee colony algorithm. The algorithm is designed based on the collective behavior of honey bees to find food sources. Bee colony algorithms provides simple, random, robust optimization on the basis of aggregate behavior. Our main goal was to improve the page ranking algorithm using web mining techniques and base colony algorithms.

The contents of this paper are organized in five sections. First section deals with literature in the ranking of web pages and search engines. The second part presents the method use in this paper, and the idea of improving PageRank method and combine it with the bee colony algorithm. The results and analysis of the results are provided in the third part; conclusions and possible future research are discussed in Section 4.

2. Related Works

Currently there are two main methods, "content-based" ranking (used in traditional information retrieval) and "structure-based" (used in the current web). In content-based methods that are used in traditional information retrieval, models like probabilistic, vector space and language modelling methods are used to rank documents based on their content. The most important of such techniques in modelling is IF-IDF algorithm and among

probabilistic models, algorithms BM25 is the most important one. In fact, links indicates the quality of the content of a page from the perspective of external pages. (As opposed to text content of page that is totally dependent on its creator). In other words, in link-based ranking, content of other pages is used to evaluate a page. This property makes the ranking algorithm using information extracted from links show less sensitive. Among link-based algorithms are PageRank, Hit [3], host rank [4] and the distance-rank [5].

In recent years, other ranking methods have been proposed including combined methods, learning-based on user click-based methods; the following will described each separately. Murthy et al. suggested and studied Bee Colony Algorithm to search and find the optimal solution; they acknowledged that many Internet users are not only consumers of information, but also information provider. There is a lot of information on the web and many people find what they want via search of the web. One of the problems with a lot of data on the web is that we often spend much time to find correct results from the search results. Often one recommendation is based on bee colony algorithm, which is a population-based search algorithm. Bee colony algorithm is an optimization algorithm inspired by the behaviour of a search for food (food gathering) of natural bees to find the optimal solution. Forsati et al. showed that honeybee colony optimization algorithm is one fast, robust and effective general search method for overcoming various practical problems. Considering this algorithm in this paper, it is used to classify data, which is a fundamental problem that often arises in many applications [6]. Liu and Li proposed a non-symmetrical weighted k-mean classification algorithm to improve the accuracy of the result of classification. The distance of original k-mean algorithm is corrected by adding the non-symmetric weights to distance measure. That is, different weights are used for the attributes in clusters, so that the contribution of the features can be adjusted during the classification process in an adaptive manner. In this work, weights were provided through an optimization process using a rank-based artificial bee colony algorithm. Then, five sets of patient data for medical diagnosis, including breast cancer, heart disease, diabetes, liver disease and hepatitis were applied to assess the effectiveness of the proposed algorithm [7].

Tyagi and Sharma provided weighted PageRank algorithm based on link visit for search engines, in which the number of visits to incoming links of web pages was calculated. The proposed algorithm assigned a higher rank to the outgoing links most visited by users and received the highest number of incoming links. In this algorithm, the popularity of outbound links that are included in the original algorithm is not considered but users' browsing behaviour is calculated based on by link visits [8].

Xing proposed a new type of page ranking algorithm suggests that uses a combination of classification tree and static PageRank algorithms; this algorithm can create classified tree in accordance with a large number of similar search results of, and can significantly reduce the obvious problems of PageRank, and the problem of old web pages cut. In this algorithm, using the keywords, the results obtained by different users can be developed to improve the user calls. Based on this, classification tree can dynamically change in accordance with the change of user search. Finally, the proposed algorithm can efficiently improve search without slowing speed [9].

Xing and Qorbani introduced weighted PageRank algorithm, which is developed form of PageRank algorithm. The weighted PageRank considers importance of inbound links and outbound links and distributes rank scores ranked by popularity of pages. The weighted PageRank has the ability to identify a larger number of pages of a search compared to the standard PageRank [10].

3. Proposed Method

The method proposed here is to use the bee colony procedure to improve PageRank algorithm. Before addressing the steps of the proposed method, some of the settings must be done so that the proposed method can be compared with rival methods under equal conditions. The steps are as follows. Description of the steps that follow are modelled on [11].

