Management Science Letters 3 (2013) 1707–1712

Contents lists available at GrowingScience

Management Science Letters

homepage: www.GrowingScience.com/msl

Users' recognition in web using web mining techniques

Hamed Ghazanfaripoor[a], Ali Harounabadi[b*] and Amir Sabaghmolahoseini[a]

[a]*Department of computer engineering, Science and Research Branch, Islamic Azad University, Kerman- Iran*
[b]*Department of computer engineering, Islamic Azad University, Central Tehran Branch, Iran*

C H R O N I C L E

Article history:
Received February 16, 2013
Received in revised format
6 May 2013
Accepted May 7 2013
Available online
May 8 2013

Keywords:
Web usage mining
Web personalizing
Web structure and Petri net

A B S T R A C T

The rapid growth of the web and the lack of structure or an integrated schema create various issues to access the information for users. All users' access on web information are saved in the related server log files. The circumstance of using these files is implemented as a resource for finding some patterns of user's behavior. Web mining is a subset of data mining and it means the mining of the related data from WWW, which is categorized into three parts including web content mining, web structure mining and web usage mining, based on the part of data, which is mined. It seems necessary to have a technique, which is capable of learning the users' interests and based on the interests, which could filter the unrelated interests automatically or it could offer the related information to the user in reasonable amount of time. The web usage mining makes a profile from users to recognize them and it has direct relationship to web personalizing. The primary objective of personalizing systems is to prepare the thing, which is required by users, without asking them explicitly. In the other way, formal models prepare the possibility of system's behavior modeling. The Petri and queue nets as some samples of these models can analyze the user's behavior in web. The primary objective of this paper is to present a colored Petri net to model the user's interactions for offering a list of pages recommendation to them in web. Estimating the user's behavior is implemented in some cases like offering the proper pages to continue the browse in web, ecommerce and targeted advertising. The preliminary results indicate that the proposed method is able to improve the accuracy criterion 8.3% rather static method.

1. Introduction

During the past few years, the rapid growth of web has increased existing data, significantly. These information should be available for the users to gain their requests among the numerous pages and it creates various problems for the users, which leads them to have web recommendation system (WRS). WRS estimates the user's requirements and recommends them. A WRS is a web-based interactive software agent, which attempts to estimate the priorities of the users notifying the user's data. This task simplifies and personalizes the user online experiences by using the recommendation

*Corresponding author.
E-mail addresses: a.harounabadi@gmail.com (A. Harounabadi)

doi: 10.5267/j.msl.2013.05.014

lists from some suggested items. The offered item can be artifacts like books, films, music and online resources. A WRS is a combination of two modules including online as well as offline modules. Offline module preprocesses the data for producing the user's models. Online module uses the user's model for recognition the user's objectives, estimates the list of recommendation and updates it. WRS is divided in to two parts in web including WRS based on collaborative filtering as well as WRS based on web usage. On the other hand, we can offer a high abstract by implementing the model. A model is a simple description of a system, which is associated with especial characteristics of it. Formal models act based on mathematic theories. These models often support the non-functional requirements. Petri nets are implemented to determine the important information about the structure and the behavior of modeled system. The features and concepts of Petri net introduce it as a simple and strong method for describing and analyzing the information flow and control the systems, which work with non- synchronization activities. Also Petri net includes a strong tool to describe and to study the information process systems specially the systems, which works with discrete, concurrent, distributed, parallel and uncertain events (Heylighen & Bollen, 2002). In this study, first, the system finds the user's interests by considering the user's behavior and these interests are registered in a profile for the user. WRS recommends the next pages to the user by using the profile. This subject has various usages like recommending the goods, which are interested by the user in ecommerce, recommending the proper pages in search engines and targeted advertising on web. In the second part of the paper, we review some related literature. The third part considers the previous related works. The fourth part presents the proposed method by using the user's profile. The fifth part discusses the implementation of the proposed method and compares it with others works in a case study frame. Finally, the sixth part is allocated to conclusion and future works.

2. Background

2.1. Web mining

Generally, data mining, also called data or knowledge discovery, is the process of analyzing data in terms of various perspectives and summarizes it into useful information, which could be applied to increase revenue, cuts costs, or both. Web mining is a subset of data mining techniques divided into three parts including web content mining, web structure mining and web usage mining. Data and web texts are in web content mining. Discovering and reviewing a model based on links named structure mining. Web usage mining is the process of extracting useful information from server logs i.e. user's history. Web usage mining contains three steps of data preprocessing, templates discovery and analysis of templates.

