International Journal of Computer (IJC)
ISSN 2307-4523 (Print & Online)

http://ijcjournal.org/

An Algorithm for Ranking Web Pages Based on Links and Ant Colony Algorithm

Asma Khonji[a*], Dr. Ali Harounabadi[b]

[a]*Department of Computer , Kish International Branch, Islamic Azad University, Kish Island*
[b]*Professor at Islamic Azad University ,Central Tehran Branch*
[a]*Email: asma_khonji@yahoo.com*
[b]*Email: a.harounabadi@gmail.com*

Abstract

With the exponential web growth, techniques of recommender systems and pages ranking algorithms have gained importance over time. Web mining, which is considered as a subset of data mining, is being emphasized in three categories: web mining based on application, web mining based on content and web mining based on structure. Pages ranking algorithms operates mainly based on structure-based web mining. In the current study it has been tried to maximize accuracy of pages ranking with combining function and structure-based techniques. Improvement of Page Rank algorithm is performed using user profiles and important attributes to page ranking algorithm (number of inbound and outbound links to pages). Due to the problem broadness, the use of meta-heuristic algorithms (such as ant colony algorithm) has been highlighted in the current study and the fitness function is set so that the increase of iterations will increase the accuracy of PageRank algorithms. The results of the study explain accuracy of the proposed method compared to other methods.

Keywords: PageRank algorithm; ant colony algorithm; structure-based web mining.

1. Introduction

Nowadays, the Web is a popular and dialogic medium for disseminating information. Traditional web search engines return lots of results for a search which is very time consuming for users to surf. Web mining is used to overcome these problems. Web mining is a data mining technique which automatically extract information from Web documents and is divided into three categories of web structure search, web content search and web using search. Web content search talks about usable and useful information discovery and knowledge extraction from web content deals.

---

* Corresponding author.

Discovery of interesting and useful patterns from web users data usage are in the web usage discovery field. Web structure mining is a process that uses the patterns of the graphs to analyze nodes and structural connections on a website. To support users who move in search engine results, numerous ranking algorithms have been applied on web pages so far. In fact, pages ranking algorithms are a fundamental component of web search engines. They aim to provide a rating for each web page which is a measure that predicts how important and valid the page that users will see is. Ranking algorithms greatly reduce the search space [1]. One of the most important web pages ranking algorithm is PageRank algorithm which is considered by Google to calculate the relative importance of Web pages. In this article we will focus on this algorithm and will provide an extended version of it. In this article, 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 PageRank algorithm

PageRank algorithm which is presented by Page and Brin's for the first time [2] is used by Google as one of the factors to calculate the relative importance of Web pages. The value of PageRank of a web page depends on PageRank values of pages pointing to it and the number of outgoing links from those pages [2]. The performance of this algorithm is that pages with further references are more important. PageRank algorithm is superior as it does not suffice to the number of citations alone to determine the importance but also considers the importance of the referral page [2].

PageRank of a page is calculated as:

$$1) \qquad PR(u) = (1-d) + d * \sum_{v \varepsilon B(u)} PR(V) / N_V$$

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. Nv denotes the number of outgoing links of page v. 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 Ant colony algorithm

Ant algorithm was presented for the first time by Durygu and colleagues as multi-factor solution for optimization problems. In fact, ant colony algorithm is a sub-field of swarm intelligence. Swarm intelligence talks about the collective intelligence resulting from community of smart and simple factors [4]. An ant on the move, leaves a chemical called pheromones on the ground from itself and thus defines the way by odor. When

an ant moves alone and by accident and faces the path that has more pheromones, most likely chooses that pathway and with the pheromones which leaves behind itself in the same direction reinforces that pathway. Pheromone evaporates over time and therefore less pheromones accumulate on the less frequently used routes. If the amount of pheromone of some routes are the same, randomly selects a route. Pheromones update is based on the following equation:

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

τij shows the amount of pheromone between nodes i and j. ρ is the pheromone evaporation ($0 < \rho \leq 1$). Δτij is the amount of pheromone that ant k leaves on the routes that has passed.

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 [3]. Tyagi and Sharma proposed the Weighted PageRank algorithm based on Visits of Links (VOL), for the search engines [4]. 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 [5]. Peng and and his colleagues 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 [6]. Scarselli and and his colleagues 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]. Sara Setayesh, Ali Haroonabadi and Amir Masoud Rahmani present a developed version of PageRank algorithm, in which the interest level of web page users and ant's colony algorithm are used [1].

4. The proposed method

The suggested method in this article is based upon the features of web search algorithm founded on structure and usage. First we get the users interest in web pages, then inspiring from ants colony algorithm and using the suggested version of ranking algorithm, we obtain each page's rank.

The structure of the suggested method is illustrated in Figure 1. In this research we have used NASA's web server recording file. There is raw data in recording files, in order to use them, they should be preprocessed. In this article we use data refinement, user distinguishing and recognizing each user's settlement. After the preprocessing and distinguishing each user's settlements, we determine each user's interest level to vectorize the settlements using the following equation:

$$4)\ \text{Interest(page)} = \frac{2*\text{Frequency(page)}*\text{Duration(gepa)}}{\text{Frequency(page)}+\text{Duration(Page)}}$$

After determining the average of each user's interest level in each page in its settlements, inspiring from ants colony algorithm we act as follows:

- We consider each user as an ant. We take the users interests in pages as the ants' pheromones.
- We consider pages, the routs which ants pass.