3.1 Pre-processing of server logs

This step is the first step in methods to identify web access sessions on main web logs. These web servers are able to record all web user activities. That is why they are called Web server. However, it is to be noted that because different parameters are used in server settings, various types of logs have been created. It can be argued almost that the log files have same basic information. This information can be IP address of the client, the time of request, URL requested or HTTP status code.

3.2. Data Cleaning

Web servers perform multiple explorations and record different inputs. Not all such inbound logs are suitable for exploring the web. To take advantage of inputs, we should only consider inputs that provide us with appropriate information. Then, as the second step, we remove inappropriate inputs. Inputs that should be removed include [12]:

- Because we consider the clear requests of users; inputs related to video, audio and graphic files that are

associated with requests for specific pages must be removed.
- Inbound log that corresponds to the non-fulfilled requests. For example, the requests that faced with HTTP error re removed from the log files.
- Inbound logs that have responses other than "GET" and "POST".

3.3 User Identification

The best and most trusted method to identify users is to use the IP address. To be more precise, IP address of users who have request a web page can be examined assuming that every IP is that of a separate visitor. However, this assumption is not always true. Because these two states can also exist:

1. Some users may use the same IP.
2. A user may use different computers resulting in different IPs.

So we must find a solution to this problem or establish a general convention on which user should be recorded as a new user. The selected solution is as follows:

- If referrer IP address of user is same as previous input in the log file, we go to the next user factor.
- If the user factor is not similar, it is a new user; if it is identical, we refer to referrer URL and site topology.
- If referrer URL is not same, it is recorded as a new user. However, if it is the same, we go to the page requested by the user.
- If the page requested by the user is not directly accessible from any pages that have been visited by this IP; the user is identified as a new user with the same address. Otherwise, it is not recorded as the same user.

3.4 Session Identification

The user session means all the attributes of the user that can be collected during his visit of a page. Since it is possible to collect from the user's visits of a different page user behavioral patterns, several methods have been established in this regard. All methods in this field are divided to two groups: time-based and subject-based. Since the classification of web pages on the subject seems a bit difficult and measuring user interest in a particular subject is difficult; researchers are more inclined to use time-based method. This paper also used this method. In time-based approach, pages are considered as a session when they are requested within a time period less than or equal to a specified length of time. This time in this paper is set at 30 minutes.

3.5 Making session vectors

The session vectors are a series of transactions that includes a series of weighted pages over a user's visit of a page. This means that the user's session can be expressed in a vector of weights related to pages. Here, user's session is shown as follows: if P denotes the set of user-accessible pages of a site; each page Pi (i = 1,2,3, ..., m) has a unique URL. In addition, a set S is considered as set of user access sessions, which will be n in size. Each session Si (i = 1,2,3, ..., n) is shown by an m-dimensional as Si = {ω (P1, Si), ω (P2, Si), ..., ω (Pm, Si)} where ω (Pj, Si) is the weight of j-th page visited in the session Si. However, it should be noted that each page Pi can be accessed again at any user session Si.

The most important point is that the weight ω (Pj, Si) should be able to accurately show interest of user in a web page. This parameter can properly contribute to the PageRank algorithm's efficiency and makes practical use of user profiles. To give weight to pages, factors were incorporated in this paper to measure the amount of interest of user in that page:

Page Frequency: Frequency of a page is the number of times a Web page is visited. It is obvious that the higher the number of times a user visits a page in one session; the more interested the user is in that page and the more importance of that page in that session.

Duration of visiting a page: Time spent by a user on a page is called the page visit duration. If the user visits the page and does not take an interest in it; he will leave the page for another. However, if user is interested, he will spend more time visiting the page.

