Presenting a Way to Improve Web Page Ranking Algorithm Using Firefly Algorithm

Fariba karimi
Department of Computer engineering ،Damavand science and Research Branch ،Islamic Azad University ،Damavand ،Iran . and
Department of Computer engineering، Damavand Branch ،Islamic Azad University ،Damavand ،Iran

Ali Harounabadi
Department of computer, Islamic Azad University,Central Tehran Branch Tehran, Iran

Seyed javad mirabedini
Department of computer, Islamic Azad University - Central Tehran Branch Tehran, Iran

Abstract
According to daily growth in the volume of data and web development, the need for methods and techniques in order to derive useful information, have risen more than before. So to enhance Website Performance, Web server activities based on interests and the profits of users have been changed. In order to provide better and more efficient results to users, many ranking algorithms on the web pages have been used. In this study, we try to make changes in the standard algorithm to achieve developed version that unlike standard algorithm, considers the users' interests in web pages to calculate Page Rank, which leads to better and more relevant results. The simulation results show that the proposed algorithm presents more suitable page ranks and generates better and more distinctive ratings. In comparison, the proposed algorithm is improved 1.35% compared to the standard Page Rank algorithm and 1.66% compared to Page Rank (VOL).

Keywords
Web mining, ranking, PageRank algorithm, fireflyy colony algorithm

1. Introduction

Today Web is a popular environment to disseminate information. Unique features of the World Wide Web and high availability of the information disseminated on it, caused an impressive growth in its volume of data. As well as other information environments, organizing information on the Web environment is essential in order to facilitate and accelerate access to it. In order to organize websites and to achieve the mentioned goals, the ranking of web pages is one of the important techniques. Several ranking algorithms have been used on web pages, such as: Standard Page Rank algorithm, Weighted Page Rank algorithm, HITS and .... in fact the ranking algorithms of Web pages are fundamental components of search engines.Their goal is to generate a rank for each web page. This means to provide a criterion that predicts how important and valid are the pages visited by the users. These algorithms, greatly reduce the search space. In this paper, considering users' interests, the standard Page Rank algorithm is improved. The Web server logs are accessed through studying the users' interests and firefly colony is used to improve the ranking algorith. In the proposed algorithm, we considerour web pages as fireflies and users' interests in pages as a factor which represents the attractiveness of pages and the amount of accumulated luciferin. In the following, these issues are discussed:
In the second part, the basic concepts and the context for our work are discussed. The third section explores some of the most important ranking algorithms of web pages. In the fourth section, the proposed method is introduced. Inspired by the firefly colony algorithm, the method to update users' interests is described and in the last step, the developed version of Page Rank algorithm to rank pages is presented. In the fifth part, the results of experiments and comparing them to the other methods are discussed.

2. Context of the research

In this section it is necessary to introduce the basic concepts of web mining and the algorithms used:

A. Web Mining

In [1] Johnson and Kumar Gupta have suggested web mining as a process of discovering the knowledge or unknown and potentially useful information from the Web data. Web mining is used to record relevant information, to create new knowledge among irrelevant data, to personalize information, to learn about the customer or individual users and user groups. Web mining uses data mining techniques to automatically extract and discover information on the Web. According to [2] web mining process can be divided into four steps:

Manuscript received December 5, 2016
Manuscript revised December 20, 2016

- Discovering the Source: to recover and choose a suitable data source from web documents.
- Choosing information and preprocessing: After choosing a suitable source to do the web mining, we need to change and process the raw data of the source.
- Generalization: General patterns are discovered automatically by the machine learning algorithms or data mining techniques.
- Analysis: Finally, output of previous steps is used to interpret and analyze the results in order to obtain useful results.

Web mining is done through three aspects of content, usage and structure: According to [1] web content mining studies the content of web pages (content includes text, audio data or graphical data) in order to discover useful information through Web content and documents. As described in [3], web structure mining tries to analyze nodes and structural links in a website through perception of structure of the hyperlinks. According to [4] web usage mining is used in order to discover patterns from Web logs file. The primary sources of data in order to web function mining include categorized text files from multiple web servers. Web usage mining includes four phases of data collection, data preprocessing, pattern discovery and analysis.

