Presentation of an Extended Version of the PageRank Algorithm to Rank Web Pages Inspired by Ant Colony Algorithm

Sara Setayesh
Department of Computer Engineering, Science and Research Branch, Islamic Azad University, khouzestan, Iran

Ali Harounabadi
Department of Computer Engineering, Tehran Center Branch, Islamic Azad University, Tehran, Iran

Amir Masoud Rahmani
Department of Computer Engineering, Science and Research Branch, Islamic Azad University, Tehran, Iran

ABSTRACT
The general search engines represent various results in their lists, which is very time consuming to check. . One way to limit the search engine results is to use ranking pages algorithm in web. One of the most important ranking algorithms of web pages in the internet is known as "PageRank", which works on the web-graph structure.

In this article, an extended version of the "PageRank" algorithm taking into consideration the degree of user interest in web pages and the ant colony algorithm is presented. Simulation results indicate that in the recommended algorithm ranks are closer to the real data; they produce more distinguished ranks, and have a less amount of errors.

Keywords
Web mining, Ranking, PageRank, Ant Colony

1. INTRODUCTION
Nowadays, web is a favored environment of conversation and information publication. The traditional web search engines return a lot of results for searching, which reviewing all of them is very time consuming for users. To overcome these problems, the web mining technique is used. Web mining, is a technique of data mining that automatically extracts the data from web documents. Based on the data gathered from the World Wide Web, the web mining is divided into 3 groups: web content mining, web structure mining, and web usage mining [1]. The web content mining deals with information or the science discovered from web pages content. The web structure mining discovers the relations between web pages, by analyzing the web structure. According to the hyperlinks, the web structure mining organizes the web pages in groups and produces relative patterns, such as similarities and relations between various web sites. The web usage mining is the process of discovering the fact that what are the users searching for in the internet.

Various ranking algorithms have been applied on the web pages till now to support the users who surf the search engine result lists. Actually, the ranking algorithms of web pages are a fundamental element of the search engines. Their goal is to determine a rank for every web page, which means that they would determine the importance and validity of each web page. The ranking algorithms greatly reduce the search area.

One of the most important ranking algorithms of web pages is the "PageRank" algorithm, which is considered by Google to determine the relative importance of the web pages. In this article, we focus on this algorithm, and represent an extended version of it.

The rest of this paper is organized as follows:

In the second section the research background has been described. In the third part the previous ranking algorithm will be investigated and in the fourth, the suggested method of this study is to be introduced. The fifth section explains the details regarding the implementation and analysis of the suggested method and eventually on the sixth part, the result and the strong points of the suggested method are going to be explained.

2. BACKGROUNDS
In this section the PageRank and the ant colony algorithm used in this article is explained.

2.1 The PageRank Algorithm
The PageRank algorithm presented by Brin, Page and et al [2, 3], is one of the factors used by Google to calculate the relative importance of the web pages. The PageRank value of a Web page depends on the PageRank values of pages pointing to it and on the number of links going out of these pages [4]. In this algorithm those web pages with more citations are more important. The advantage of the PageRank is that it does not only depend on the count of referrals, but also considers the importance of the cited web page. The PageRank of a page is calculated as below:

$$PR(u) = (1 - d) + d * \sum_{v \in B(u)} PR(v)/N_v \quad (1)$$

Where u represents a web page, B(u) is the set of pages that point to u. PR(u) and PR(v) are rank scores of page u and v, respectively. N_v denotes the number of outgoing links of page v. d is the damping factor that is set to a value between 0 and 1. It is usually set to 0.85 for the web graph. d can be thought of as the probability of users' following the links and could regard (1 − d) as the page rank distribution from non-directly linked pages.

2.2 The Ant Colony Algorithm
The ant colony algorithm was first represented by Dorigo and et al [5], as a multi-agent solution for optimization issues. The ant colony algorithm was inspired by the research and observations from the ant colonies. A moving ant, leaves a chemical agent on its path known as Pheromone, and therefore marks its path via the smell of this material. While the ant moves randomly and alone, when facing a path with higher amounts of pheromone, it is more likely to choose that path, and strengthen it even more via the pheromone it adds itself. The pheromone evaporates over time; therefore less pheromone would be accumulated to the less used paths. Updating the pheromone is based on this formula:

$$\tau_{ij}(t+1) = (1-\rho).\tau_{ij}(t) + \Delta\tau_{ij}(t) \quad (2)$$

τ_{ij} is the amount of pheromone between nodes i and j. ρ is the evaporation rate ($0 \leq \rho \leq 1$). $\Delta\tau_{ij}$ is the amount of pheromone left by the ant k on the path it met.

