IDENTIFYING IMPORTANT FEATURES OF USERS TO IMPROVE PAGE RANKING ALGORITHMS

Amir Hossein Eskandari [1] and Ali Haroun Abadi [2]

[1] Department of Computer Engineering, Kish International Branch, Islamic Azad University, Kish, Iran
[2] Department of Computer Engineering, Islamic Azad University, Science and Research Branch, Tehran, Iran

ABSTRACT

Web is a wide, various and dynamic environment in which different users publish their documents. Web-mining is one of data mining applications in which web patterns are explored. Studies on web mining can be categorized into three classes: application mining, content mining and structure mining. Today, internet has found an increasing significance. Search engines are considered as an important tool to respond users' interactions. Among algorithms which is used to find pages desired by users is page rank algorithm which ranks pages based on users' interests. However, as being the most widely used algorithm by search engines including Google, this algorithm has proved its eligibility compared to similar algorithm, but considering growth speed of Internet and increase in using this technology, improving performance of this algorithm is considered as one of the web mining necessities. Current study emphasizes on Ant Colony algorithm and marks most visited links based on higher amount of pheromone. Results of the proposed algorithm indicate high accuracy of this method compared to previous methods. Ant Colony Algorithm as one of the swarm intelligence algorithms inspired by social behavior of ants can be effective in modeling social behavior of web users. In addition, application mining and structure mining techniques can be used simultaneously to improve page ranking performance.

KEYWORDS

web mining, application mining, web page ranking, page rank algorithm, ant colony algorithm

1. INTRODUCTION

Most people use search engines to access information available in the Internet. Search engines are programs which find specific pages matching with users' search. Ranking web pages is the most important factor for search engines [1]. Web search engine is a program which returns a number of web pages in response to search of a user. Currently, Google, Bing and Yahoo are the largest search engines throughout the world [2].

PageRank is a representative link-based ranking method in modern Web information retrieval [3]. Ranking web pages is a technique which assigns a rank to web pages based on different indices and parameters. However, web data has some characteristics which makes web mining difficult. Some features which challenges extracting and mining information and useful knowledge from web include [4]:

- Web information and data are heterogeneous.
- A significant amount of web information are related to each other (links)
- Web information are noisy.
- web is dynamic and its information vary continuously.
- Web offers services
- Web is a virtual society

DOI : 10.5121/ijwest.2018.9303

Search engine optimization (SEO) helps search engines to assign highest rank to a specific website. Therefore, higher rank websites are located at the top of search engine result page in respond to a user search. This technique indicates procedure of improving traffic through increasing site visits. In fact, if a website is located at the top of search engine result page, probability that more users visit that page is higher [2].

Page ranking algorithms are the most important algorithms used by search engines to rank web pages. Ranking algorithm of a search engine is one of the most important elements which determines ability and quality of a search engine [5]. Thus, necessity of conducting this study lies in application of web page ranking algorithms in web search engines.

In the following, this paper presents a developed version of Page Rank algorithm by focusing on Page Rank algorithm. In the proposed algorithm, web page users' interest and Ant Colony algorithm are employed.

2. WEB MINING

Web is a wide, various and dynamic environment in which different users publish their documents. Web-mining is one of data mining applications in which web patterns are explored. Studies on web mining can be categorized into three classes: application mining, content mining and structure mining.

In web mining application, patterns of users' interests in using web data are extracted and analyzed. Application mining procedure consists three phases including [5]:

- Preprocessing data/ data preparation
- Pattern detection
- Pattern analysis

purpose of web content mining is to obtain useful information from text and web contents. But, in web structure mining, purpose is to analyze web structure using graph theory principles. In fact, web is considered as a graph where each node corresponds to a page in web and each edge is a link between two corresponding pages [7].

3. PAGE RANKING

Web mining techniques offer additional information about linkage of different documents through hyper-link. Web can be considered as a labeled graph where its nodes are documents or pages and its edges are hyper-links between pages. Directed graph structure is known as web graph. Among algorithms presented based on link analysis, Page Rank, HITS and Weighted Page Rank are the most important algorithms which are studied in the following [8].