3.6 Calculating the new page ranking function based on combining of PageRank and user profile information

We already have relation related to PageRank method, and we should just incorporate user profiles in it. User profiles is incorporated as shown in relation (1). The user profile value is denoted by PU.

$$P(i) = (1-d) + d\left(\frac{(1+\gamma^2)\varphi_{out}(PU)}{\gamma^2\varphi_{out}+PU}\right) \quad (1)$$

where

$$\varphi_{out} = \sum_{i,j \in E} \frac{P(j)}{outdegree(j)} \quad (2)$$

The reason for introducing parameter of user profiles in relation (1) is to make it possible to determine influence of the result of standard PageRank and user profiles on final result of the proposed method. This is done by the parameter. Three modes can be considered for the parameters:

- $\gamma > 1$: Weight related to the number proposed by user profile increases.
- $\gamma = 1$: Weight is same in both cases.
- $\gamma < 1$: Weight related to number proposed by standard PageRank increases.

We selected the two factors of page frequency and duration of visiting the page as parameters that must be involved in the calculation of the user PU. Page frequency is given by equation (3) and duration of visiting the page is given by the equation (4).

$$\text{Page frequency} = \frac{\textit{Frequency of page visits}}{\textit{frequency of all web pages visits}} \quad (3)$$

$$\text{Duration of visiting the page} = \frac{\tau_1}{\tau_2} \quad (4)$$

where

$$\tau_1 = \frac{\textit{time of stay on the page}}{\textit{length of the page}} \quad (5)$$

$$\tau_2 = MAX_{\textit{all pages on web}}\left(\frac{\textit{time of stay on the page}}{\textit{length of the page}}\right) \quad (6)$$

The cause of divisions in relations (3) and (4) is becase there is no ceiling for any parameters

- Frequency of page visits
- and "time of stay on the page".

So, whether "time of stay on the page" that is 15 seconds is or too low or too low can determined according to the length of time that we stayed on other pages. The only remaining case is that with use of the frequency and duration of visiting the page, PU parameters is estimated. PU is calculated via the harmonic mean. However it is best for us to have higher frequency and longer visit time. The higher they are, the better the condition. Thus, the harmonic mean of these two parameters equation (7) was used. Because the use of the harmonic mean implies that mean increases when both parameters increase.Thus, the harmonic mean of these two parameters (equation (7)) was used. Because the use of the harmonic mean implies that mean increases when both parameters increase.

$$\textit{Harmonic mean} = \frac{2}{\frac{1}{\alpha}+\frac{1}{\beta}} \quad (7)$$

This is why the harmonic mean was used. So, PU is given by (8).

$$\text{PU} = \frac{2}{\frac{1}{\textit{page visit frequency}}+\frac{1}{\textit{time of stay on page}}} \quad (8)$$

Use of the bee colony in this paper is to identify the appropriate values for the parameters d and γ that exist in relation (1). Bee Colony Algorithm searches for appropriate values for these parameters by creating appropriate random values within a given interval. Suitable values of these parameters can greatly contribute to better functioning of the proposed method to help us in finding sustainable solutions. The only question that remains is how fit function is defined for bee colony to measure the value fitting each bee, which contain values for the two parameters d and γ. This function is as follows:

$$\text{Fitness} = \frac{\sum_{u=1}^{n} \textit{ERROR}(u)}{n} \quad (9)$$

The rank of each bee is given by average error of all pages in all the days. The Error function is the figure given by difference of ranks of a page according to the two methods, standard PageRank and the proposed method. In Error relation, page rank obtained using standard PageRank is denoted by RealRank function and rank of page as obtained using the proposed method is shown by BeeRank function.

$$\text{ERROR(u)} = \sum_{T=1}^{t} \frac{\textit{RealRank}(u,t) - \textit{BeeRank}(u,t)}{T} \quad (10)$$

Page ranking using the proposed method is done using relation (1). For the time T, given PageRank of previous times, equation (11) is used.

$$\textit{BeeRank}(u,t) = \textit{BeeRank}(u,T-1) * \frac{T}{t} * PU \quad (11)$$

4. Experimental Results

This section details the implementation and results of the proposed method compared with rival methods. Among the most important bases for scientific research paper is assessment and evaluation of the results of the proposed method. Database that is used in this paper is NASA database, which is a server log file of size of 195 MB. It has already been used as the standard database in many analyses in the field of ranking web pages [13]. General specifications of the database in the context of the user's profile are as shown in Table 1.

