International Journal of Advanced Computer Science and Information Technology (IJACSIT)
Vol. 5, No. 3, 2016, Page: 36-44, ISSN: 2296-1739

www.elvedit.com

Proposing a Mechanism to Improve Web Usage Mining Automatically using Semantic Repository of the Data

Authors

Azizollah Ghazi
Department of Computer Engineering/ Islamic Azad University, Zahedan Branch, Iran
ghazi1c4@gmail.com
Zahedan, Iran

Javad Hosseinkhani
Department of Computer, Zahedan Branch, Islamic Azad University, Zahedan, Iran
jhkhani@gmail.com
Zahedan, Iran

Abstract

Users encounter with some troubles finding the information they need to access easily at the right time because on the one hand they need to examine the relevance of each page with their needs and on the other hand must assess reliability of pages. In recent decades retrieval systems and search engines have been created to fix this problem which index the content of web pages and pages relevant to the user query will be returned. Burdensome of information in current web is a major problem. Personalization systems were provided to deal with this problem which compatible the content and services of a web site based on interests and behavior of people. An essential element in any web personalization system is its user model. The content of web page can be used to create more precise models of the user , but approaches based on key words does not have a deep understanding of website . Nevertheless, manually creating a hierarchy of concepts is time-consuming and costly. On the other hand the public literal meaning resources are suffering from low coverage of specific phrases for domains. In this article we're going to resolve both of these defects. Our main achievement is providing a mechanism to improve the user views automatically in Web site using a comprehensive lexical meaning source. We are using today's largest encyclopedia Wikipedia as a rich source for meanings to improve automated modeling manufacture of user's interests. The proposed architecture includes a number of components that include: pre-primary processing, mining concepts website domain, extracting keywords from web site, creator of keywords vector and key words mapping to concepts. Another important achievement is using the structure of website to limit specific concepts of the domain.

Key Words

Web Mining Applications, User Profile, Source of Lexical Meaning, Automated Modeling

I. INTRODUCTION

In general, data mining can be given into account as web mining on content, structure and web usage. Web mining aims to discover hidden patterns and models in Web resources. Web mining application is specifically aimed at discovering web users' behavior patterns [1]. The discovery of such patterns in vast amounts of data generated by the web servers has important applications. Among them systems that evaluate the effectiveness of web sites to meet user expectations, techniques for dynamic load balancing and optimizing web server to achieve more effective web access to users and applications relevant to restructure and adjust a website based on the user's anticipated needs [2].

Approach based exclusively on the personalization of the web has a major flaw and it is because recommendation process to the user takes place only on the basis of his existing transactional data and hence the items or pages that have been recently added to the website cannot be recommended. On the other hand, however, discovered patterns related to the application of web resources through the web usage mining are helpful in discovering the items communication with each other or to users and determining similarities in user sessions but without using a deeper knowledge of the target web domain, these such patterns gives us little understanding about the reasons why the items or users are put together. We propose a comprehensive view of the personalization process based on Web usage mining. A general framework for this process is shown in Figures 1 and 2 [3]. We will use this framework as a guide.

