Bulletin de la Société des Sciences de Liège, Vol. 85, 2016, *p.* 321 - 328

Web mining and Web usage mining techniques

Nasrin JOKAR[1], Ali Reza HONARVAR,[2] Shima AGHAMIRZADEH[3], Khadijeh ESFANDIARI[4]

Department of Electrical and Computer Engineering, Safashahr Branch, Islamic Azad University, Safashahr, Iran

Department of Electrical and Computer Engineering, Safashahr Branch, Islamic Azad University, Safashahr, Iran

Department of Electrical and Computer Engineering, Safashahr Branch, Islamic Azad University, Safashahr, Iran

Department of Electrical and Computer Engineering, Safashahr Branch, Islamic Azad University, Safashahr, Iran

Abstract

Computers promise that be as a repository of knowledge and wisdom, but instead, they sent us large amounts of data, web mining is the process of information discovery and knowledge from the Web data. The data is collected from the server, client, proxy server or database in Web mining. Web mining methods are divided into three categories: web content mining, web structure mining and web usage mining. There are several functional areas including e-commerce web mining, text mining, and management of customer behavior. Web mining research focuses on developing knowledge extraction techniques which are used for data analysis. 3 main methods that are used for data mining in web include: association or association rules, sequential patterns, and clustering requirements. The main objective of the web mining is to collect information about the user navigation patterns. Of course, web mining is faced with various challenges and constraints. And many researches are currently doing research in the field of web mining that aim to solve this problem.

Keywords: data mining, web mining, association rules, clustering and sequential pattern

1. Introduction

We are living in information era, the era in which humans produce and publish data and information more than any other time in the past, in fact, there is more information that we can not analyze it. Therefore, methods and techniques are required to achieve data efficiently, sharing data and data mining and use of this information. Due to the vast amount of information on the web, its managing is nearly impossible with traditional tools and new

[1] Adresse email: nasrin.jokar@gmail.com
[2] Adresse email: alireza_honarvar@shirazu.ac.ir
[3] Adresse email: sh.aghamirzadeh2014@gmail.com
[4] Adresse email: khadijaesfandiari@gmail.com

tools and methods are needed to manage it. Generally, web users using it are faced with the following problems:

Finding relevant information, creating new knowledge using information available on the web, information privatization. There are several reasons for the emergence of Web mining: first of all, the World Wide Web (www) is the greatest and most influential source for data mining and data warehousing [1]. The size of the Web is constantly rising. It has been shown in the source [2] that there are more than 10 million pages with public access on the Web. In addition, nearly 6 terabytes of new information are added to the web each month. The web mining is defined in source [12] as follows:

Web mining is using data mining technique to discover and extract information automatically from documents and Web services.

Web mining aims to discover and retrieve useful and interesting patterns from large data sets, as well as in the classic data mining [3]. Big data act as data sets on web mining. Web data includes information, documents, structure and profile. Web mining is based on two concepts defined, process-based and data-driven. (Based on data routinely and commonly used). In the view of Web mining data web is used to extract knowledge [4]. In general, the use of web mining typically involves several steps: collecting data, selecting the data before processing, knowledge discovery and analysis [9].

1-1 Types of web mining

Web mining can be generally divided into three categories, as seen in Figure 1:

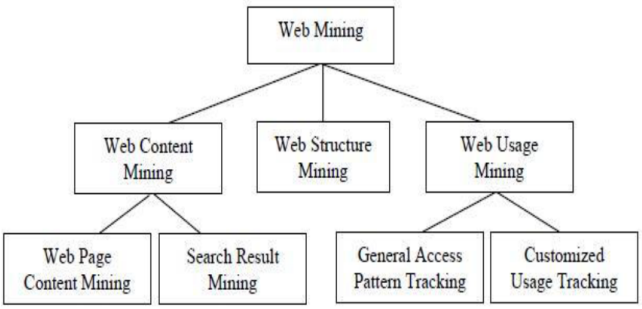


Figure 1: Web mining

- **Web content mining**

Web content mining is the process of extracting useful information from the content of Web documents. The contents of a web document is corresponding to the concepts that that the document sought to transfer it to users. This content can include text, image, video, sound or records such as lists and tables. The text mining has been studied more than other areas.

- **Web structure mining**

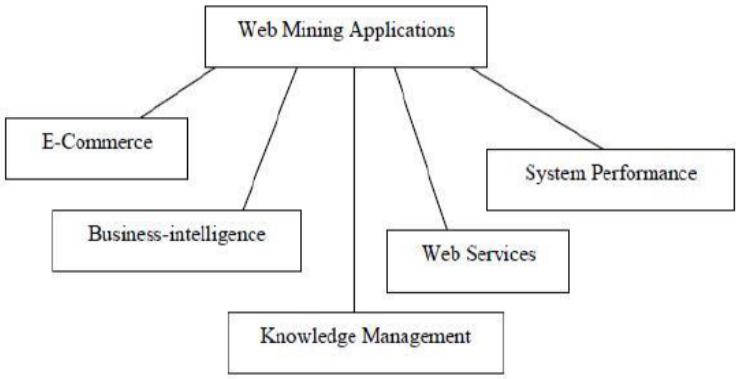
The web can be represented as graph which its nodes and edges are the links between documents. Web structure mining is the process of extracting structural information from the web.

