Volume 4, Issue 3, March 2014 ISSN: 2277 128X

International Journal of Advanced Research in Computer Science and Software Engineering

Research Paper

Available online at: www.ijarcsse.com

# A Review on Promising Development of Research based on Web Mining

**Sachin Balvir**
*Assistant Professor*
*Dept. of Computer Science & Engg*
*DMIETR, Wardha, India*

**Prashant Dahiwale**
*Assistant Professor*
*Dept. of Computer Science & Engg*
*RGCER, Nagpur, India*

**Abhiram Gandhe**
*PhD Scholar*
*Dept. of Computer Science,*
*VNIT, Nagpur, India*

***Abstract— From its very beginning, the prospective of extracting precious knowledge from the web has been relatively marked. Web mining i.e. application of data mining techniques to mine knowledge from web content, structure & usage is the collection of technologies to accomplish this potential. Interest in web mining has grown rapidly in the short period of time. Today there are several billions of HTML documents, images and other files available on www via internet and the number is still getting bigger. But taking into consideration the exciting diversity of the web, retrieving interesting content has become a very difficult task. So, the World Wide Web is a fertile area for data mining research. This paper provides the concise outline about the research and application in web mining.***

***Keywords— Web mining, Content Web Mining, Structure Web Mining, Web content Mining, Web Mining Applications***

## I. INTRODUCTION

The wide use of the Internet has essentially changed the ways in which we communicate, collect information and make purchases. As the utilization of the World Wide Web (WWW) and exchange of email, data over internet increased radically, the computer scientists hurried to describe this new phenomenon. In the beginning they were shocked by the incredible mixture the Internet confirmed in the size of its features, they soon exposed a general pattern in their capacity: there are a lot of small elements enclosed within the Web, but only some large ones. A few sites have of millions of pages, but millions of sites only have a handful of pages. Few sites have millions of links, but many sites contain one or two. Millions of users gather to a few select sites, giving little awareness to millions of others.[1]With the volatile increase of information sources offered on the World Wide Web, it has become more and more essential for users to utilize computerized tools in order to find, extract, filter, and evaluate the preferred information and resources. In addition, with the revolution of the Web into the main tool for electronic commerce, it is essential for organizations and companies, who have invested millions in Internet and intranet technologies, to follow and evaluate user behavior. These factors give rise to the need of creating client-side and server-side intelligent systems that can successfully extract knowledge. Many organizations make available information and services on the web such as on-line shopping, customer support, web based applications etc. are becoming common practice. The WWW is becoming everywhere and a regular tool for daily activities of common people, from a child to a senior across the world [2].

## II. OVERVIEW OF WEB MINING

Web Mining is based on knowledge discovery (KD) from web. It extracts the knowledge & represents in a proper way. Web mining is like a graph & all pages are node & each connects with hyperlinks. Web mining is useful to extract the information, image, text, documents etc. If we want search any topic from web, then it is difficult to get exact topic information because as there are lots of data available on web related to that topic but now's day it is very easy to get the proper information about any things. Web mining is based on data mining technique by using data mining technique we discover the hidden data present in the web. Thus, web mining is considered as an application of data mining.[1]

### *A. Task of Web Mining:*

To clarify the uncertainty to find out what forms Web mining. Kosala and Blockeel [4] had recommended a decomposition of Web mining in the following tasks:

1. *Resource finding*: the task of retrieving intended Web documents.

2. *Information selection and pre-processing*: automatically selecting and pre-processing specific information from retrieved Web resources.

3. *Generalization*: automatically discovers general patterns at individual Web sites as well as across multiple sites.

4. *Analysis*: validation and/or interpretation of the mined patterns.

 

The fig.1 shows the various tasks of web mining.

***B. Web mining Categories:***

In general, Web mining tasks can be classified into three categories: Web content mining (WCM), Web structure mining (WSM) and Web usage mining (WUM). All of the three categories focus on the process of knowledge discovery of implicit, previously unknown and potentially useful information from the Web. Each of them focuses on different mining objects of the Web. As follows, provide a brief introduction about each of the categories.