After all ants tours are finished, we use the amount of pheromones placed on each page in the developed version of PageRank, to calculate the page ranks. We add a new parameter to the suggested algorithm, named the ratio of number of entry and exit links to a page. The higher number of entry and lower number of exit to a page, the more desirable results. We take the pheromones changes, which is the user's interest, equal to the weight of the pages multiplied by the ratio of entry to exit links. The developed version of the PageRank algorithm is as follows:

$$3) \qquad PR(U) = (1-d) + d * (\sum_{v \epsilon B(u)} PR(V)/N_v) + P_u * N_v / N_u)$$

P_u is the amount of pheromones or user interest, and Nv and Nu are entry and exit links of page, respectively. The ther components are the main PageRank.

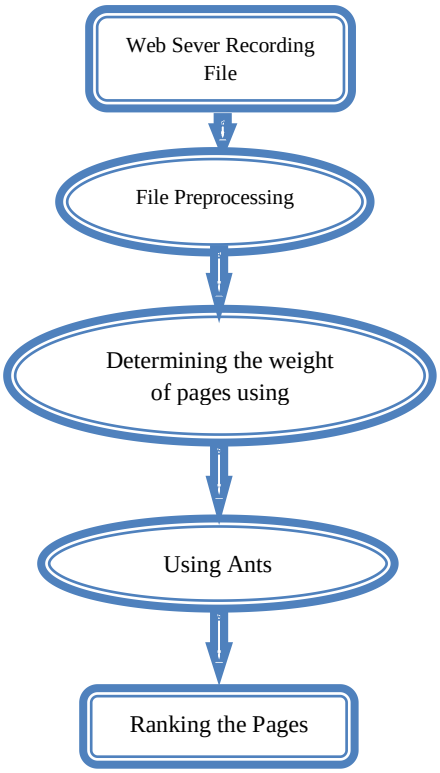

Figure 1: Structure of the suggested algorithm

5. Evaluation and implementation of the proposed method

In this section implementation details of the suggested method is explained. MATLAB and C# programs are used for implementation of the components of the suggested system.

To evaluate the suggested system, we compared the PageRank algorithm to the suggested method and real data which represents the users' interest in each page. We selected the PageRank algorithm as the base method since we suggested a developed version of this algorithm. The results from analysis of the suggested method and PageRank algorithm and real data are shown for 50 pages in the Figure.

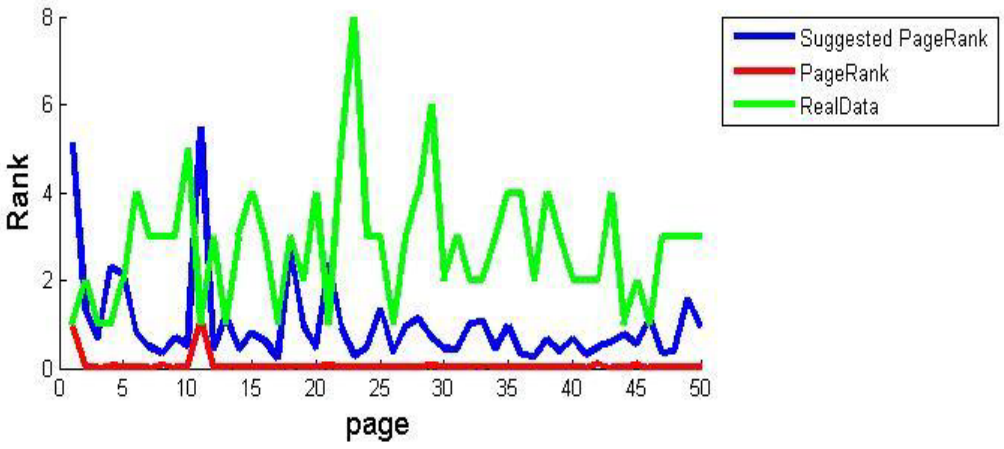

Figure 2: comparison of ranking of 50 pages in the suggested method and PageRank algorithm

In the suggested method, number of individual ranks of the pages are so many, and this is from advantages of the suggested method. In fact, with the higher ratio of the entry to exist links, we see improvement of the algorithm. For 100 pages subjected to ranking, 94 different ranks were generated for the pages. While in the PageRank method, 25 different ranks were generated for 100 pages.

6. Conclusion

In this paper, combination of structural web mining and ant colony algorithm is used to rank web pages. First calculated the frequency of page views and page view duration by different users then inspired by ant algorithm and used updates pheromone part of the algorithm to update out to users' interests. We calculated rank of each page at a time based on previous times and also calculated the ratio of number of incoming links to outgoing as new parameter added to the proposed algorithm. Simulation results show that the proposed method produces more distinguished ranks compared to the PageRank algorithm. When some pages have the same rank, they do not have any preference to each other and one of them will be placed randomly in the results list. Generated rankings in the proposed procedure are closer to real data. The error of our approach is less than main PageRank method.

References

[1] Sara Setayesh , Ali Harounabadi, Amir Masoud Rahmani, 2014, Presentation of an Extended Version of the PageRank Algorithm to Rank Web Pages Inspired by Ant Colony Algorithm, International Journal of Computer Applications (0975 – 8887) Volume 85 – No 17

. [2] Brin, S., Page, L. , 1998. The anatomy of a large-scale hypertextual web search engine, Proceedings of

the 7th International World Wide Web Conference. Briabane, Australia, Apr 14-18, pp.107-117.

[3] 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.

[4] 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.

[5] 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..

[6] 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.

[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.