Table 1: Specifications of NASA database of user profiles

Day	No. of Entries	No. of IP address	No. of Unique Users	No. of Hits	Failures
1	64567	7597	576	20893	2931
2	60264	6630	613	21995	2463
3	89565	19193	1766	17390	1738
4	65536	9340	868	8545	795
5	6535	17638	1039	12564	1421
6	68342	24706	2033	22398	2304
7	87233	33657	1765	2324	2897

First, the parameters of the bee method are mentioned, by which implementation was performed. The number of iterations was equal to 30. Also, the size of the bee population was set at 20. To be more precise, 20 bees were randomly generated with each bee assuming two proposed values for the parameters γ and d. It should be noted that d denoted the adjustment factor as used in the equation (1) is adjustable in the interval [0, 1]. Also, the parameter γ

denotes the result from standard PageRank and determines user profiles in the final result.

After determining the best bee from among these 20 randomly generated bees; all the bees were updated using such good bees to get adequate levels of γ and d. Values obtained by using bee colony method for γ and d are shown in Table 2. Bee Colony Algorithm searches for appropriate values for these parameters by creating appropriate random values within a given interval. Suitable values of these parameters can greatly contribute to better functioning of the proposed method to help us in finding sustainable solutions.

Table 2 : Appropriate values for parameters γ and d

Parameter	Value
γ	1.02
D	0.82

In the following, ranks derived from the proposed algorithm are compared with those obtained from the standard PageRank algorithm. The number of distinct ranks in each method and the error of both methods are specified.

Before dealing with any subject, error criteria should be used to compare the two methods, standards PageRank and the proposed method. Two conventional criteria used by researchers in the past things were number of distinct ranks and total error:

- Number of distinct ranks: Numbers obtained from ranking of different pages; the more the number of them, the better the efficiency. To be more precise, if multiple pages have the same rank, the proposed method does not know which one is superior to another page and proposes by chance one of them. So, as a result, the more different the values obtained for page ranks, the clearer the prioritization of the pages will be.
- Total resulting error: Because the number of pages to which the proposed approach and method of Standard Page Rank are applied is 224; calculation error for pages was obtained by calculating of the difference of the obtained rank and the actual ranks. After calculating the error rate for the entire 224 pages, the error values were summed up to get the total error of the method. Error is given by the equation (9).

$$\text{ERROR} = \sum_{i=1}^{224} |y_i - x_i| \quad (12)$$

X_i is obtained rank and y denotes one mentioned in the database. Error obtained for the proposed method and the standard PageRank method along with distinct ranks for each method is summarized in Table 3.

Table 3: Error obtained with distinct errors of rival methods

Method	No. of distinct errors	Error
Proposed method	212	0.30
Standard PageRank	102	0.43

The only thing that remains are the two basic diagrams. The first diagram relate to ranks obtained for each page as shown in Figure 1. Figure 1 shows ranks obtained by Standard PageRank method along with ranks obtained by the proposed method. Difference between the ranks is clear; obviously, the proposed method has completely changed the ranking system and reached better results.

Blue denotes standard PageRank method and red the proposed method. The second diagram shows the error obtained for each page as shown in Figure 2. Blue and red denote standard PageRank method and the proposed method, respectively. It is clear from Figure 2 that the errors related to standard PageRank method are more than errors of the proposed method.

According to the results obtained and the items listed, it can be argued that the proposed method has good capability and is a good development for standard PageRank method, which have better results.

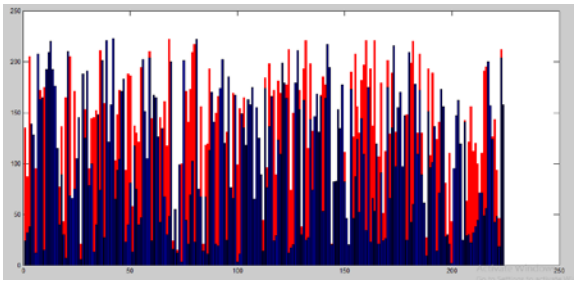

Fig. 1. Diagram of ranks proposed for each page. Blue color relates to standard the PageRank and red color relates to the method proposed.