- **Web usage mining**

Web usage mining is the application of data mining techniques to discover patterns using the Web to better understand and meet the needs of the user. This type of web mining explores data relating to the use of web users. It should be noted that there are no clear boundaries between web mining groups. For example, web content mining techniques can use user information in addition to using the documents. It can also be achieved to better results by the combination of above techniques [1].

1.2 Web mining applications

There are many applications in web mining which the most important of them is seen in Figure 2. Most popular web applications in the mining area are e-commerce and customer relationship management (CRM). Most major Web usage mining is in these two areas [4].



Figigure 2: Web mining applications

2. Web usage mining

Business web usage mining uses statistical methods to explore. But researches focus on developing knowledge extraction techniques that are used analyze the web usage mining data. Three main methods that are used to web usage mining include: Association rules, sequential patterns, and clustering. In this section, each of these methods are described in [1].

- **Association rules**

Association rule is the most basic rule of data mining methods which is used more than other methods in the web usage mining. This method enables web site for more efficient organization of content or provide recommendations for effective cross-selling product [10]. These rules are statements in the form X => Y where (X) and (Y) are the set of available

items in a series of transactions. The rule of X => Y states that, transactions that contain items in X, may also include items in Y. association rules in the web usage mining are used to find relationships between pages that frequently appear next to one another in user sessions. For example, a rule can be obtained in the following format:

A.html, B.html => C.html

This rule shows, if user observes A and B pages, most likely will observe page C at the same meeting. A common algorithm to extract association rules is Apriori algorithm. Some criteria are presented to assess the rules extracted from the web usage data .Also, a method is presented by using association rules and fuzzy logic to extract data using the web fuzzy association rules [1].

- **Sequential patterns**

Sequential patterns are used to discover the subsequence in the large volume of sequential data .In web usage mining , sequential patterns are used to find user navigation patterns which appear frequently at meetings. A sequential pattern is often as follows:

70% of users who have first observed the page A.html and then page B.html, have observed page C,html in the same session , too. The sequential patterns may seem to association rules. Actually, algorithms that are used to extract association rules, can also be used to generate sequential patterns [10]. But the sequential patterns are included the time, it means that the sequence of events occurred is defined in sequential patterns. In the above example, pages A, B, C are respectively seen in a user session. But in the example mentioned before, anything has not been considered about the sequence of events. Two types of algorithms are used for mining sequential patterns. The first type of algorithms is based on association rules mining. In fact, many common algorithms of mining sequential patterns have been changed for mining association rules. For example, GSP and AprioriAll are two developed species of Apriori algorithm which are used to extract association rules. But some researchers believe that association rules mining algorithms do not have enough performance in the long sequential patterns mining .For this reason , the second type of sequential patterns mining algorithms have been introduced in which the tree structure and Markov chain are used to represent survey patterns. For example, in one of these algorithms that is called WAP-mine, the tree structure which is called WAP-tree is used to explore access patterns to the web. Evaluation results show that its performance is higher than an algorithm such as GSP.

- **Clustering**

Clustering techniques diagnose groups of similar items among high volumes of data. This is done based on distance functions which measures the degree of similarity between different items. Clustering in web usage mining is used for grouping similar meetings. What is important in this type of search, is contrast of the user group and individual group. Two types of interesting clustering can be found in this area: 1- user clustering, 2- page clustering. Clustering of user records is usually used to analyze the tasks in web mining and web analytics [10]. More knowledge derived from clustering is used to partition the market in e-commerce. Different methods and techniques are used for clustering which include:

- Using the similarity graph and the amount of time spent viewing a page to estimate the similarity of meetings.
- Using genetic algorithms and user feedback.
- Clustering matrix.
- K -means algorithm, which is the most classic clustering method [10].

In other clustering method, first the repetitive patterns are extracted from the user's sessions by using association rules. Then, these patterns used to construct a graph where the nodes are the visited pages. The edges of the graph connect two or more pages, if these pages exist in a pattern extracted so the weight will be assigned to the edges that shows the relationship between the nodes. Then, for clustering, this graph is recursively divided to user behavior groups are detected [1].

3. Web usage mining applications

The main objective of web usage mining is to collect data about the user's navigation patterns. This information can be used to improve the Web sites in the user view. Three main applications of this mining are studied in this section [1].

3.1 The privatization of web content

Web usage mining techniques can be used for personalization of web users. For example, user behavior can be immediately predicted by comparing her current survey patterns with survey patterns extracted from the log files. Recommendation systems which have a real application in this area are, suggest links that direct the user to his favorite pages. Some sites also organize their product catalogues based on predicted interests of specific user and represent them.

3.2 Pre - recovery

The results of web usage mining can be used to improve the performance of Web servers and Web-based applications. Web usage mining can be used for retrieving and caching strategies and thus reduce the response time of Web servers.