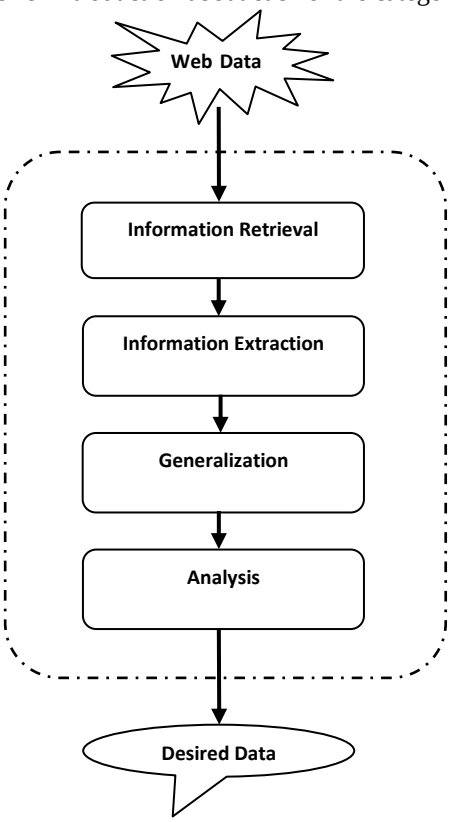


Fig.1. Various Tasks of Web Mining

***i. Web Content Mining(WCM):***

Web content mining is also known as text mining. Content mining is the scanning and mining of text, pictures and graphs of a Web page to determine the relevance of the content to the search query. This scanning is completed after the clustering of web pages through structure mining and provides the results based upon the level of relevance to the suggested query. Web content mining describes the automatic search of information resources available online [5], and involves mining web data contents. With the massive amount of information that is available on the World Wide Web, content mining provides the results lists to search engines in order of highest relevance to the keywords in the query. The main uses for this type of data mining are to gather, categorize, organize and provide the best possible information available on the WWW to the user requesting the information.

The various contents of Web Content Mining are Web page, Search page and Result page [6]. A Web page typically contains a mixture of many kinds of information, e.g., main content, advertisements, navigation panels, copyright notices, etc. For a particular application only some part of the information is useful and the rest are noises. A search page is typically used to search a particular Web page of the site, to be accessed numerous times in relevance to search queries. A result page typically contains the results, the web pages visited and the definition of last accurate result in the result pages of content mining.

***ii. Web Structure Mining(WSM) :***

The majority of the web information retrieval tools only uses the textual information and ignores the link information that could be very important. Web structure mining is a tool used to recognize the relationship between Web pages linked by information or direct link connection. This connection allows a search engine to pull data relating to a search query directly to the linking Web page from the Web site the content rests upon. This completion takes place through use of spiders scanning the Web sites, retrieving the home page, then, and linking the information through reference links to bring forth the specific page containing the desired information. The main purpose for structure mining is to extract previously unknown relationships between Web pages. This structure data mining provides use for a business to link the information of its own Web site to enable navigation and cluster information into site maps.

The structural information generated from Web structure mining includes the follows [2]: the information measuring the occurrence of the local links in the Web tuples in a web table containing links that are internal and the links that are

 

within the same document. The information measuring the frequency of web tuples in a web table that contains links that are global and the links that span different web sites, web structure mining has a nature relation with the web content mining, since it is very likely that the Web documents contain links, and they both use the real or primary data on the web

***iii. Web Usage Mining(WUM):***

Web usage mining is used to discover user navigation patterns and the useful information from the web data present in server logs, which are maintained during the interaction of the users while surfing on the web. Most existing Web analysis tools provide mechanisms for reporting user activity in the servers and various forms of data filtering. Using such tools it is possible to determine the number of accesses to there server and to individual files, the times of visit and the domain names and url's of users.[7]

Web mining usage aims at utilize data mining techniques to discover the usage patterns from web based application. It is technique to predict user behavior when it is interact with the web. [8] Web usage mining is categories in three phases:

*a) Preprocessing*- According to client, server and proxy server it is first approach to retrieves the raw data from web resources and processed the data .it is automatically transformed the original raw data.