3.1. HITS ALGORITHM

HITS algorithm classifies web pages into two categories; Authority ad Hub. Hubs are pages which operate as source lists and Authoritys are pages which contain important information. This algorithm assumes that a document which indicates other documents more is a good Hub and document which indicates more documents in a good Authority. A page might be a good Hub and a good Authority. Equations (1) and (2) show how weight of Hub(Hp) and Authority (Ap) are calculated in which Hp and Ap is Hub score of a page and Ap is Authority score of a page. I(p) is a set of visits to page p. in addition, B(p) is proportional to sum of Hub weights of pages to which they are linked. Similarly, Hub of a page is proportional to sum of Authority weights of pages to which they are linked [9].

$$H_p = \sum_{p \epsilon l(p)} Aq \tag{1}$$

$$A_p = \sum_{p \epsilon B(p)} Hq \tag{2}$$

3.2. WEIGHTED PAGE RANK ALGORITHM

This algorithm is an improved version of Page Rank. This algorithm assigns a higher rank to most important pages instead of dividing rank of a page among pages originated from that page. Each exiting link receives a value proportional to its importance. Importance is assigned with weight values and include input links and output links. Main formula of the weighted page rank algorithm is the developed version of page rank calculated according to Eq. (3) [10].

$$WPR(n) = (1-d) + d \sum_{m \epsilon B(n)} WPR(m) W^{in}_{(m,n)} W^{out}_{(m,n)} \tag{3}$$

3.3. PAGE RANK ALGORITHM

Page Rank algorithm is used to restore web pages and rank them based on relation of user's search. This algorithm which was first presented in 1998 by Larry Page and Sergey Brin is an independent search method. This method assigns a score to each web document once and uses this score to rank documents without considering a measure regarding user's search. This algorithm obtains rank of each page by assigning weight to a link given to that page. Value of this weight depends on quality of the page in which the link is located. In this case, links of more important pages get higher weights. In order to specify quality of referred pages, rank of the page which is determined recursively and its initial value is arbitrary is used in Page Rank [11].

4. ANTS COLONY ALGORITHM

Ants algorithm was first proposed as a multi-agent solution for optimizing problems [12]. In fact, Ants Colony algorithm is a subset of swarm intelligence. Swarm intelligence studies swarm intelligence caused by community of simple and intelligent agents [12]. Ant Colony Optimization algorithm is implemented in real behavior of ants in nature. Main purpose of this algorithm is to reduce mean mining time of ants to find destination nodes. Ants are not intelligent individually; in practice, they are considered to be blind but socially their behavior is structured. Ants perform finding the shortest path without knowing difficulties along the path. While moving, ants release a chemical substance called pheromone so that other ants realize that an ant has passed this path. Amount of pheromone that an ant leaves is inversely proportional to the distance passed by the ant. Therefore, ants which move along a shorter path, leave higher amount of pheromone in unit length [13].

5. PROPOSED APPROACH

In order to improve Page Rank algorithm in ranking pages, this paper suggests several steps. Figure 1 illustrates a schematic of conceptual model of the proposed method in this research in improving Page Rank algorithm used to rank web pages.

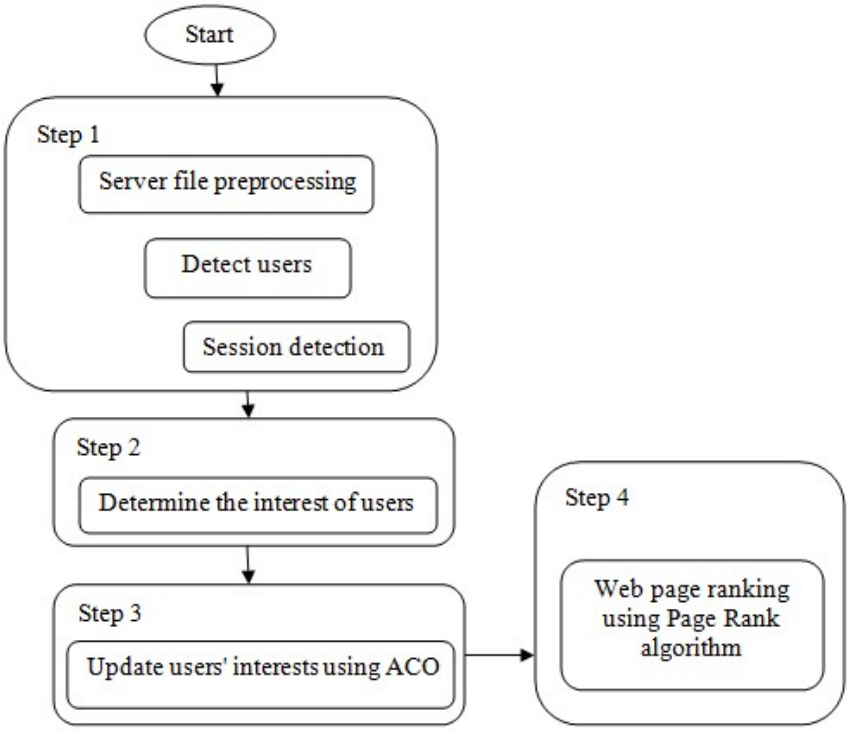

Figure 1. Conceptual model of the proposed method used to improve Page Rank algorithm