3.3 The improvement of Web site design

Usability is one of the most important issues in designing and implementing web sites. The results of web usage mining can help to appropriate design of web sites. Adaptive web sites is an application of this type of mining. Website content and structure are dynamically reorganized based on data derived from user behavior in these sites.

4. Web mining challenges

Web mining is faced with various challenges and constraints. From one perspective, these limitations can be divided into two groups: technical and non-technical. The non-technical restrictions can be included the lack of management support, inadequate fund and lack of required resources such as professional human resources. But there are many technical problems that some of them are mentioned here [1]:

4.1. Incorrect and inaccurate data

To do web mining process successfully, be sure that the collected data are correct and in the proper format. But there are usually many problems in this area. First, the data may be inaccurate. Secondly, the data may be incomplete and unavailable. Thirdly, estimation of assurance about the accuracy of the data is simply not possible.

4.2. The lack of tools

Another limitation of web mining is the lack of appropriate and complete means for it. In this regard, the experts must decide to develop an application of web mining or use available tools.

4.3. Custom tools

Available tools only support one of the web mining types such as classification or clustering. But it is better a web mining tool is able to perform several techniques to allow users to use the appropriate technique according to his requirements. Now, many researches are doing in the field of web mining which aim to solve these problems.

When web mining techniques are used in companies that have some type of personal information and business concerns, help companies to have more detailed profiles of individuals to have a more intelligent marketing, at the same time web mining can be a threat to personal information and privacy (or at least it seems). Privatization web mining made latency and non-proliferation information difficult to people. [11].

5. Related work

This paper aims to focus on discovering websites patterns in Web usage mining from the server files and also comparing memory usage and using time in Priori algorithms and pattern algorithm are common [6].

In this paper, an algorithm is presented which offers a combination of association rules and clustering that shows continuous data set as an output by collecting items purchased by the customer or website repeated information and least backup [4].

In this paper several websites that contain similar categories were recovered by clustering methods and, sorting and Kation test and suggest a method that briefly called compression of tree structure by the new tree structure, identifies commonalities between the

website that is easier to analyze the structure of web pages as well as has proposed a weight to each node [7].

This paper presents an efficient framework for Web personalization based on sequential and non-sequential patterns and use of data that the results show that limited patterns such as continuous sequence pattern is more suitable for doing initial prediction of web. While patterns with less limitations such as frequent item sets or sequential patterns are more effective alternatives in the field of personal websites and recommendation systems [8].

6. Conclusion

Methods and techniques are required for the use of this information and extract new information from them by creating and deploying Web and a significant increase in the volume of data. Web mining as new knowledge and the practical tools has been emerged to help users and webmasters. Web mining is divided into three methods: 1. web usage mining, 2. web structure mining, 3. web content mining which web usage mining techniques include association rules, sequential pattern, clustering, that can be used to develop the site. There are several functional areas in web mining, e-commerce and customer relationship management are the most ones and numerous articles have been written in these fields. Now, many researches are conducting in web mining to destroy these challenges. Although the explaining all methods and applications in this area is not possible , this article can give the reader an overview of web mining and resources and guide him to resources which are his interests. Web mining is faced with challenges such as incorrect and inaccurate data, lack of tools, custom tools, lack of the required resources, management and so on.

References

1- Han, J., Kamber, M., & Pei, J. (2011). Data mining: concepts and techniques. Elsevier.
2- Li, J., & Zaiane, O. (2004). Using distinctive information channels for a mission-based Web recommender system. In Proc. of WebKDD (pp. 22-25).
3- Sudhamathy, G., & Jothi Venkateswaran, C. (2012). Fuzzy temporal clustering approach for e-commerce websites. International Journal of Engineering and Technology, 4(3), 119-132.
4- Rajan, S. D., & Sao, M. N. (2014). An Elegant Draw Near to Improve the Design of an E-commerce Website Using Web Usage Mining and K-Means Clustering.
5 - Siddiqui, A. T., & Aljahdali, S. (2013). Web Mining Techniques in E-Commerce Applications. arXiv preprint arXiv:1311.7388.
6- Kumar, B. S., & Rukmani, K. V. (2010). Implementation of web usage mining using APRIORI and FP growth algorithms. Int. J. of Advanced Networking and Applications, 1(06), 400-404.
7- Yi, L., & Liu, B. (2003, August). Web page cleaning for web mining through feature weighting. In IJCAI (Vol. 3, pp. 43-48).
8- Mobasher, B., Dai, H., Luo, T., & Nakagawa, M. (2002). Using sequential and non-sequential patterns in predictive web usage mining tasks. In Data Mining, 2002. ICDM 2003. Proceedings. 2002 IEEE International Conference on (pp. 669-672). IEEE.

9- Singh, D. S., & Arun, A. P. (2013). Web usage mining: Discovery of mined data patterns and their applications. International Journal of Computer Science and Management Research, 2(5), 2423-2429.
10- Liu, B. (2007). Web data mining: exploring hyperlinks, contents, and usage data. Springer Science & Business Media.
11- http://epbank.ir/port-payments/web-mining
12- Kosala, R., & Blockeel, H. (2000). Web mining research: A survey. ACM Sigkdd Explorations Newsletter, 2(1), 1-15.