*b) Pattern Discovery*- According the data preprocessing discovered the knowledge and implements the techniques to discover the knowledge like as machine learning and data mining procedures are carried out at this stage.

*c) Pattern Analysis*- pattern analysis is the process after pattern discovery. Its check the pattern is correct on the web and how to implement on web to extract the information on your web search / extract knowledge from the web.[3]

Web usage mining is the application of data mining techniques to discover usage patterns from Web data, in order to understand and better serve the need of Web-based application. Comparison of web mining categories on the parameter like sources, data forms & search objects are given in table 1.

TABLE 1: COMPARISON OF WEB MINING CATEGORIES

| Category | Sources | Data forms | Search objects |
|---|---|---|---|
| WCM | Web Pages | Text | Collection of websites |
| WSM | Topology used | hyperlinks | Collection of hyperlinks |
| WUM | Server log, web log, cookies | Click forms, Access pattern | Collection of Log Data |

## III. TECHNIQUES OF WEB MINING

There are different types of techniques are offered for data mining. The most frequently used techniques are artificial neural networks, decision trees, and the nearest-neighbor method visualization, association rule, classification and clustering. each one of these approaches brings special advantages and disadvantages that have to be considered prior to their use. A good way to apply advanced data mining techniques is to have a flexible and interactive data mining tool that is completely incorporated with a database or data warehouse. In spite of of the technique used, the real value behind data mining is modeling — the method of building a model based on user-specified criteria from already captured data.

In web mining some of the techniques of data mining can be used for example association rule, classification and clustering etc. The brief descriptions of these techniques are as follows.

***Classification:***

Classification is use to build up a idea of the type of customer, item, or object by relating several attributes to categorize a particular class. For example, one can easily classify cars into different types (sedan, 4x4, convertible) by identifying different attributes (number of seats, car shape, driven wheels). Given a new car, we might relate it into a particular class by comparing the attributes with our known definition. One can apply the same principles to customers, for example by classifying them by age and social group. Classification algorithms can be used to categorize users into special classes according to their browsing behavior or pattern. The criterion by which items are assigned to different clusters is the degree of similarity among them

***Prediction:***

Prediction is a wide topic and runs from predicting the failure of components or machinery, to identifying fraud and even the prediction of company profits. Used in combination with the other web mining techniques, prediction involves analyzing trends, classification, pattern matching, and relation. By analyzing past events or instances, you can make a prediction about an event.

***Sequential Pattern:***

Sequential patterns are a helpful technique for identifying trends, or regular occurrences of similar events. For example, with customer data we can identify that customers buy a particular collection of products together at different times of the year. In a shopping basket application, you can use this information to automatically suggest that certain items be added to a basket based on their frequency and past purchasing history. In web usage mining, sequential patterns are exploited to find sequential navigation patterns that appear in users sessions frequently.

***Decision Tree:***

Decision tree learning is a technique commonly used in data mining. The objective is to build a model that predicts the value of a target variable based on several input variables. A decision tree is a simple representation for

 

classifying examples. Decision tree learning is one of the most successful techniques for supervised classification learning

*Association Rules:*

The association rule method can be used to show pages that are mainly referenced together and to discover the direct or indirect relationships between web pages in users browsing behavior. Here, we can make a simple correlation between two or more items, often of the same type to identify patterns. After transactions are detected in the preprocessing phase, frequent item-sets are discovered using the A-priori algorithm. A hypergraph is an extension of a graph where each hyperedge can connect more than two vertices. A hyperedge connects URLs within a frequent item-set. Each hyperedge is weighted by the averaged confidence of all the possible association rules formed on the basis of the frequent item-set that the hyperedge represents [2].