3. RELATED WORKS

Xing and Ghorbani represented the Weighted PageRank algorithm, which is an extended version of the PageRank. The Weighted PageRank takes into account the importance of both the in links and the out links of the pages and distributes rank scores based on the popularity of the pages. The Weighted PageRank is able to identify a larger number of relevant pages to a given query compared to the standard PageRank [6].

Scarselli and et al used the neural network model graph to calculate the PageRank amounts for the web pages. The neural network graph could learn the ranking function via examples, and is capable of generalizing over unseen data [7].

Peng and et al first off, analyze the traditional PageRank algorithm of the search engine deeply. Afterwards, According to its topic drift and putting emphasis on old web pages, suggest an improved PageRank algorithm which is based on the text content analysis and the time factor [8].

Dinkar and Kumar have considered the time factor for the ranking of each web page. The rank of each page is calculated based on the unit per time page ranking algorithm [4].

Khodadadian and et al, by using the reinforcement learning concept, proposed RL_Rank algorithm which is a novel connection based algorithm for ranking web pages. Also, they use RL_Rank algorithm in a hybrid algorithm and demonstrated that the hybrid algorithms are effective and have satisfying results [9].

Keong and Anthony proposed the behavior of Markov chain involved in a random surfer model from the original PageRank. This model could lead to a more predictable time for computing PageRank [10].

Chong proposed a new type of algorithm of page ranking by combining classified tree with static algorithm of page ranking-PageRank, which enables the classified tree to be constructed according to the large number of users' similar searching results, and can obviously reduce the problem of Theme-Drift, caused by using PageRank only, and problem of outdated web pages. It improves the searching efficiency without reducing the searching speed, which provides the users with the abundant expanded information relevant to searching content [11].

Kumar and et al proposed the PageRank algorithm based on Visits of Links (VOL) for search engines, which takes the number of visits of inbound links of web pages into account [12].

Tyagi and Sharma proposed the Weighted PageRank algorithm based on Visits of Links (VOL), for the search engines [13].

4. THE PROPOSED METHOD

The proposed method is an extended version of the PageRank algorithm, which the stages will be explained hereafter.

In this study the log file of NASA web server was utilized. Since there are unprocessed data in the log files, they must be subjected to some preprocessing in order to be used in web mining, which in this article, data cleansing, distinguishing the users from each other and identifying sessions per user is considered [14].

After the preprocessing operations and determining the sessions per each user, the degree of each user's interest of each page is calculated via [15]. Afterwards with inspiration from the ant colony algorithm [5]: each user is considered as an ant and each user's interest to web pages is assumed to be the pheromone that the ants left from themselves on the path and the web pages as the path that the ants are going to pass. After completing the tour from all ants, the amount of pheromone on each page is used to calculate the rank of pages. The extended version of the PageRank algorithm is as follows:

$$PR(u) = (1-d) + d * ((\sum_{v \in B(u)} PR(v)/N_v) + P_u) \quad (3)$$

In the equation above, P_u is the amount of pheromone on page u, the rest of the variables are similar to the standard PageRank equation.

5. EVALUATION AND IMPLEMENTATION OF THE PROPOSED METHOD

In this section, the details of implementation of the suggested method are explained. To implement the components of the suggested system, MATLAB and Microsoft SQL Server 2008 software were used.

Figure (1), indicates the page rankings using the suggested method and the PageRank algorithm. It also indicates the real data, which represent the accumulated interest of the users in each page. The ranking method was conducted on 332 pages. In the extended PageRank algorithm, the pheromone's evaporation coefficient was considered as 0.01 ($\rho = 0.01$), because firstly, the page ranking would be much closer to the real data, and secondly, more distinguished ranks would be produced.

If several pages had the same rank, none is better than the others, and one of them is randomly suggested. Therefore, the more distinguished the ranks, the priority of the pages would be determined better. In the suggested method, the count of distinguished ranks is more than the PageRank method. In the suggested method, 304 distinguished ranks were provided for 332 pages, while this count was 84 distinguished ranks for 332 pages, via the PageRank algorithm.