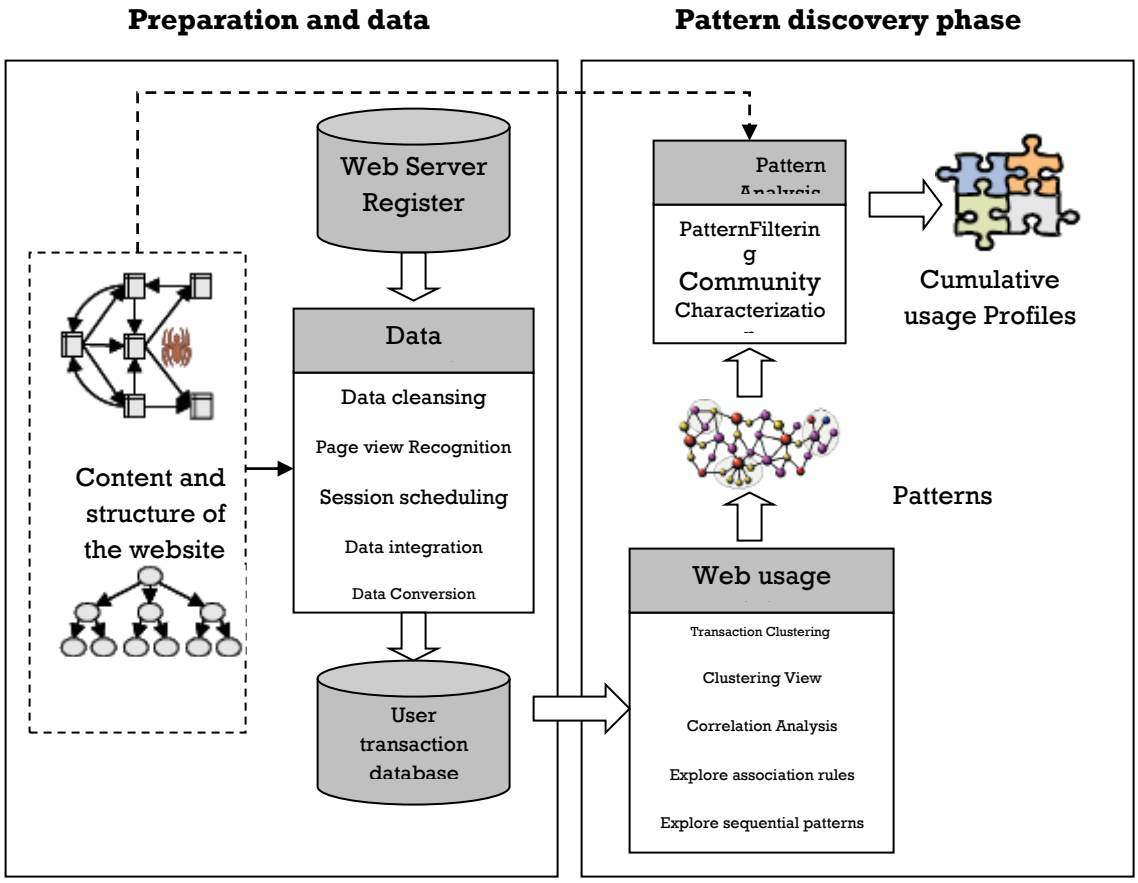

FIGURE 1. THE COMPONENTS OF THE OFF-LINE DATA PREPARATION AND DISCOVERY OF THE PATTERN [3]