B. Web pages ranking:

A search engine's ranking algorithm is one of the main issues determinining the ability and quality of a search engine. According to [5], web pages search algorithms are divided into two categories of query-independent algorithms and query-dependent algorithms:

- Query-independent algorithms: It gives a point to a document once (independent of the user's query) and then uses this point for all query results. In fact, carries out the query considering all web pages [5].
- Query dependent algorithms: It needs to analyze the link, and assigns a point to each page regarding the subset of web pages relevant to the issue of the query [5].

C. Firefly Colony

This method is proposed by an inspiration from the behavior of Fireflies in the nature and indicates that fireflies live in colonies and cooperate for the survival of the colony. Generally, in order to model the behavior of fireflies, three following assumptions will always be considered [6]:

- All fireflies are homogeneous.
- Attractiveness of each firefly is related to its level of brightness.
- Brightness of firefly is determined with an exponential objective function.
- Singh and Arora [7] state that each firefly always emits a kind of light that by which attracts other fireflies. The amount of accessed light (the attractiveness of each firefly for other fireflies) depends on parameters such as distance and absorption coefficient of the surroundings. The longer the distance the lesser the amount of accessed light will be. Also in surroundings with high light absorption coefficient such as foggy weathers, the intensity of light decreases. The certain issue is that every firefly regardless of its gender has always been attracted to and moved toward the brighter firefly.

The factor of producing light in fireflies is secretion of a substance called luciferin, which is actually a factor to assess the attractiveness of each firefly and it is updated according to equation (1) [8].

$$\ell_i(t+1) = (1-\rho)\ell_i(t) + \gamma J(x_i(t+1)) \qquad (1)$$

In which $\ell_i(t)$ and $\ell_i(t+1)$, are respectively quantities of the luciferin for the moments t and t + 1, ρ is evaporation constant of luciferin)0<ρ<1, γ is replacement rate of luciferin and $J(x_i(t+1))$is value of the objective function in i 'th place at the moment of t + 1[8].

3.An overview of the web page ranking algorithms

Hence, many web page ranking algorithms have been presented based on the Web usage, structure and content mining that in the following they will be briefly discussed.
According to [9], Page Rank algorithm, which is a query-independent algorithm, is one of the most important factors used by Google to calculate a Web page rank. The Page Rank value of each web page depends on the Page Rank values of pages referring to it and the number of external links of those pages.
Page Rank of a page is calculated by equation (2) as:

$$PR(u) = (1-d) + d * \sum_{v \epsilon B(u)} PR(v)/N_v \qquad (2)$$

In the mentioned formula, d is a probability that a consistent user clicks on the links and d-1is the probability that a user jumps to a random page. PR (v) is Page Rank of page v and $N_v$is the outbound degree of page v. Also B (u) is a set of nodes that has an inbound link to u. it means u has been referred by them. d is a modifying factor adjusted to a value between 0 and 1 (usually 0.85) [9].
In [10], According to the standard Page Rank algorithm, weighted Page Rank algorithm is presented. Weighted Page Rank algorithm unlike Page Rank algorithm considers the importance of inbound links and outbound links of the pages in calculating the Page Rank of a page. It also distributes rank scores based on the popularity of pages. The popularity is determined by the number of inbound links

and outbound links respectively as $W^{in}_{(v,u)}$ and $W^{out}_{(v,u)}$. $W^{in}_{(v,u)}$ is the weight of the link (u, v) calculated by the number of inbound links of page u and the number of inbound links of all reference pages of page v.

$$W^{in}_{(v,u)} = \frac{I_u}{\sum_{p \in R(v)} I_p} \quad (3)$$

I_u and I_p respectively *show* the number of inbound links of page u and page v and R (v) shows the list of reference pages of page v[10].

$$W^{out}_{(v,u)} = \frac{O_u}{\sum_{p \in R(v)} O_p} \quad (4)$$

$W^{out}_{(v,u)}$ is the weight of link (u, v) calculated based on the number of outbound links of page u and the number of outbound links of all reference pages of page v.

