A Web Mining Process for Knowledge Discovery of Web usage Patterns

P. Madhura[1]
Head, Dept. of Computer Science,
RIMS, Tirupati, India [1].

M. Padmavathamma[2]
Professor, Head, Dept. of Computer Science,
S.V. University, Tirupati, India[2].

***Abstract*— A number of recent studies are trying to improve the quality and effectiveness of web mining. Web mining is mining of data related to the World Wide Web, this may be the data actually present in Web pages or data related to Web activity. This paper present the frame work for usage pattern discovery in order to knowledge management. Here is proposed a new reference architecture based on reusable building blocks. The system is designed to support a decision maker in making decisions by adopting a clear separation of tasks. It allows the analysis of web information by extracting, selecting, processing and modelling huge amounts of data.**

Keywords- Data mining, web usage mining, information retrieval, pattern extraction.

I. INTRODUCTION

The digital universe known as the world wide web is a very huge place that includes literally billions of web pages and it estimated to continue growth in it. Moreover with this amount of data available online, the WWW is today considered a popular and interactive medium to disseminate information. Web mining is the application of data mining techniques to extract knowledge from web data, including web documents, hyperlinks between documents, usage logs of web sites.

web mining comprises four different steps:

• Resource identification, in which the resources needed for information extraction are identified.

• Pre-processing, in which relevant information is selected from found information sources. This step is directly related to information extraction techniques

• Generalization, in which automatic pattern discovery is made on several web documents. This step uses data mining techniques as well as clustering and classification trees.

• Analysis, in which pattern discovery is validated and interpreted.

These four steps are put together and applied in different ways, according to the type of information source upon which they are made to act.

II. WEB MINING TAXONOMY

Web mining can be broadly divided into three distinct categories. Figure 1 shows the taxonomy.

1. Web content mining
2. Web Structure Mining
3. Web Usage Mining.

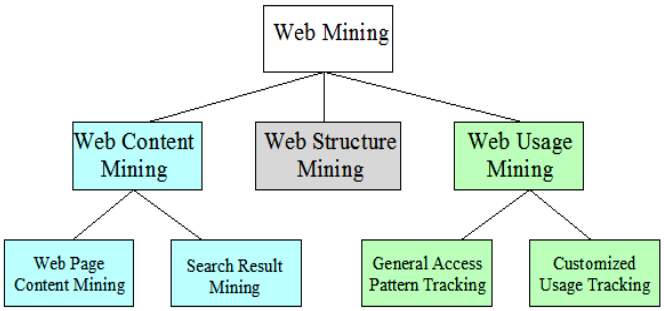

Figure1. Web Mining Taxonomy.

A.Web Content Mining

Web content mining is the process of extracting useful information from the contents of web documents. Content data is the collection of facts a web page is designed to contain. It may consist of text, images, audio, video, lists and tables. Application of text mining to web content has been the most widely researched. Issues addressed in text mining include topic discovery and tracking,Extracting association patterns, clustering of web documents and classification of web pages.

B.Web Structure Mining

Web structure mining can be viewed as creating a model of the Web organization. This can be used to classify Web pages or create similarity measures between documents. The structure of web graph consists of web pages as nodes and hyperlinks as edges connecting related pages. Web structure mining is the process of discovering structure information from the web. This can be further divided into hyper links and document structure.

C.Web Usage Mining

Web usage mining is the application of data mining technique to discover interesting usage patterns from usage data in order to understand user. It performs mining on web usage data, Web logs. A web log is a listing of page reference data. Sometimes it is referred to as click stream data .Usage data captures the identify of origin of web users along with their browsing behaviour at a web site.

III. WEB MINING PROCESS FOR KNOWLEDGE DISCOVERY

The primary objective of a Web Mining process is to discover interesting patterns and rules from data collected within the Web space. In order to adopt generic data mining techniques and algorithms to Web data, these data must be transformed into a suitable form. The idea is to connect specific research domains such as Information Retrieval, Information Extraction, Text Mining and so on, and to put them together in an innovative process of workflow defining several phases and steps moreover they can share common activities, facilitating reuse and standardization.