## IV. APPLICATIONS OF WEB MINING

There are various application areas of web mining [9] some of which are shown in fig.2 and briefly discussed as follows:

- ***E-Learning:*** Web mining can be used in e-learning environment. The process of learning in e-learning environments is improved & enhanced by using web mining techniques. Applications of web mining to e-learning usage based.
- ***Search Engine:*** In development of search engine applications web mining technique plays a main role. With the help of different web mining techniques users are able to get the interested data from the search engine from the huge amount of related data.
- ***Personalization:*** Web personalization can be defined as the process of customizing the content and structure of website to the specific and individual needs of each user taking advantage of the user's navigational behavior. The steps of web personalization process include: Collection of web data, the modeling and categorization of these data (preprocessing phase), the analysis of the collected data and determination of the actions that should be performed. The ways that are employed in in order to analyze the collected data include content based filtering, collaborative filtering, rule based filtering and web usage mining.[10]

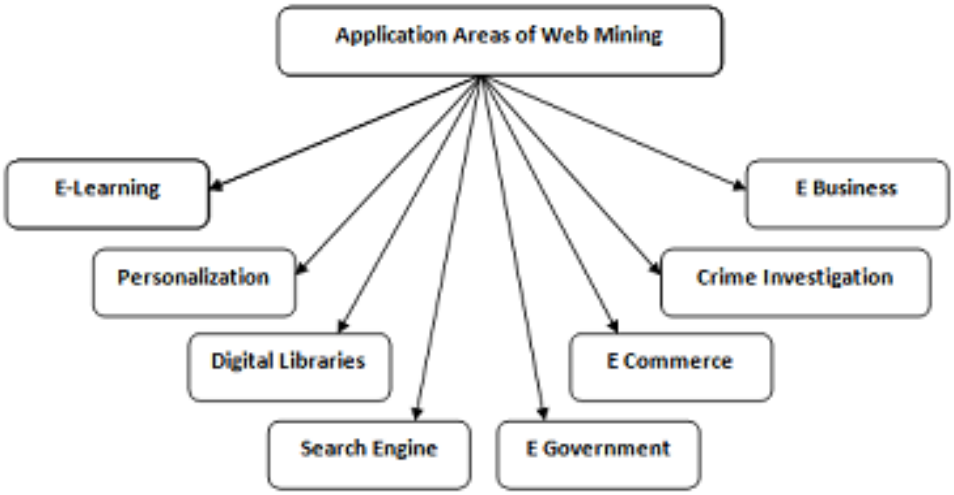


Fig.2. Application Areas of Web Mining

- ***Digital Libraries:*** Digital libraries provide services to distribute valuable information all around the world, eliminating the necessity to be physically present at different libraries in different part of the world.
- ***E-Government:*** With the rapid growth and development of electronic government as well as the ease and speed with which government affairs can be carried out over the web, one of the application fields of web mining is electronic government systems. Electronic government is one of the most appropriate applications of data mining: richest and the most common source of data, automatically generated data. The use of web mining can be quickly converted into the government behavior, at the same time the policies of government derived from the mining can be evaluated in time.
- ***E-Commerce:*** Web is the best medium of communication in modern business. Many companies are redefining their business strategies to improve the business output. Business over internet provides the opportunity to customers and partners where their products and specific business can be found. Nowadays online business breaks the barrier of time and space as compared to the physical office. Big companies around the world are realizing that e-commerce is not just buying and selling over Internet, rather it improves the efficiency to compete with other giants in the market. For this purpose web mining techniques are very much useful. Web mining is a technique that is applied to the WWW as there are vast quantities of information available over the Internet.
- ***Crime Investigation:*** Cyber criminals exploit opportunities for anonymity and masquerade in web-based communication to conduct illegal activities such as phishing, spamming, cyber predation, cyber threatening, blackmail, and drug trafficking. One way to fight cyber crime is to collect digital evidence from online documents