O_u and O_p respectively are, the number of outbound links of page u and v and R (v) shows the list of reference pages of page v [10]. And accordingly the relationship (5) has been represented:

$$PR(u) = (1-d) + d \sum_{v \in B(u)} PR(v) W^{in}_{(v,u)} W^{out}_{(v,u)} \quad (5)$$

In [11], HITS algorithm in order to rank pages using the hyperlink feature is proposed. The pages are divided into two categories of Authority and Hub. Authorities are the pages containing important content and receive large number of inbound links from other pages. Hub pages are the pages containing useful links to other pages and have outbound links to many other pages. In this way, according to Authority and Hub's scores, a weight will be assigned to each page and pages are arranged based on their importance. In Fig (I) a view from a structure of hypothetical set of Hubs and Authorities is shown.

So, Authority and Hub scores are calculated according to formulas (6) and (7):

$$a_p = \sum_{q \epsilon B(p)} h_q \quad (6)$$

$$h_p = \sum_{q \epsilon F(p)} a_p \quad (7)$$

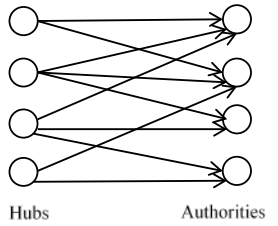

Fig. I. Structure of a hypothetical set of Hubs and Authorities

In [12], page ranking algorithm inspired by the standard PageRank algorithm called Page Rank based on visits of links (VOL) is proposed, which calculates the number of visits of inbound links for web pages. The proposed algorithm allocates more scores to the outbound links that have the most visits by the users. Consider Equation (8):

$$PR(u) = (1-d) + d \sum_{v \epsilon B(u)} L_u (PR(v))/TL(v) \quad (8)$$

L_u determines the number of link's visits referring to the page u through page v. TL (v) determines the total number of visits of all links of the page v. Other parameters are similar to the original Page Rank formula in this equation. Figure (II) shows an example of a Web graph along with its visits of links. The rank of each page is also expressed as follows:

$PR(A) = (1-d) + d(PR(C) * 2/2)$

$PR(B) = (1-d) + d(PR(A) * 1/3)$

PR(C) = (1-d) +d ((PR(B)*2/2)+(PR(A)*2/3))

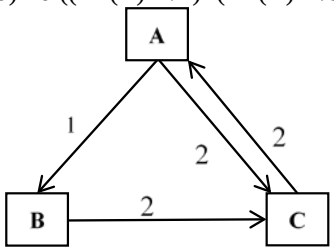

Fig. II. An example of a Web graph along with its visits of links. [12]

Many other algorithms also have been proposed, but because of comparing the proposed method with mentioned algorithms, only these ones are mentioned.

Table (I) shows the Comparison of three primary and main ranking algorithms that other methods have been developed based on them, [13] and [9].

Table I.A comparison of Page Rank, Weighted Page Rank, HITS algorithms.[9] and[12]

Benefit	limit	The complexity	Input parameter	Technique	algorithm
fights against spam	Query-Independent	O(log N)	Inbound links	web Structure mining	Standard Page Rank
fights against spam	Query-Independent	< O(log N)	inbound and outbound links	web Structure mining	Weighted Page Rank
In addition to the structure of the links, also considers the concept of the pages.	Topic drift flow and efficiency problems occur	< O(log N)	inbound and outbound links and concept	web Structure and concept mining	HITS

4. The proposed method

The steps of proposed algorithm are as follows:

A. Pre-processing of web server logs in order to extract user sessions .(Pre-processing of web server logs including data cleaning, user identification and the user sessions identification).
B. User Feature Extraction using their series of sessions
C. Creating a vector of sessions
D. Creating user profiles
E. updating interests on each page inspired by the firefly algorithm
F. rank pages according to the results

In this study, the set of web server log file data CTI [14] has been used for a period of two weeks. The file has been pre-processed using Microsoft SQL Server. After pre-processing the raw data and separating users based on their IPs and identifying them along with their sessions, user sessions with threshold duration of 30 minutes were determined [15].