2.2. Colored Petri Nets (CPN)

Petri nets and queue systems are considered as two strong formal models, which support more usages because of supporting the concurrency concepts. These nets are implemented to model the system's behavior. Petri nets describe the system structure with place and transition concepts. Tokens demonstrate the system behavior too and literally there are numerous versions of Petri nets considered where each has its own especial usage. The colored Petri nets explain the classification concept to the tokens by allocating various colors to the tokens. The formal definition of a colored Petri net includes nine tuples (Jensen, 1994) as follows,

$CPN = (\sum, P, T, A, N, C, G, E, I)$

where

- $\sum$ is a limited and non-empty set that is called color set,
- P is a limited set of places,
- T is a limited set of transitions,

- A is a limited set of arcs that: $p \cap T = p \cap A = T \cap A = \phi$,
- N is a Node function that is defined from A to P×T U T×P,
- C is a color function that is described from P to $\sum$,
- G is a guard function,
- E is an arc expression,
- I is an initialization function.

3. Related work

Presently, there are various works accomplished in web mining. Selvan et al. (2012) performed a comprehensive review on some related works. Some algorithm used by Google for ranking the web pages. Page ranking is achieved from a mathematical algorithm based on graph where the graph of web pages is represented as nodes. The value of the rank indicates the relative importance of a special page. The rank of a page is associated with the number of paper links. The page linked to several pages with high page rank is received the high page rank itself. If there were not any links to a web page, there would not any support for this page. Page rank is a probability distribution where a person visits a page, shows his random click probability on the links of that page. We can use page ranking for a set of documents with various sizes and a probability is explained by a numerical value between zero and one.

Forsati and Meybodi (2009) considered two elements of "duration of page observation" and "page frequency" as criteria for weighting pages. In one session, the user may visit a page for various reasons. If the number of these refers is high, that page is more important compared with other pages in that session (Breeding, 2005). Also in a session, in comparison between two observed pages with the same number, the page with the lower links is more important because, this page has the lower potential observation probability. The duration of observation a page by a user represents the relative importance of that page for the user because if, a page is not attractive for him, he/she rejects it and refers to another one. If a user likes the page, he/she spends noticeable time in observing it (Mobasher et al., 2001). Of course, we should consider whether the size of web page is small, the time of observation is decreased proportion to the size. Therefore, in computing page importance, this proportion should be considered:

$$f_p(P) = \frac{\text{visit(P)}}{\sum_{Q \in T} \text{visit(Q)}} \times \frac{1}{\text{In degree(P)}} \tag{1}$$

$$d_p(P) = \frac{\frac{\text{duration(p)}}{\text{size(p)}}}{\max_{Q \in T}\left(\frac{\text{duration(p)}}{\text{size(p)}}\right)}, \tag{2}$$

where *Fp(P)* is page frequency and *Q* is a session from a set of sessions. The relative importance of page is computed by combination of two criteria. The harmonic average of these two criteria is considered for weight and the importance of "page frequency" and "duration of page observation" are considered the same (α=2).

$$W(p) = \frac{\alpha \times f_p(p) \times d_p(P)}{f_p(P) + d_p(P)}. \tag{3}$$

4. The proposed method

This part, like other web counselor systems, consists of off-line phase, which ranks web pages, ordered based on content, based on the number of the surveys on the previous users. Then, in the on-line phase, the degree of user's interest in any class of pages is determined by following the current session of user and noticing the time (s) user allocates to each page symmetrical the size of page.

Now, according to the degree of user's interest, some pages of the relative group would be proposed to user. The rationale behind selecting these criteria, including user behavior of the previous user, the page size, and the time allocated to surveying each page, are as follows:

- Following up the users' surveys leads to information employed in page assessment. Among these cases, we may consider the number of surveys. The pages of web, which have been frequently surveyed by pervious users receive higher value.

- When a user allocates extended time to surveying the page, possible (s) he/she is interested in that page and if a page is not attractive, the user quickly jumps to another page. It should be considered that a quick jump to another page is probably the cause of short length of the page.

Thus, the user interest in a page can be evaluated from the time spent to survey normalized by the length of the page. In addition, the user may leave the page during surveying the page and performs the other tasks. In this case, such extended pause on the page is not a sign of the page value. Therefore, in order to prevent such states, a threshold (20 minutes) is recognized.