Step 1

In this step, server register file is preprocessed to eliminate information which are not used in this study; in addition, users are identified and session is detected.

Requests transmitted by clients to the server are stored in a file called web event record. Requests might be issued either by human or search engine crawlers. Server records are considered as a large information source for extracting knowledge in web mining context. Data stored in server records show that multiple users access a web site [12]. In this step, unnecessary items including activities of search engine crawlers, i.e. GET method accesses are not considered. Identifying users is the most important task in preprocessing step. Simplest method in identifying users is using IP addresses and cookies [14]. In this study, users are identified using IP addresses.

Sessions demonstrate activities of each user. Two general methods have been presented to detect user sessions which include time-based method and scrolling-based method [12]. In this study, time-based method is used to detect user session. In this method, session might be calculated and detected based on duration between entrance and exit of a user in a system or based on the time that each user has spent on each page.

Step 2

In this step, interest of user to each web page is determined using time spent to visit that page and number of visits. It is assumed that P is a set of all accessible pages, therefore, each si indicates a session which is a subset of P. Moreover, si is a member of set S as set of user sessions. Therefore, interest of each user is determined by assigning a weight to each page at each session such that W(pi,si) is weight of page i at session i.

Weights indicate users' interest to pages using page frequency and visit duration. Equation (4) and (5) calculate page frequency and visit duration.

$$Frequency(page) = \frac{Number\ of\ visits(page)}{\sum_{page\ \epsilon visited\ pages}(Number\ of\ visits(page))} \quad (4)$$

$$Duration(page) = \frac{Total\ Duration(page)/Lenght(page)}{Max_{page\ \epsilon visited\ pages}(Total\ \frac{Duration(page)}{Lenght(page)})} \quad (5)$$

Weights show users' interests calculated through combining average weights, frequency and duration which is calculated using Eq. (6). Accordingly, weight of each page increases proportional to increase in both criteria [15].

$$Interest(page) = \frac{2 * Frequency(page) * Duration(page)}{Frequency(page) + Duration(page)} \quad (6)$$

Step 3

Updating users' interests is performed using ants colony algorithm; to this end, each user is considered as an ant and pages visited by users at each session are considered as paths which the ants pass; therefore, interest of each user to each page is interpreted as amount of pheromone left by each ant at its path. In addition, the higher is amount of existing pheromone, pages are visited more and users have more interest and vice versa, i.e. pages with lower number of visits are paths with less amount of pheromone.

In the ants colony algorithm, p_ij^k is the probability that kth ant goes from i to j while j is a member of neighbors of i. ijΓ which is defined as amount of pheromone on edge ij in ants colony algorithm; it is used as number of visits from each page on link ij in this research. ijΠ in the ants colony algorithm is the amount of background information known from path ij; in this research, it is used as information known from features of users which is the information obtained from Eq. (6). Alpha and beta are impact factors. Eq. (7) represents probability function of ants colony algorithm.

$$P_{ij}^{k} = \frac{[\Gamma_{ij}]^{\alpha}[\eta_{ij}]^{\beta}}{\sum_{l\in N_i^k}[\Gamma_{ij}]^{\alpha}[\eta_{ij}]^{\beta}} \quad , \quad if\ j \in N_i^k \quad (7)$$

J is the cost function on edge ij where its value is in matrix D where its main diagonal is zero. Eqs (8) to (10) show calculation of this function.

$$J(l_{ij}) = d_{ij} \quad (8)$$

$$D = [d_{ij}]_{n\times n} \quad , d_{ii} = 0 \quad (9)$$

$$\eta_{ij} = 1/d_{ij} \quad (10)$$

$_{ij}\Pi$ is the background information known from path ij. In this problem, background information is weights $W(p_i,s_i)$ defined as weight of ith page at ith session and users interest. In addition, in order to calculate pheromone evaporation in ants colony algorithm, it is assumed that weight of pages with less visits is reduced gradually.