Also we have to remove inappropriate pages. Pages appearing in less than 10% and more than 80% of the total number of accesses in the sessions (such as home pages) have been removed. Also all user sessions with a length less than three were eliminated and 20946 sessions remained. Finally, the number of pages that ranking operation was performed on were 359 pages [16].

The session vector is a set of transactions that includes weighted pages that were visited during a specified time interval. In other words, user sessions can be expressed as a vector form of weight of the pages.

Consider P as a set of pages accessible by users of a site such that P=$\{p_1, p_2, \dots, p_m\}$and each $p_i is\ a$ page with a unique URL. S is a series of user access sessions that is defined as S=$\{s_1, s_2, \dots, s_m\}$ where each s_i is a subset of P. Each s_i session is shown with an m-dimensional vector as $s_i = \{w(p_1, s_i), w(p_2, s_i), \dots, w(p_m, s_i)\}$ that each $w(p_j, s_i)$ is determined weight for jth Web page visited in $the\ s_i$session. It should be noted that each web page $of\ p_i$ can be repeated at the session.

Weight of $w(p_j, s_i)$ indicates users' interest in a web page. To determine this according to [17], we use the parameters of "frequency" and "duration". The frequency is defined as the number of times a page is visited. User may see a page several times in a session.

The more number of observations of a page at a session means that page is a more important page in the session.The visiting duration of a page is the time spent on a page that can express the importance of the page. Because if a page is attractive for the user, they spend more time visiting it otherwise they close the page and go to another page. According to [17] to calculate two mentioned criteria the formulas (9) and (10) are used.

$$\text{Frequency(page)}=\frac{\text{Number of visits(page)}}{\sum_{\text{page}\epsilon\ \text{visited pages}} \text{Number of visits(page)}} \quad (9)$$

$$\text{Duration(page)} = \frac{\text{Total Duration(page)/Lenght(page)}}{\max\limits_{\text{page}\epsilon\ \text{visited pages}} (\text{Total Duration/Lenght(page)})} \quad (10)$$

Since the importance of the entire page depends on both mentioned parameters, and an interest in one page is high when both parameters are high, according to [18], in order to weight pages, harmonic mean of mentioned parameters is used according to equation (11).

$$\text{Interest(page)}=\frac{2*\text{Frequency(page)}*\text{Duration(page)}}{\text{Frequency(page)}+\text{Duration(page)}} \quad (11)$$

Then user profiles are built and session vectors of different users are separated. Consider $s_1, s_2, \dots, s_k$as a set of sessions of i'th user (u_i). So the mean vector of $S_{ui}$for the user u_i is calculated as an indicator of favorite pages of the user and the weight of each web page on mean vector will be obtained from the average weight of that web page in all sessions of the user $(s_1, s_2, \dots, s_k)$.

So far, we have gained the amount of interest of each user in a web page. Now, we will apply the firefly algorithm. So there are following assumptions:

1. Consider each page as a firefly.
2. The amount of users' interest in a web page is considered as the amount of luciferin secreted by each firefly.

Each user visits pages during his sessions and leaves some amount of interest on that pages which, as previously described, is obtained through two criteria of "frequency" and "duration".

As a result, the higher the average interest on a page is, that page is more attractive. This amount of attractiveness reflects the amount of luciferin of each firefly. So, the pages that have not been visited so much will have lesser amount of luciferin. According to equation (1) extended version of Page Rank algorithm is expressed as (12).

The amount of luciferin on each date can be considered as the rank of each page on that date, which is directly related to average amount of users' interest and also depends on the page rank of that page on the previous date.

$$PR(u)_{(t)} = (1-\rho)(PR(u)_{(t-1)} * Avr(interest)_{(t)}) + \gamma[(1-\text{d}) + \text{d} * (\sum_{v\epsilon B(u)} \frac{PR(v)}{N_v})] \quad (12)$$

In the mentioned equation, $PR(u)_{(t-1)}$ and $PR(u)_{(t-1)}$, are respectively, the rank of page u on the date of t-1 and on the current date of (t).p is evaporation constant of luciferin ($0<\rho<1$) and γ is replacement rate of luciferin. $Avr(interest)_{(t)}$represents the average amount of users' interests who have observed the page u on the date of t. d shows the probability of a user clicking constantly on the links and (1-d) is the probability of a user jumping to a random page. In fact, d is a modifying factor and is adjusted usually between values of 0 and 1, (0.85) for the web graph. PR (v) is rank of page v and B (u) is set of pages that have an inbound link to page u. N_v is a set of pages having an outbound link to page v.