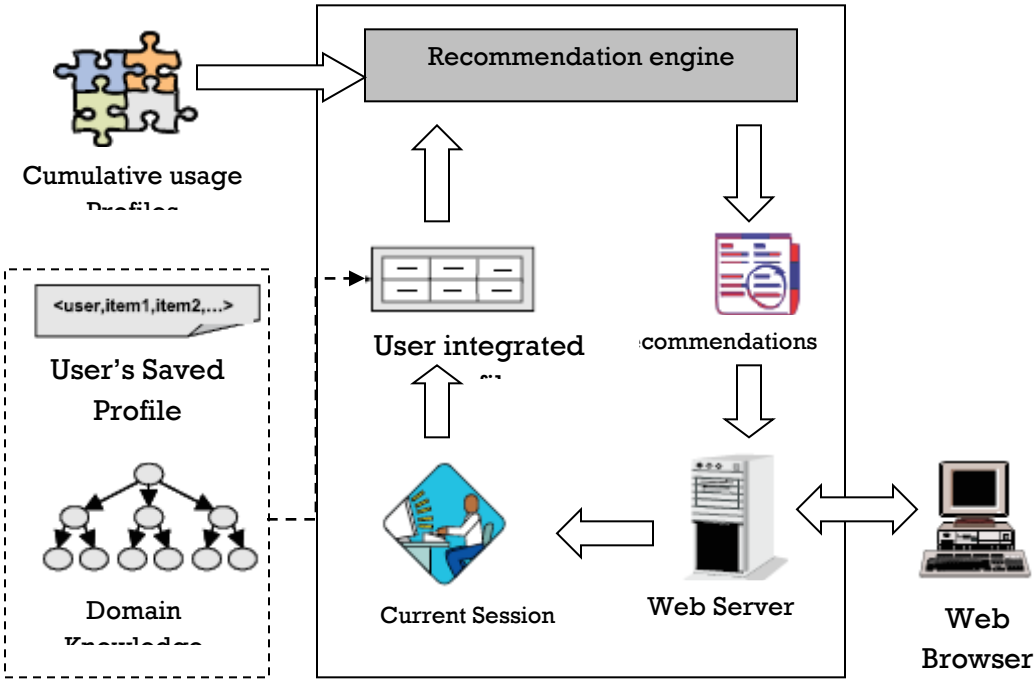

FIGURE 2. ONLINE WEB PERSONALIZATION COMPONENTS [3]

User models can be classified in different dimensions [4] For example, short-term or long-term, express or implied, individually or collectively, and so on. System will not be able to provide the right and relevant resources to the user without having a precise representation of the user, even if the most advanced algorithms would be applied. One of the most important parts of it is the user's display. A major weakness in most existing approaches that use web content to improve user model is that methods usually use the expressions vector to display user's interests and ignore semantic relationships between them while this methodology can be improved by using meaning. The main purpose of this article is to improve the user model in website in such methods that use web content benefit from available meaning in web domain and it is desirable that this process would operate automatically as much as possible.

II. RELATED WORK

There have done researches in various fields such as user modeling and web usage mining in order to use the user behavior in web due to implicitly create a model of his interests. The approaches done based upon these techniques were evaluated in the field of user modeling in Web site. These actions can be divided into two categories. The first category are approaches based on key words which uses vector space model techniques such as term repeat and inverse document frequency (tf-idf) to extract terms from Web documents and create a user model. The second category is semantic approaches that try to use of semantic techniques from the field of information retrieval to improve their models. These approaches can be divided into two categories. The first category is statistical methods that seek hidden semantic relationships

between those terms that occur together. The second category are methods based on hierarchy which try to improve the user model by generating a cognitive or taxonomy vocabulary. A significant portion of repetition in log files arise from HTTP protocol specifications which requires a separate request to the server for every available file, image, audio, video etc. in Web pages. As a result, such data is usually are deleted from log files [5].However, as mentioned earlier, this step is dependent on the domain and deleting such cases from log files could lead to the loss of valuable data. An example might be a website that mainly consists of multimedia content.

Spiders and Crawlers are often identified through user agent field in registration. Most of crawlers are introduced through this field. Tan and Kumar [6] proposed a method for identifying spiders' sessions based on several features such as percentage of requested media files, percentage of incoming requests via HTTP methods and features that indicate the search is in first level. It may also be necessary at this stage to combine several web server log files which requires global synchronization between these servers. Identifying users who have visited the website is one of the most important factors for the success of a personalized website. The easiest way is to assign distinct IP address to distinct users. Several web usage mining tool despite the low degree of precision due to existence of proxy server have selected this way, [7]. Cookies are also a useful way to identify visitors to a website. They store an ID in themselves which is generated by web server for any user visiting the Website, but users often disable or remove them, because cookies are a threat to security and privacy. In addition, if a user connects to the internet using different machines, then the user cannot be properly detected by the cookie. Because of these problems, heuristic techniques are presented. One of these techniques is to use certain internet services such as intend and fingered which provides various information about a client who intend to access to the web server. One of these approach problems is that these services may also be disabled for security reasons. In addition, if users are accessing the website through a proxy server, identifying them by fingered and intends services are impossible because proxy servers hosting the IP address for a large number of users.