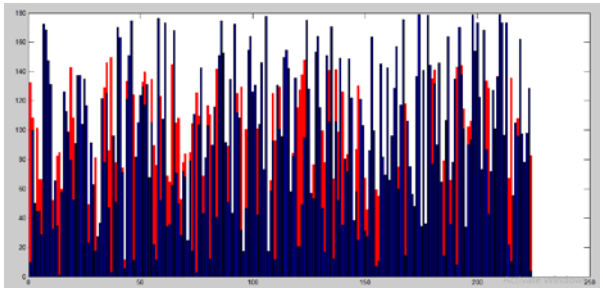

Fig. 2. Diagram of error obtained for each page. Blue color relates to standard the PageRank and red color relates to the method proposed.

5. Conclusion

In this paper, an extended version of PageRank algorithm was presented, mostly applicable in the field of web usage mining. In the proposed version, the pre-processing

operations are first performed on the log file. In our work, operations included data cleansing, separation of user and identification of sessions. To delete inappropriate inputs from the log file, data cleansing was performed. Time-based method was used to identify sessions and pages were considered as a session if they were requested in a period of less than or equal to 30 minutes. As a next step, users' interests in pages were derived based on two criteria of "frequency" and "time of stay on a page". Then, inspired by bee colony algorithm, it was used for estimation. For this purpose, each user was viewed as a bee consider. The proposed version adds a new parameter to the PageRank, which parameter is to incorporate user profile in PageRank method. The results indicate that the proposed method has good power compared to rival methods.

References

[1] Almasifard, M. (2013). An Econometric Analysis of Financial Development's Effects on the Share of Final Consumption Expenditure in Gross Domestic Product (Master's thesis, Eastern Mediterranean University (EMU)-Doğu Akdeniz Üniversitesi (DAÜ)).

[2] Khorasani, S. T., & Almasifard, M. (2017). Evolution of Management Theory within 20 Century: A Systemic Overview of Paradigm Shifts in Management. International Review of Management and Marketing, 7(3), 134-137..

[3] G. R.Xue, Q. Yang, H. J. Zeng, Y. Yu, Z. Chen, "Exploiting The Hierarchical Structure For Link Analysis", Proceedings Of The 28th Annual International Acm Sigir Conference On Research And Development In Information Retrieval, 2005.

[4] A. M. Bidoki, N. Yazdani, "Distancerank: An Intelligent Ranking Algorithm For Web Pages". Information Processing & Management, 2008.

[5] M. Murthy, G. G. Sriram, B. Abhiram, "Enhanced Bee Colony Optimization Mechanism In Content Recommendation System, 2016.

[6] R. Forsati, A. Keikha, M. Shamsfard, "An Improved Bee Colony Optimization Algorithm With An Application To Document Clustering". Neurocomputing, 2015.

[7] T. Liu, "Learning To Rank For Information Retrieval", Springer Science & Business Media, 2011.

[8] N. Tyagi, S. Sharma, "Weighted Page Rank Algorithm Based On Number Of Visits Of Links Of Web Page", International Journal Of Soft Computing And Engineering (Ijsce), 2012.

[9] C. Tian, "A Kind Of Algorithm For Page Ranking Based On Classified Tree In Search Engine", International Conference On Computer Application And System Modeling (Iccasm 2010), 2010.

[10] W. Xing, A. Ghorbani, "Weighted Pagerank Algorithm. Communication Networks And Services Research, Proceedings. Second Annual Conference On, 2004.

[11] S. Setayesh, A. Harounabadi, A. Rahmani, "Providing a developed version of page rank for ranking webpages", 8th conference on computer engineering & sustainable development, 2013.

[12] S. F. Rashidi, A. Harounabadi, M. Abasi Dezfouli, "Prediction Of Users' Future Requests Using Neural Network", Management Science Letters, 2012.

[13] K. R. Suneetha, R. Krishnamoorthi, "Identifying User Behavior By Analyzing Web Server Access Log File", International Journal Of Computer Science And Network Security, 2009.