Step 4

In order to rank web pages using page rank algorithm, it is assumed A is the page which has T1 to Tn input pages. Therefore, Page Rank can be calculated using Eq. (11) for page A.

$$PR(\mathrm{a}) = (1-d) + d[(PR(T_1)/\mathrm{C}(T_1)) + \cdots + (PR(T_n)/\mathrm{C}(T_n))+\mathrm{P}A_a] \quad (11)$$

In Eq. (8), d is impact factor between 0 and 1; it is usually considered to be 0.85. C(A) is the number of links exiting from page A. PA_a is page importance in terms of users' interests obtained using ants colony algorithm.

5.1. IMPLEMENTATION

The proposed method has been evaluated in MATLAB. In order to evaluate the proposed system, real datasets obtained from NASA web server are used. These data indicate interest of users in each page and represent real rank of pages. Other parameters are considered as other papers. Parameters include parameters of ants colony algorithm and impact factor (d) in the Page Rank algorithm where its values are given in Table 1.

Table 1: values of ants colony algorithm and Page Rank algorithm

Value	Parameter	Algorithm
0.05	Evaporation rate	ACO Algorithm
40	Number of ant	
40	Repeat number	
0.85	Impact factor	Page Rank Algorithm

5.2. RESULTS

Results obtained from ACO Rank algorithm (the proposed algorithm) and Page Rank algorithm and real data for 128 pages are shown in Figure 2. Page Rank algorithm is used for comparison because purpose of this research is to improve this algorithm.

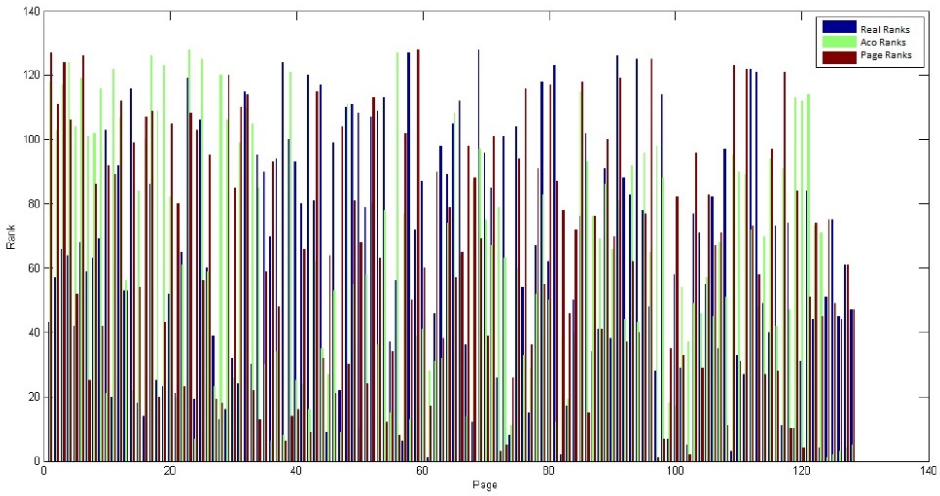

Figure 2. Comparing AcoRank and PageRank ranking

Table 2 shows ranking of 128 web server pages. As can be seen in this table, ranks of the proposed method are closer to real data. Considering this table in AcoRank algorithm for 128 pages to which ranking is applied, 120 different ranks are generated for pages while in PageRank method, 30 different ranks are obtained for 128 pages. If several pages have the same rank, none is superior over others and one of them is proposed randomly. Therefore, the higher is the number of distinct ranks, pages are prioritized better.

Table 2. Results obtained from ranking

Average error	Number of distinct ranks	Algorithm
32	120	ACO Rank
40	30	Page Rank

Average error calculated for AcoRank and PageRank is 32 and 40, respectively which shows 20% improvement in the proposed method.

5.3. EVALUATION AND ANALYSIS OF THE RESULTS

Evaluation of the proposed algorithm and PageRank algorithm shows that the proposed algorithm has higher accuracy compared to PageRank algorithm. The proposed algorithm has higher number of distinct ranks compared to PageRank algorithm. In addition, the proposed algorithm shows ranks more obviously with higher rank difference compared to Page Rank Algorithm. Moreover, the proposed algorithm performs effectively in satisfying users by offering page ranks through considering interest of users. Therefore, more real results are offered to the new users considering total interest at each page.