Generally, Web mining is the application of data mining algorithms and techniques to large Web data repositories Web usage mining refers to the automatic discovery and analysis of generalized patterns which describe user navigation paths (e.g. click streams), collected or generated as a result of user interactions with Web site. constraint-based data mining algorithms applied in Web Usage Mining and developed software tools .One of the most common algorithm applied in Web Usage Mining is the Apriori algorithm. Web user navigation patterns were represented by association rules in. Sequence mining can be also used to mine Web user navigation patterns. The association rules holds information of forward the sequence of requested pages (e.g. if user visits page A, and then page C, it will visit page D). Based on this, users activity can be determined and predictions to the next page can be calculated. The sequence mining algorithms inherited much from association mining algorithms to discovered pattern.

IV.WEB USAGE MINING

The objective of the knowledge discovery from databases(KDD) process is to extract new, interesting and useful knowledge using a variety of data mining methods and techniques such as clustering, association rule mining and sequential pattern discovery. In the case that the data origin is the Web, the process is called Web mining instead of KDD. Web mining concerns a varied range of applications that aims at discovering, evaluating and employing hidden knowledge from Web data sources using Web usage mining (WUM). The web usage mining operates on the data from server access logs, information from users' registration application forms, users profiles and transactions

To put it simply, Web Usage Mining (Figure 2) is the process is the process of extracting interesting patterns in Web access logs.

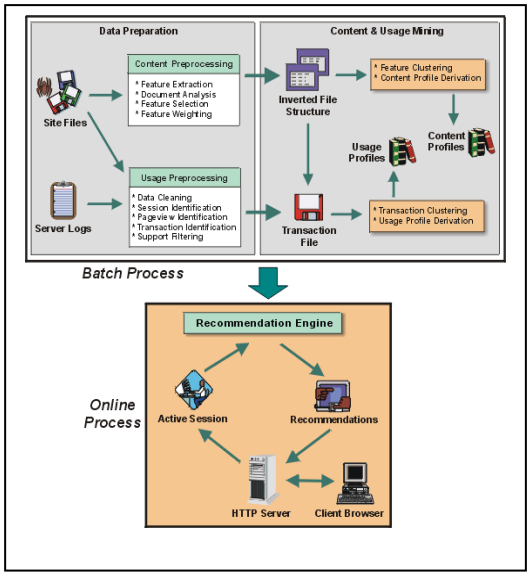

Figure.2 Web Usage Mining.

Web usage mining consists of three main tasks (Figure 2)
1) Preprocessing, contains three separate phases:
(a)cleaning which means that useless entries (e.g. graphic and multimedia objects) have to be removed,
(b) session identification by assign all requests from one user to one unique session
(c) data conversion into the format specific for the software tool.
2) Pattern discovery, means applying the presented algorithm with defined constraints by the user to the data.
3) Pattern analysis is a human domain task and means understanding the results obtained by the algorithm and drawing conclusions.

V. WEB USAGE MINING APPLICATIONS

Web Usage Mining application areas are:
1) Personalization is the ability to tailor content and recommend objects. It implies that PRS system must be able to anticipate users' needs and provide them with objects which they might appreciate based on previous interactions of other users and interactions with current user. Therefore, the personalization task can be viewed as a prediction problem: the system attempts to predict the user's interest in specific content. Recommendation and personalization techniques are classified into three different categories ,rule based filtering, content-based filtering and collaborative filtering.
2) System improvements concern of analyzing collected web data due to provide understanding web traffic behaviour. Such improvements may bring in advanced load balancing, data distribution or polices for web caching and higher security .
3) Modification of web site based on discovered web user navigation patterns—will be possible which means internal links rearrangement due to improve their visibility.

4) Business intelligence, noticeable in e-commerce activities like email marketing campaigns, cross- and up selling techniques developed and observed in e-markets
5) Characterization of use, web server log files combine with additional information.

VI. PATTERN DISCOVERY TECHNIQUES

The objective of mining process is to discover sequential association rules. This knowledge will form the knowledge base which can be used in recommendation and personalization systems.
Associations Rules, Association rules can be used to find what pages are accessed together and finding large itemsets. A page is regarded as an item, and session is regarded as a transaction.

Sequential Patterns ,a Sequential patterns is defined as an ordered set of pages that satisfies a given support and is maximal. Generating sequential patterns use k-sequence, k is the no of pages.
The WAP-tree(web access pattern) has been proposed to facilitate efficient counting . This tree is used to store the sequences and their counts.