Figures (III) and (IV) respectively, show ranking of 100 pages and all pages of Web server. As shown in the diagrams, the proposed algorithm obtained better results of the ranking scores.

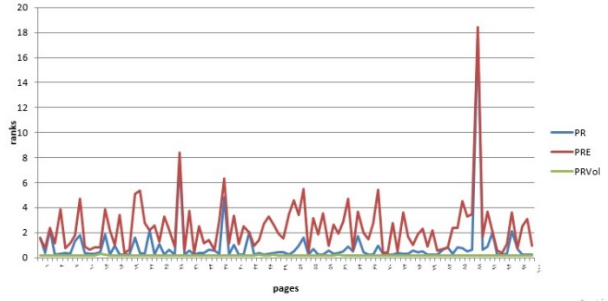

Fig. III. Comparing the ranking of 100 pages in the proposed method, and Page Rank algorithm and Page Rank (VOL) algorithm

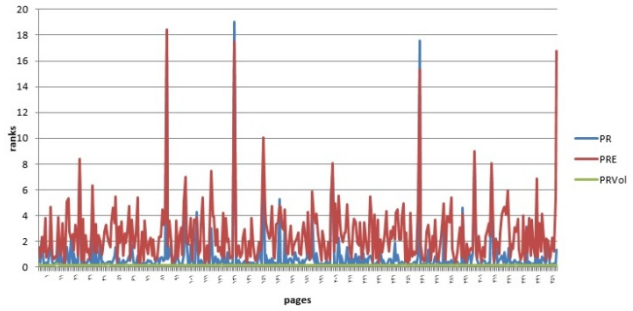

Fig.IV. Comparing the ranking of all pages in the proposed method, Page Rank algorithm and Page Rank (VOL) algorithm

According to [19] when several pages have the same rank, none of them are superior to other pages, for example, among 15pages, the algorithm considering 15 different ranks for these pages is more efficient than the algorithm considering 10 different pages.

In fact, this means that the second algorithm does not distinguish between different pages and identifies lower number of relevant and important pages. So, obtaining unique and distinguished results is a useful feature for ranking.

Table II. The results of the ranking of all pages

The number of unique ranks	algorithm
277	Page Rank
213	Page Rank(VOL)
357	Proposed method(PRE)

5. Conclusion

In this study, inspired by the firefly algorithm, a developed version of Page Rank algorithm is presented that is a compilation of web usage mining and web structure mining. First we did preprocessing operation on log file (including data cleaning, user identification and session identification) and then tried to create vector of sessions and profiles for users. Then, inspired by the firefly algorithm, we used luciferin update part of this algorithm to update web page ranks. In the next step using the developed version of Page Rank, we tried to rank pages.

As can be seen, the formula presented in this study, provides better ranking results using the average amount of users' interest in a page. Also it's been determined that in addition to the structural parameters influencing the Page Rank, the rank of each page on each date depends on its attractiveness for users on that date, too.

The results show that, the proposed method, compared to the standard Page Rank algorithm and Page Rank (VOL) produce more unique ranks which lead to propose more important and more relevant pages and to make pages for users easier to access.

6. Future Works

According to studies and researches conducted and the results of implementing this study, the followings are suggested:

- You can use other characteristics that can be extracted from the log file, such as history of visiting the pages, and sequences of access to pages and …in order to identify users' interest.
- You can do the ranking process as personalized to obtain a different web page rank score for each user. This method especially will be useful in the systems in which the users have profiles on the website.
- You can also measure the influence of users' interest on page ranks in other ranking algorithms or in their combination.
- You can do this process on social network websites having demographic information of users, and thus its influence on the rank of the pages can be investigated.