Figure (2), indicates the error graph for both the suggested and PageRank methods. Calculating the error for each page is achieved by subtracting the real data and ranking pages of each method, and afterwards accumulating all these error to achieve the total error. The error is calculated via this equation:

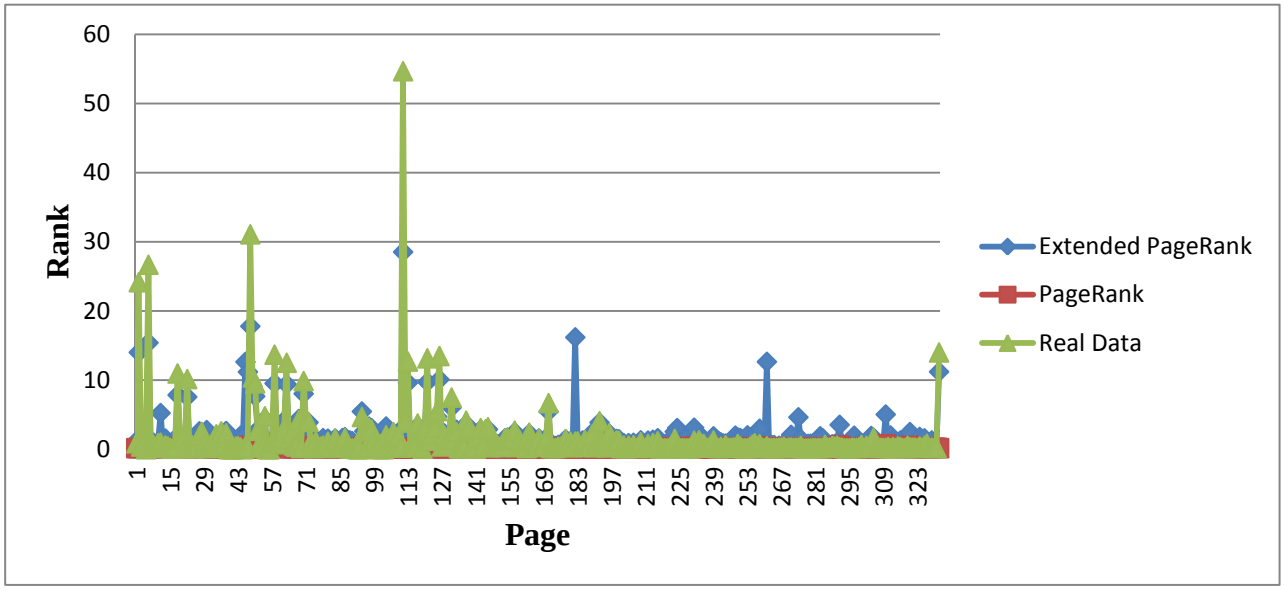

Fig 1: Comparison of the ranking by the suggested method and the PageRank method

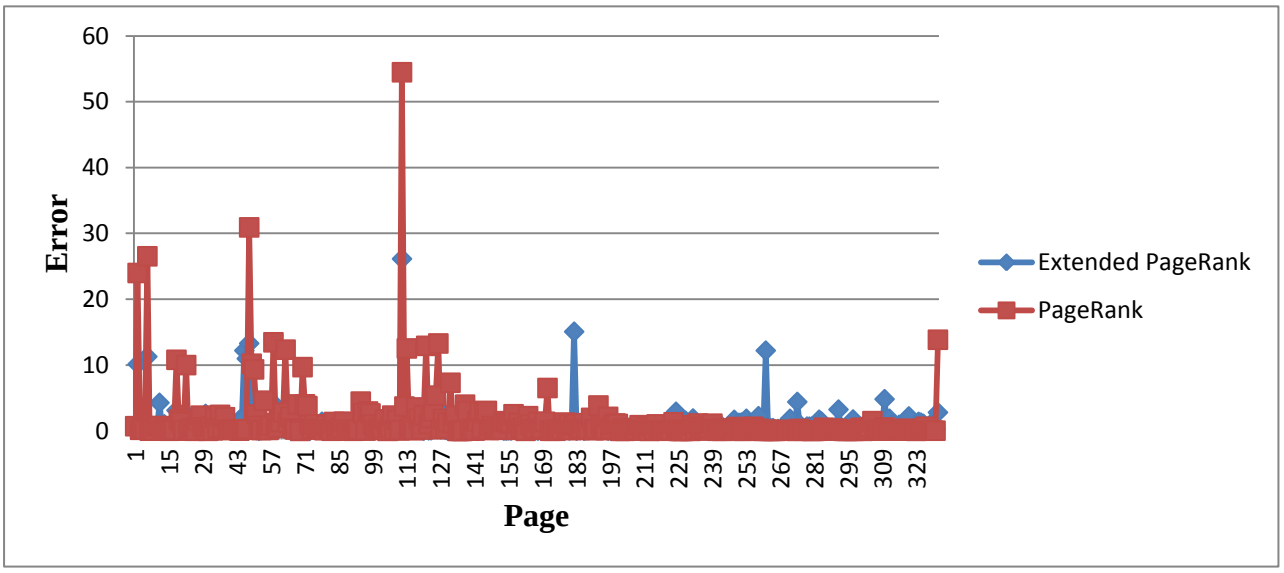

Fig 2: Comparison of the error by the suggested method and the PageRank method

$$\text{Error} = \sum_{i=1}^{332} |x_i - y_i| \qquad (4)$$

x_i Represents the real data, y_i is replaced by ranking pages. The error calculated for the suggested method was 320.82, while it was at 475.94 for the PageRank method.

6. CONCLUSION

The present information on the web is growing over a decentralized process, and this process leads to the production of a huge volume of connected pages, which lacks any kind of logical organization. Therefore analyzing the real interests of each user, and ranking pages, are of utmost importance.

In this article, an extended version of the "PageRank" algorithm was presented in which, users interest to web pages were utilized along with inspiration from the ant colony algorithm to update the users interests.

Simulation results indicate that the suggested method produces more distinguished ranks in comparison with PageRank algorithm, also in this suggested method the produced ranks are closer to the real data and the errors made are much less compared to the PageRank method.

7. REFERENCES

[1] Mustapasa, O., Karahoca, D., Karahoca, A., Yucel, A., Uzunboylu, H. 2010. Implementation of semantic web mining on e-learning. in: proc. of Social and Behavioral Sciences, vol.2, Issue 2, pp. 5820-5823.

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

[3] Brin, S., Page, L. 1998. The Anatomy of a Large-Scale Hypertextual Web Search Engine. Proceedings of the 7th International World Wide Web Conference, pp.107-117.

[4] Dinkar, S.K., Kumar, H. 2012. Interaction Information Retrieval and Improved Page Rank Algorithm Based on Aceess Duration of Page. International Journal of Engineering Research & Technology (IJERT), vol.1, pp.1-5.