Online Adaptive Traversal Patterns algorithm designed to find maximal frequent sequence(MFS).it uses a suffix tree to store patterns.
Mining systems and technologies are currently considered as enablers for business intelligence systems, because they improve the quantitative and qualitative value of the knowledge available to decision makers. Nowadays, the architecture for a mining system has a remarkable impact especially for large business environments, where data from numerous sources needs to be accessed and combined to provide comprehensive analyses, and work groups of analysts require access to the same data and results.
Web Mining Focus holds the phase of the Information Extraction, where preliminary steps of the Knowledge Discovery in Text, such as selecting and pre-processing of hypertext, hyperlinks or logs, take part.
Pattern Extraction where a set of Web Content, Structure and Usage Mining techniques together with other Data Mining algorithms, are collected for the analysis. Evaluation interprets the utility and the carefulness of the extracted patterns.

VII. E-LEARNING SCENARIO USING WEB USAGE MINING

Personalization in an E-learning system can be achieved through two levels of personalization. Level 1 allows the personalization of learning contents and structure of the course according to a given personalization strategy and level 2 defines the personalization strategy. Teacher has to choose and apply the personalization strategy which matches the learner's characteristics and the specifics of the courses.

TABLE 1. personalization parameters

Personalization	**Set of values**
Learner's level of knowledge	Beginner, intermediate, advanced
Learning goals	Knowledge, comprehension, Application
Motivation levels	Low, moderate, high
Navigation	Breadth-first, Depth first

This system allows teachers with a similar profile to share their research results as a consequence of applying mining locally on their own courses. Furthermore a new interactive and iterative association rule mining algorithm was developed, using a new weight-based evaluation measure for the rules discovered and taking into account the opinion of experts and the teachers themselves in order to produce increasingly effective recommendations. Experimental trials were performed taking into account two points of view: that of the teacher making the changes based on the recommendations provided by the system; and that of the student doing the course, after it had been modified by the teacher.

VIII. CONCLUSION

As the web and its usage continues to grows the opportunity to analyze web data and extract all manner of useful knowledge from it. The past few years have seen the emergence of web mining as a rapidly growing area, due to the efforts of the research community as well as various organizations that are practicing it. In this chapter we have briefly described the key computer science contributions made by the field, a number of prominent applications ,and outlined some areas of future research. This system allows teachers with a similar profile to share their research results as a consequence of applying mining locally on their own courses. Our hope is that this overview provides a starting point for fruitful discussion.

REFERENCES

[1] B. Mobasher, N. Jain, E.-H.S. Han and J. Srivastava, Web Mining: Pattern Discovery from World Wide Web Transactions, (1996).
[2] Agrawal, R., Imielinski, T., Swami, A.N.: Mining Association Rules between Sets of Items in Large Databases. In: Proceeding SIGMOD, pp. 207–216 (1993)
[3] Agrawal, R., Srikant, R.: Mining Sequential Patterns. In: Proceedings of the Eleventh International Conference on Data Engineering, pp. 3–14 (1995)
[4] Alcalá, J., del Jesús, M.J., Garrell, J.M., Herrera, F., Hervás, C., Sánchez, L.: Proyecto KEEL: Desarrollo de una Herramienta para el Análisis e Implementación de Algoritmos de Extracción de Conocimiento Evolutivos. Tendencias de la Minería de Datos en España,Eds. J. Giraldez, J.C. Riquelme, J.S. Aguilar, pp. 413–423 (2004)
[5] Brusilovsky, P., Peylo, C.: Adaptive and Intelligent Web-based Educational Systems. International Journal of Artificial Intelligence in Education. 13, 156–169 (2003)
[6] J. Borges and M. Levene, Data Mining of User Navigation Patterns,Book Data Mining of User Navigation Patterns, Series Data Mining of User Navigation Patterns, Springer-Verlag, 92-111, (2000).

[7] V. Chitraa and A.S. Davamani, A Survey on Preprocessing Methods for Web Usage Data, International Journal of Computer Science and Information Security, vol. 7, no. 3, 78-83, (2010).

[8] 1. Lu, J.:2004, '***Personalized e-learning Material Recommender System'***. In Proceedings of the International Conference on Information Technology for Application, London, England, pp. 374–379.

[9] 2. P. Dolog and M. Sintek, ***"Personalization in Distributed Learning Environments***," WWW 2004, ACM 1-58113-912-8/04/0005, New York, USA, May 17-22, 2004