To recognize the importance of the noticed pages, higher amounts are devoted to the most surveyed pages in ranking. The target of this research is to specify the user interest to counsel the pages related to his interest dynamically. We assume that number of web pages related to the database under study is constant during a day. In simulation tool of CPN, all pages are placed in token format in a place. Each token includes of page information including page size, class, and number of surveys. The session of every user is also considered as a token, which contains information elicited from the sequence of the followed pages by user and time allocated to each page.

$$Frequency(page) = \frac{Number\ of\ visits(page)}{\sum_{page \in visited\ pages}(Number\ of\ visits(page))}, \tag{4}$$

$$\text{Duration(page)} = \frac{Total\ Duration(page)/Lenght(page)}{Max_{page \in visited\ pages}\ (Total\ Duration(page)/Lenght(page))}, \tag{5}$$

$$\text{Interest(page)} = \frac{2 * Ferequency(page) * Duration(page)}{Ferequency(page) + Duration(page)}. \tag{6}$$

The above relationships are similar to the mentioned rules in the before works with a difference: The date of page surveying is considered in it and fuzzy approach can be noticed in this part too (Phatak & Mulvaney, 2002). In this scenario, with each step the user take, more information is extracted about his/ her interests and more related pages are counseled to him (her).

- Surveying the first the page by the user among the counseled pages or one of the determined group.

  First scenario: if the user selects one of the counseled pages: the name of group, and size of the page of the token related to page, and also the time spent to survey the page are noticed of the token related to the user session and then based on these agents, the user's interest in this class is computed using Eq. (4) and Eq. (5). Several new pages are replaced various final pages of the proposed list and the list pages are arranged based on the number of surveys.

  Second scenario: If the user selects one of the given classes, the pages related to other classes are filtered and the user can see only the pages related to that certain group. Then (s) he chooses a page related to that group and the process continues similar to the first state.

- The user selects the next page (second page). The name of the class and the page size of the token related to page, and the time spent to survey the page are elicited from the token related to the user session.

If the selected page is of the same pervious class, the value of new page is added to the pervious one and sums of both of these shows the value of the related group to user. In this case, two new pages of the related group from the same higher rank class are again added to set up the counseled pages and two final pages of the proposed list are deleted. This is performed before adding new pages of the classes to the counseled list, pages of the list are checked. If the number of pages counseled from a group is more than a threshold, pages of that group are no longer added to that list. For example, if the length of the counseled list is assumed as 25, if the number of the counseled pages of public class attains to 25 and the user selects the social pages again, the number of social pages of the counseled list does not change. If the chosen page was from another class, the page value in session is computed and based on this value, pages of new class with higher rank in that class are added to the list and the same number of pages is deleted from the end of the counseled list. In order to specify the number of the counseled pages of each class added to the list at each step, the following process is followed:

Let us suppose that the amount of interest of the user in each of the four counseled pages has been computed as 0.11, 0.38, 0.51, and 0, respectively. We want to contain six new pages for him in the counseled list. The computed amounts are multiplied by six respectively and the result is rounded to the integer. So, from the first to the fourth page, one, two, three, and zero page(s) are counseled, respectively.

5. Simulation and results

All pages are placed in the template of tokens in a place. Each of these tokens includes information related to page including size, name of class, and number of surveys. The users' interests are shown as token containing sequence information of the pages followed by the user and the time allocated to each page. There is another place including counseled pages and different classes that the user has access to them while entering the website.