 

and to prosecute cyber criminals in the court of law. A lot of facts have proved that it is not enough to manage the information on the Internet simply through traditional administrative models. In this concern, Web mining is a novel research direction for the information gathering and analyzing on the Internet, which is explosive and unstructured. The focuses of Web mining research are to develop new web mining techniques and to extract the features of texts to represent them. [11]

- ***E-Business:*** The drastic development of internet and information technology made E-business which is a new type of commercial channels developing prosperously. How to analyze data of E-business users for mining users' information that enterprises interest is critical for their development. Web mining has advantages that it can mine data efficiently and intelligently, so it is becoming more and more important in modern E-business.

## V. CONCLUSION

We, review the researches in the area of web mining. Three standard types of web data mining are introduced generally. Web data is budding at a significant rate. Web mining is a fruitful area of research with many booming applications. As the Web and its usage continues to develop, so develops the chance to explore Web data and pull out useful knowledge from it. Web mining is a revolution that the Internet has grown from a simple search tool to a gold mine. It is having its own benefits and successful applications with which we can overcome the problems or difficulties faced in data mining. In this paper a clear idea of how web mining is efficiently use has been highlighted. Companies find a new and better way to do business: E-commerce through the Internet. Companies have to implement Web mining systems to understand their customers' profiles, and to identify their own strength and weakness of their E-marketing efforts on the web through continuous improvements.

## REFERENCES

[1] Kavita Sharma, Gulshan Shrivastva and Vikas Kumar, "Web Mining: Today and Tomorrow", 3rd International Conference on Electronics Computer Technology (ICECT 2011), In IEEE, 2011.

[2] Manoj Padia, Shubhendhu Kumar Pani, Sanjay Kumar Padhi, Lingaraj Panigrahy and R. Ramakrishna, "A Review of Trends in Research on Web Mining", International Journal of Instrumentation , Control & Automation (IJICA), volume 1, Issue 1, 2011.

[3] TIAN Meirong and CHEN Xuedong, "Application of Agent Based Web Mining in E-Business", 2nd international Conference on Intelligent Human-Machine Systems and Cybernetics, In IEEE, 2010.

[4] Raymond Kosala and Hendrik Blockeel, "Web Mining Research: A Survey", ACM SIGKDD Explorations Newsletter, June 2000, Volume 2, Issue 1.

[5] S.K.Madria, S.SBhowmick, W.K, Ng, and E.P.Lim."Research issues in web data mining", In Proceedings of data ware housing and knowledge Discovery, first International conference, DaWak'99, pages 303-312, 1999.

[6] Mr. Dushyant B. Rathod and Dr. Samrat Khanna, "A Review on Emerging Trends of Web Mining and Its application", International Journal of Engineering Development and Research (IJEDR).

[7] S. U. Balvir, G. N. Tikhe and L. M. Barapatre, "Positioning Webpage Using Rank", International Journal of Computer Technology and Electronics Engineering (IJCTEE), 2011.

[8] K. L. Wu, P. S. Yu and A. Ballman, "Speed Tracer: A Web usage mining and analysis tool", IBM Systems Journal, Volume 73, No. 1, 1998.

[9] S. Yadav, K. Ahmad and J. Shekhar," Analysis of web mining application and Beneficial Areas", IIUM, Engineering Journal Volume 12, No. 2, 2011.

[10] Magdalini Eirinaki and Michalis Vazirgiannis, "Web Mining for Web Personalization", ACM Transactions on Internet Technology, Volume 3, No.1, February 2003, Pages 1-27.

[11] Javed Hosseinkhani, Mohammad koochakzaei, SolmazKeikhaee and Javid Hosseinkhani Naniz, "Detecting Suspicion Information on the Web Using Crime Data Mining Techniques", International Journal of Advance Computer Science and Information Technology (IJACSIT), Vilume 3, No. 1, 2014.

 