Two other revelation methods proposed to overcome the user identification problems [5]. The first method provides an analysis on the web server log file and look for different browsers or if the IP is the same, it looks for different operating systems based on their type and version. These data together show the existence of different users. The second method which the subject is presented in combines the website topology with referring input access. . If the IP address request to a page with a page request to another IP address are the same and there is no direct link between these pages, then perhaps a new user has access to the website. Even these two methods are not without problems. Regardless of their computational cost, there are times that this revelation is not properly operating. For example, when a user uses different browsers at the same time or using different tabs of a browser to open two different pages of a same website which are not linked directly to each other. Another method is also suggested for user identification [7]. A unique identifier that is generated by the web server for each user enters the URL of pages delivered to the user. Instead of storing the IDs in a cookie file, the user is requested to bookmark one of the pages that have that ID as part of the URL. This is a very simple procedure that does not have problems of cookies, however, this technique also has problems of its own.The main

problem is because the user should bookmark a page and have access to the page by using that bookmark its identification method is semi-automatic. Otherwise, the user ID will be useless. Moreover, a situation in which a user has access to a site from different devices continues to remain a problem.

III. PROPOSED METHOD

First the exact definition of the problem is essential to provide proposed method in order to improve the space model vector of user behavior on website then system architecture is presented and its components and performance will be explained. This paper aims to provide a method for improving the space model vector of user behavior keywords on the website which have the following characteristics:

- The model should be implicitly made which means without the direct intervention of user.
- The model should be individually made which means there a specific model for each user.
- The model should be made by circulating behavior of the user in a certain period of time, i.e. two weeks of surfing on the website
- The model creation process is automated as much as possible.
- The model should be containing specific domain concepts along with their value in user visits.

To be mere formal assuming that the user U in period T visited the website and had {s1, s2,...,sm} sessions. The aim is to create a vector of important domain concepts and their weight, namely V = <(c1, w1), (c2, w2), ..., (cn, wn)>, so that the concepts in it are properly chosen (ie have good precision and recall) and as much as possible their weights indicate the importance of user behavior on that time period .The proposed system architecture is shown in Figure 3. As seen in the figure, the system includes several components which each of them described in the following sections.