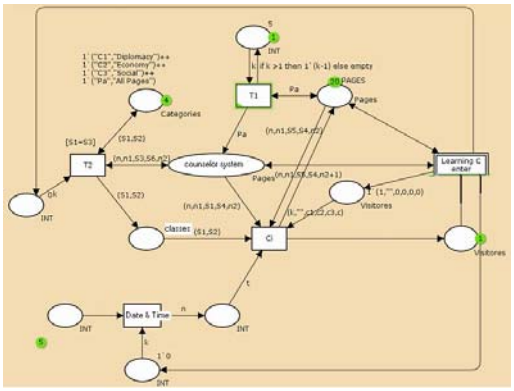

Fig. 1. counselor system

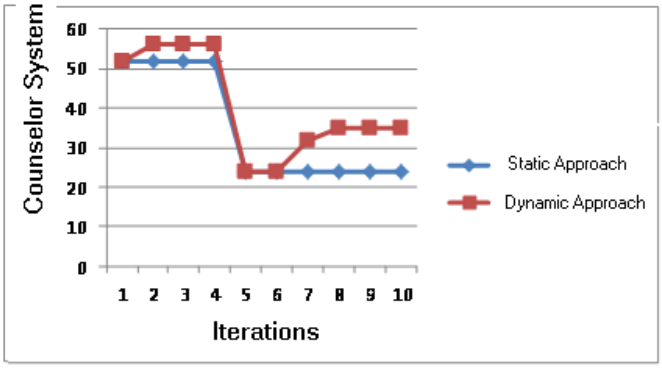

Fig. 2. The comparison of static and dynamic approach

By reference to modeling of the counselor system presented in Fig. 1, while entering the site, the user can choose one of the counseled pages or the available classes (transition T1). If he/she selects a specific class, the pages within the place containing the page are filtered and only the pages of that specific class are available to the user (transition T2). If one of the counseled pages is selected, the features of that page are included in user's interests and its computations are done based on relations 4 and 5 in learning center subnet. Then, the counseled list is updated and transferred to counsel list place through the subnet output and it is given to user for the next choice. In user's next steps, the same cycle is repeated.

6. Results and conclusion

In this research, the number of relevant detected documents is equal to the number of pages in counseled list, which are the same as the content with the user's selected page. “The total number of detected documents” is the number of the counseled factors. Given the fact that, a counselor system evaluates and regains the documents relevant to user's interest among a set of documents, to assessment a counselor system, evaluation criteria related to information retrieval can be used. In information retrieval concepts, the accuracy is computed relevant to two sets of documents. These two sets are “set of detected documents” and “set of documents relevant to the given topic”. The result of division of “the number of relevant retrieved documents” by “the total number of detected documents” is called accuracy. The initial static algorithm used in the studied website, offers a fixed counseled list to all users and elements of this list have been selected just cause of the number of surveys of previous users and it does not notice the user's current session and interests. Fig. 2 shows the results relevant to comparison of precision average of the dynamic counseled algorithm and the aforesaid static algorithm. In this figure, the horizontal pivot shows the order of steps of users' result surveys and vertical pivot represents the average of computed accuracy of the studied users. As we can observe from Fig. 2, the result does not follow a constant tendency and it experiences a fluctuating tendency. This is because of the fact that the manner of users' web surfing different a lot. When the user follows pages of the same class as result, the counseled algorithm increases the value of that class form the users' point of view and the number of the counseled pages of that class increments in the proposed list. When the user goes to a page of another class, the efficiency undergoes considerable fall. One statement is that the pages have been classified in a strict way, i.e. a page is a member of a class or not.

References

Adomavicius, G., & Tuzhilin, A. (2005). Toward the next generation of recommender systems: A survey of the state-of-the-art and possible extensions. *Knowledge and Data Engineering, IEEE Transactions on*, *17*(6), 734-749.

Breeding, M. (2005). Analyzing Web Server Logs to Improve a Site's Usage. The Systems Librarian. *Computers in Libraries*, *25*(9), 26-28.

Forsati, R., & Meybodi, M.R. (2009). Algorithmic based on structure pages and user’s usage information for recommending web pages. *The 2nd Iranian data mining conference*.

Heylighen, F., & Bollen, J. (2002). Hebbian algorithms for a digital library recommendation system. In *Parallel Processing Workshops, 2002. Proceedings. International Conference on* (pp. 439-446). IEEE.

Jensen, K. (1994). *An introduction to the theoretical aspects of coloured petri nets* (pp. 230-272). Springer Berlin Heidelberg.

Mobasher, B., Dai, H., Luo, T., & Nakagawa, M. (2001). Effective personalization based on association rule discovery from web usage data. In *Proceedings of the 3rd international workshop on Web information and data management* (pp. 9-15). ACM.

Phatak, D. S., & Mulvaney, R. (2002). Clustering for personalized mobile web usage. In *Fuzzy Systems, 2002. FUZZ-IEEE'02. Proceedings of the 2002 IEEE International Conference on* (Vol. 1, pp. 705-710). IEEE.

Selvan, M. P., Sekar, A. C., & Dharshini, A. P. (2012). Survey on web page ranking algorithms. *International Journal of Computer Applications*, 41(19), 1-7.