6. CONCLUSIONS

Since word wide web is being used widely, methods and techniques are required to extract useful and efficient information from large volume of information. Responding to users search in the shortest possible time is one of the objectives of using machine learning techniques. Data mining is a machine learning branch offered in web mining context, content mining, application mining and structure mining; each technique and integration of techniques can be used to extract new information. Page Rank algorithm is one of the algorithms based on link analysis in structure mining which assigns an importance degree to each web page. The proposed approach shows that using swarm based algorithm improves Page Rank performance. Ants colony algorithm as one of the swarm intelligence algorithms inspired by social behavior of ants can be effective in modeling social behavior of web users. Therefore, the current study shows that using application mining and structure mining techniques simultaneously can improve performance of page ranking.

REFERENCES

[1] Seema Rani , Upasana Garg,(2014), “A Ranking Of Web Documents Using Semantic Similarity And Artificial Intelligence Based Search Engine”, International Journal of Science, Engineering and Technology Research (IJSETR), Volume 3, Issue 12, ISSN: 2278 – 7798 ,page 3354-3357.

[2] Nisha , Dr. Paramjeet singh, (July 2014) ,“A Review Paper on SEO based Ranking of Web Documents “,International Journal of Advanced Research in Computer Science and Software Engineering, Volume 4, Issue 7, ISSN: 2277 128X, page1136-1140 .

[3] Hee-Gook Jun, Dong-Hyuk Im, and Hyoung-Joo Kim, (2016),An RDF metadata-based weighted semantic pageRank algorithm, International Journal of Web & Semantic Technology (IJWesT) Vol.7, No.2, pages:11-24.

[4] Pranit B. Mohata,(April 2015) ,"Web Data Mining Techniques and Implementation for Handling Big Data", International Journal of Computer Science and Mobile Computing , ISSN 2320–088X, Vol. 4, Issue. 4, pg.330– 334.

[5] Prerna Rai, Arvind Lal,(2016), " Google PageRank Algorithm: Markov Chain Model and Hidden Markov Model" , International Journal of Computer Applications (0975 – 8887) Volume 138 – No.9, March 2016, pages:9-13.

[6] M. Sathya, Dr. P. Isakki, (2017), " Eclat Algorithm on Web Log Data for Mining the Frequent Link", International Journal of Innovative Research in Computer and Communication Engineering, Vol. 5, Special Issue 1, March 2017, pages:85-92.

[7] Abha Joshi , Avani Jadeja ,(May, 2015) , "Improving Algorithm for Calculation of Page Rank " , The International Journal Of Science & Technoledge (ISSN 2321 – 919X) , , pages 23-25.

[8] Rekha Jain , Dr. G. N. Purohit,(2011), "Page Ranking Algorithms for Web Mining" , International Journal of Computer Applications (0975 – 8887) Volume 13– No.5.

[9] Phyu Thwe,(2013) ," Proposed Approach For Web Page Access Prediction Using Popularity And Similarity Based Page Rank Algorithm ",International Journal of Scientific & Technology, ISSN 2277-8616 ,pages 240-246.

[10] A.M. Sote , S. R. Pande,(2014), "Application of Page Ranking Algorithm in Web Mining", International Conference on Advances in Engineering & Technology , IOSR Journal of Computer Science (IOSR-JCE) , p-ISSN: 2278-8727 ,Pages 47-51.

[11] M. Dorigo and G. Di Caro, (1999),"The ant colony optimization meta-heuristic" In: New Ideas in Optimization, D. Corne, M. Dorigo and F.Glover Eds. London, UK: McGraw Hill, pp. 11-32.

[12] Moghimi. M, Zare, R and Noruzi, Sima,(2016), A hybrid method for preprocessing a web server record file, third International Conference on Applied Research in Computer Science and Information Technology,

[13] M. Dorigo and G. Di Caro, (1999),"The ant colony optimization meta-heuristic" In: New Ideas in Optimization, D. Corne, M. Dorigo and F.Glover Eds. London, UK: McGraw Hill, pp. 11-32.

[14] K. Etminani and M. Akbarzadeh-T and N. Raeeji Yanehsari,(2009), "Web Usage Mining: users', navigational patterns extraction from web logs," IFSA-EUSFLAT, pp. 396-401.

[15] H.Hannah Inbarani and K. Thangavel and A. Pethalakshmi, (2007), "Rough set based Feature Selection for Web Usage Mining," Conference on Computational Intelligence and Multimedia Applications, Vol 1, pp. 33-38.