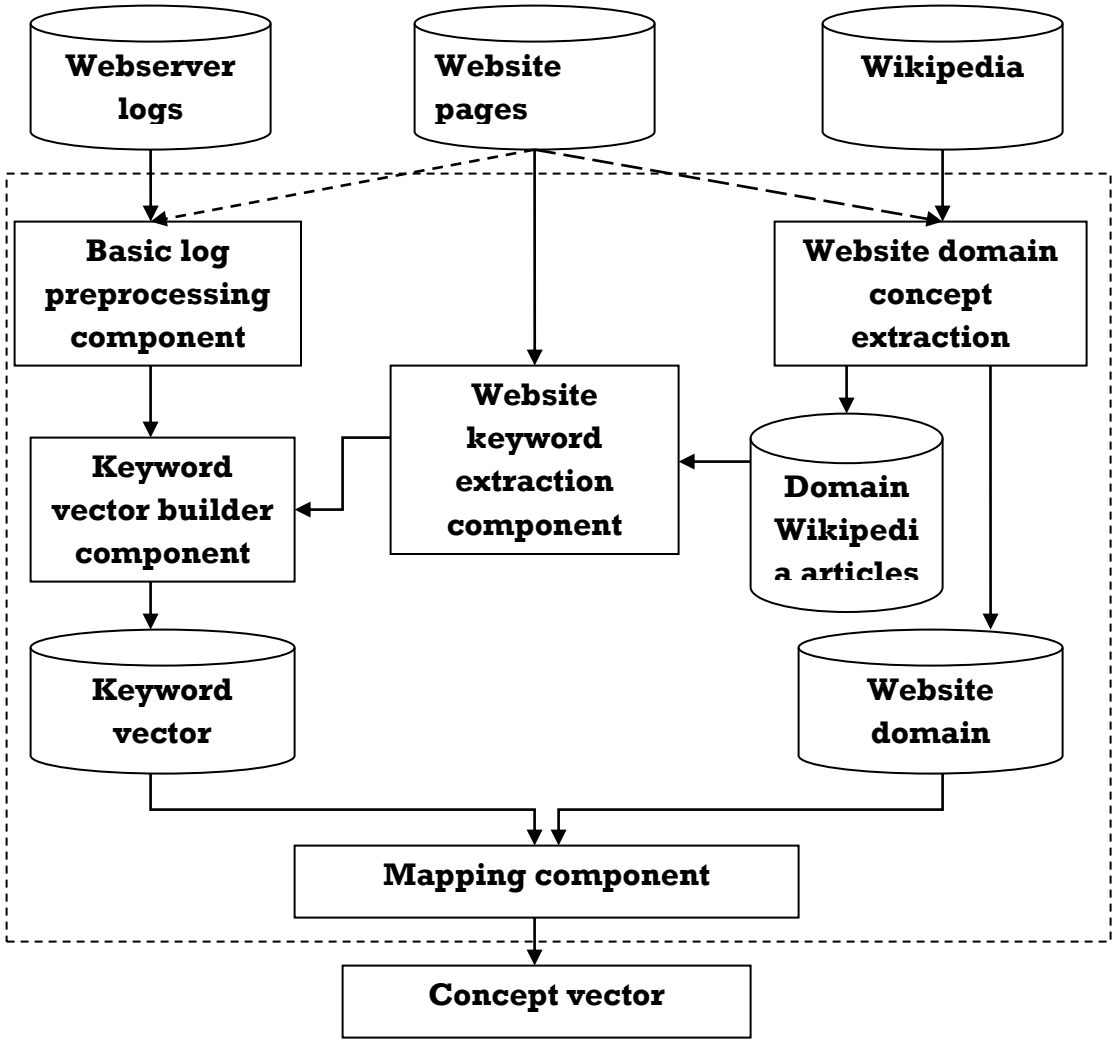

FIGURE 3 OVERALL ARCHITECTURE OF THE PROPOSED SYSTEM

I. Initial pre-processing component registration

This component task is to apply the initial processing of the web server registration such as data wiping, user identification, user session identification etc. This component input is registering web servers and web pages and its output is user sessions.

II. Extracting keywords from web pages component

Keywords in a web page are those words and phrases that are important for that page. These words generally include technical words, named entities, new terms and other keywords relevant to the content on that page. Keyword extraction is automatically identifying keywords on a page.

III. Extraction concepts of web domains component

The task of this component is the extraction of specific domain concepts from Web pages. In this article, any Wikipedia article is intended as a concept. The reason for this is that each article focuses on a particular event or entity and provides important information and related topics in proper structure.

IV. Manufacturer Keywords Vector Component

The task of this component is keyword extraction with an appropriate weight of user sessions. Component output is Keyword extraction for pages like pi is as follows:

$$p_i = < (k_{i1}, w_{i1}),...,(k_{in_i}, w_{in_i}) >$$

Which their weight is their key phrasemes.

V. Mapping Component

The task of this component is creating user model by mapping keywords of visited pages by the user to blocked domain concepts in intended time period. The purpose of this step is to find the most relevant concepts for key words.

IV. Evaluation of Proposed Method

Three tests have been conducted in this article. First test is related to the proposed system with Wikipedia source. Second test have been done by replacing WordNet instead of Wikipedia and a little change in components performance. As mentioned in Chapter 3, WordNet in the absence of manual cognitive words to the website count as a state-of-the-art meaning source. The third test is related to the know method based on key words tf-idfand in fact proposed model (regardless of the source of its meaning) is assessed in trials 1 and 2 and its comparison to the results of trial 3, is the main purpose of assessment in this article. Comparison of results of trials 1 and 2 can be seen in as the comparison of Wikipedia and WordNet as two semantic sources. Table 1 shows the results of three trials. The average precision values and average recall after calculating the precision and recall for every session out of 100 sessions were evaluated and the average was obtained.