References

[1]. F, Johnson, K, Gupta, S.,"Web Content Mining Techniques: A Survey". International Journal of Computer Applications. Vol.47, No.11. pp 44-50, 2012

[2]. T, Munibalaji, C, Balamurugan, "Analysis of link algorithms for web mining". International Journal of Engineering and Innovative Technology (IJEIT). vol.1, Issue 2, pp 81-86, 2012.

[3]. T., Pamutha, S. Chimphlee, C. Kimpan, P. Sanguansat. "Data preprocessing on web server log files for mining users access patterns". International Journal of Research and Reviews in Wireless Communications (IJRRWC). UK, Jun, vol.2, no.2, pp.92-98, 2012.

[4]. V.R.R.Nagarjuna, B., Ratna babu, A., Markandeyulu, M., A.S.K.Ratnam."Web mining: methodologies, algorithms and

applications". International Journal of Soft Computing and Engineering (IJSCE). Jul, vol.2, Issue 3, pp 164-167, 2012.
[5]. M. Punjani, V, Gupta,. "A Survey on Data Preprocessing in Web Usage Mining". IOSR Journal of Computer Engineering. Vol.9, Issue4, pp 76-79, 2013.
[6]. Yang, Xin-She. Engineering Optimization: An Introduction with Metaheuristic Applications. Chapter 17. John Wiley & Song Publishing, 2010.
[7]. S. Arora, S. Singh,. "The Firefly Optimization Algorithm: Convergence Analysis and Parameter Selection". International Journal of Computer Application. Vol.69, No.3. pp 48-52, 2013.
[8]. K, Krishnanand, N., Ghose.. "Glowworm Swarm Optimization for Searching Higher Dimensional Spaces". Innovations in Swarm Intelligence, SCI 248, pp 61-75, 2009.
[9]. M, Selvan, P., Sekar, A, C., Dharshin. "Survey on Web Page Ranking Algorithms". International Journal of computer Applications. Vol.41, No.19. pp 1-7, 2012.
[10]. W., Xing, A, Ghorbani. "Weighted PageRank algorithm". Proceedings of the Second Annual Conference on Communication Networks and Services Research (CNSR'04), IEEE. May 19-21, pp 305- 314, 2004.
[11]. M., Kleinberg, " Authoritative Sources in a Hyperlinked Environment". Journal of the ACM, Vol. 46, No. 5. pp 604 – 632, 1999.
[12]. G. Kumar, N. Duhan, A.K, Sharma. "Page ranking based on number of visits of links of web page". International Conference on Computer & Communication Technology (ICCCT), IEEE. Allahabad, India, Sep 15-17, pp 11-14, 2011.
[13]. R., Fotoohi, D, Abdipour, "Efficiency assessment of the rankings page for the extraction of web pages". Journal of Computer Engineering and Sustainable Development, 1 to 13, Mashhad, 1392.
[14]. www.facweb.cs.depaul.edu. Access Time: winter 2016.
[15]. G,Forsaty, D, Meybodi,. "Algorithm based on link structure of the web pages and using data users to offer". Iran Data Mining Conference, 1 to 12, Amirkabir University of Technology, 1387.
[16]. G,Castellano, A.M. Fanelli, M.A, Torsello. "Newer: a system for neuro-fuzzy web recommendation". Applied Soft Computing vol. 11, Issue 1, pp 793-806,2011.
[17]. M, .Kolah kaj, A. Haron Abadi, M., Sadeghzadeh, "Providing a way to personalize the web using neural network." The first National Conference on Emerging Trends in Computer Engineering and retrieval of information, from 1 to 5, Islamic Azad University of Roudsar and Amlash. 1392.
[18]. S.F., Rashidi, A. Harounabadi, M. Abasi Dezfouli. "Prediction of users future requests using neural network". Management Science Letters. vol.2, Issue 6, pp.2119-2124, 2012.
[19]. H. Dubey; B,, Roy, "An Improved Page Rank Algorithm based on Optimized Normalization Technque".International Journal of Computer Science and Information Technologies. Vol.2(5). pp 2183-2188, 2011.