[5] Dorigo, M., Maniezzo, V., Colorni, A. 1996. Ant System:Optimization by a Colony of Cooperating Agent. IEEE, vol.26, pp. 29-41.

[6] Xing, W., Ghorbani, A. 2004. Weighted PageRank Algorithm. Proceedings of the Second Annual Conference on Communication Networks and Services Research (CNSR'04), IEEE, pp. 305- 314.

[7] Scarselli, F., Liang Yong, S., Gori, M., Hagenbuchner, M., Tsoi, A.C., Maggini, M. 2005. Graph Neural Networks for Ranking Web Pages. International Conference on Web Intelligence. Proceedings. The 2005 IEEE/WIC/ACM, pp.666- 672.

[8] Peng, Z., Xiu, X., Ming, Z. 2011. An Efficient Improved Strategy for the PageRank Algorithm. International Conference on Management and Service Science (MASS), IEEE, pp. 1-4.

[9] Khodadadian, E., Ghasemzadeh, M., Derhami, V., Mirsoleimani, S., A. 2012. A Novel Ranking Algorithm Based on Reinforcement Learning. The 16th CSI International Symposium on Artificial Intelligence and Signal Processing (AISP 2012), IEEE, pp. 546-551.

[10] Keong, B.V., Anthony, P. 2011. PageRank: A Modified Random Surfer Model. International Conference on IT in Asia (CITA), IEEE, pp. 1-6.

[11] Chong, T. 2010. A Kind of Algorithm For Page Ranking Based on Classified Tree In Search Engine. International Conference on Computer Application and System Modeling (ICCASM), IEEE, pp. 538-541.

[12] Kumar, G., Duhan, N., Sharma, A.K. 2011. Page Ranking Based on Number of Visits of Links of Web Page. International Conference on Computer & Communication Technology (ICCCT)-2011, IEEE, pp. 11-14.

[13] Tyagi, N., Sharma, S. 2012. Weighted PageRank Algorithm Based on Number of Visits of Links of Web Page. International Journal of Soft Computing and Engineering (IJSCE), vol.2, pp. 441- 446.

[14] Cooley, R., Mobasher, R., Srivastava, J. 1999. Data preparation for mining world wide web browsing patterns. Knowledge and Information Systems, vol.1, pp.5–32.

[15] Rashidi, S.F., Harounabadi, A., Abasidezfouli, M.2012. Prediction of users' future requests using neural network. Management science letters (www.growingscience.com), vol.2, pp. 2119-2124.