TABLE 1. RESULTS OF TRIALS ON 100 RANDOM SESSIONS

	AVERAGE PRECISION	AVERAGE RECALL
TRIAL 1 : WIKIPEDIA	49.6%	51.1%
TRIAL 2 : WORDNET	37.8%	42.5%
TRIAL 3 : KEYWORD METHOD	29.5%	31.3%

Comparison results of the two first trial with the third trial indicates that use of semantic relationships in the web personalization process increases the precision of the user model that will not be considered in keywords approach. The reason of low precision in keywords method is that repetition of a word is not a good measure of its importance and may be repeated so many times but does not make a great deal importance in domain. On the other hand an important concept may appear on a web page in different ways and may also the number of iterations be a few that this has led to the lower recall of key words. Comparison of two first trials also showed the advantage of Wikipedia to WordNet. Its higher precision and recall compared to WordNet is because of good coverage of domain specific concepts and entities. For example, many of experimented concepts on web pages such as VLSI, digital circuit, Software engineering, and so

on does not exist in WordNet and hence it has fewer recalls, while for each of them a full articles in Wikipedia is available. General concepts in WordNet have led to some general concepts to be selected as important concepts and thereby reduce its precision. Another note that can be inferred from the results of the first trial, is that the precision and recall of this method is mostly dependent on the precision and recall of extracted keywords component and seems improving this component will have a significant role in overall system performance.

V. CONCLUSION

User model has a significant importance as one of the most main aspects of web personalization systems. If this model can be made better, its recommendations will be more precise and deeper. The aim of this project is to provide a method to automatically improve the user model in website using available meanings in the pages. And it is desirable to create user model implicitly (i.e. without his direct involvement).For this, a semantic source that has rich lexical which means Wikipedia was used. The proposed method is that the website domain concepts automatically and using page URLs extracted from Wikipedia and then page viewed keywords by the user is obtained using Wikipedia and keywords Vector can be built on it. In calculating the weight of a keyword in this vector, the time spent by the user on the corresponding page to that word has a direct effect. The key words will be mapped by component to the concepts of the first phase. The results of the evaluation indicate that proposed method has better precision and recall to the keyword method as well as the use of WordNet. One of the things that immediately can be made of the results of this article is embedding the obtained model from this method in a recommender system and assessment of its recommendations. In this paper, due to limited data set we were not able to perform this. The future work will be generating comprehensive cognitive vocabulary using Wikipedia and development of semantic web technology. Although work has been done in this area, but the resulting cognitive words are very common and are not suitable for specialized domains

REFERENCES

[1] Birgani, Anoosh Mansouri, Javad Hosseinkhani, Moham, and Mohammad Akhlaghpour. "Proposing an Algorithm in order to Detect Communities in Social Networks using Multi-Objective Evolutionary Algorithm." (2016).

[2] S. S. Anand and B. Mobasher, “Intelligent Techniques for Web Personalization”, LNAI 3169, Springer-Verlag ,2015, 1–37.

[3] B. Mobasher, R. Cooley and J. Srivastava, “Automatic Personalization based on Web Usage Mining”, Communications of the ACM, 2000, vol. 43, 142-151.

[4] P. Achananuparp, H. Han, O. Nasraoui and R. Johnson, “Semantically Enhanced User Modeling”, Proceedings of the 2016 ACM Symposium on Applied Computing (Seoul, Korea, March 11 - 15, 2016).

[5] R. Cooley, B. Mobasher and J. Srivastava, “Grouping Web Page references into transactions for mining World Wide Web browsing patterns”, Technical Report TR 97-021, Department of Computer Science, University of Minnesota, 1997.

[6] P. N. Tan and V. Kumar, “Discovery of Web Robot Sessions Based on their Navigational Patterns”, Data Mining and Knowledge Discovery, 6:1, 2012, 9-35.

[7] D. Pierrakos, G. Paliouras, C. Papatheodorou and C. D. Spyropoulos, “Web Usage Mining as a Tool for Personalization: A Survey”, User Modeling and User-Adapted Interaction, 13: 311-372, 2013.