Available online at http://acfa.apeejay.edu

APEEJAY JOURNAL OF COMPUTER SCIENCE AND APPLICATIONS

ISSN:0974-5742(P)
Vol. 3, January 2015

Web Mining: Intelligent way of mining Web based data

Neeru Mago
Department of Computer Science And Applications
PUSSGRC, Hoshiarpur
neerumago80@gmail.com

Abstract
As the data on the web is increasing enormously, the World Wide Web becomes a potential area for handling Big Data using web data mining. The World Wide Web acts as an interactive and popular way to transfer information. The users cannot make use of the information very effectively and easily from the vast and diverse data on the web. Web data mining is a main converging research field, which extracts previously unknown, useful patterns and information available on the web. Till date, web mining needs more discussion in order to enhance motivation among researchers to promote research in Web Mining. In this paper, we survey the recent research and existing techniques in the area of Web Mining. Various applications, issues related to web mining and current trends in the field of Web mining are also discussed in this paper.
Keywords: *Data mining, Web Mining, Web Information retrieval, BI, FL, ANN, Web mining process.*

1. Introduction

With the rapid development of the Web, it becomes necessity to provide users with tools that are efficient and effective for resource discovery and knowledge discovery on the Web [1]. Although we have Web search engines to assist in resource discovery, it is far from satisfying for its poor precision. Moreover, the purpose of the Web search engine is only to discover resource on the Web. As far as knowledge discovery is concerned, it is not equal to at all even with high precision. Therefore, the research and development of new technology ahead of resource discovery is needed. Data mining is used to identify valid, novel, potentially useful and ultimately understandable pattern from data collection in database community [2]. However, there is not much work done on unstructured and heterogeneous information on the Web. Web mining is a new research area which draws great interest from different communities and it needs more discussion among researchers in order to define it.

Web is a collection of billions of enormous, diverse, flexible, and dynamic documents [3]. The World Wide Web continues to grow in the huge volume of traffic, size and complexity of Web sites. It is difficult to identify the relevant information present in the web. Moreover, most of the contents in the web are unstructured in nature. These problems can be addressed by the emerging field of web mining aims. It aims at finding and extracting relevant information that is hidden in Web-related data, particularly in text documents published on the Web. Data Mining deals with the extraction of useful, meaningful and valuable information from huge collection of data. Web mining is the vital area in data mining where the interesting patterns are extracted from the web contents. Web mining is the one of the applications of data mining techniques to mine knowledge from web data, including web documents, hyperlinks between documents, usage logs of web sites, etc [4].

1.1 Web Mining

Web mining is an integrated technology in which several research fields are involved, such as data mining, computational linguistics, statistics, and informatics, etc. Different researchers from different communities disagree with each other on what Web mining is exactly [1]. We present a more general definition of Web mining as follows.

Definition1. Web mining is the activity of identifying patterns p implied in large document collection C, which can be denoted by a mapping:
$f : C \rightarrow p$
Although the definition of Web mining is similar to the well-known definition of data mining, it has many unique characteristics of its own.. Firstly, the source of Web mining is web documents. We consider the use of the Web as a middleware in mining database and the mining of logs; user profiles on the Web server still belong to the category of traditional data mining. Secondly, the Web is a directed-graph consists of document nodes and hyperlinks. Therefore, the pattern identified can be possibly about the content of documents or about the structure of the Web. Moreover, the Web documents are semi-structural or non-structural with little machine-readable semantic while the source of data mining is confined to the structural data in database. As a result, some traditional data mining methods are not applicable to Web mining. Even if applicable, they must be based on the pre-processing of documents.

1.2 Web information retrieval

Definition 2. Web information retrieval is the process to find a subset S of appropriate number of documents relevant to a certain query q from large document collection C, which can also be denoted by a mapping:
$f : (C,q) \rightarrow S$.
Since 1960, there have been many achievements in the field of information retrieval, such as index model, document representation and similarity measure. These achievements were applied on the Web successfully, which gave rise to search engines. In recent years, some researchers applied database concept to the Web and presented some new methods of modelling and querying the Web at a finer granularity level than pages, such as WebOQL, Lorel, etc. These methods can retrieve not only the hyperlink between Web pages but also the internal structure within a web page.
Web information retrieval and Web mining have different goals. Although Web mining is ahead of Web information retrieval, it does not intend to replace Web information retrieval. Instead these are two technologies supplement to each other. Each has its applications, uses and issues. Rather, Web mining can be utilized to increase the precision of information retrieval. It also improves the organization of retrieval results which will bring the information retrieval system into the next generation.

2. Related work

In this section, literature review related to web mining is presented [5]. A study by Chakraborti [6] focuses on the problem areas of data mining related to hypertext, as well as on the problem of learning. In this, a different methodology such as supervised/ unsupervised learning and their implementation on hypertext are presented. Srivastava et. al. [7] gives an overview of the system called "WebSIFT" which describes three phases of web usage mining as pre-processing, pattern discovery and pattern analysis. Shankar et. al. [8]highlight the limitations of data mining tools and advantages of web mining technologies in addition to a discussion on the future directions of using fuzzy logic (FL), artificial neural networks (ANN), genetic algorithms (GA) and rough sets (RSs).
Abraham [9] proposes "Intelligent Miner (i-miner)" with fuzzy clustering and fuzzy inference system algorithms for complex e- commerce applications known as Business Intelligence (BI). A study by Renata and Vajk [10] deals with "Frequent Pattern Mining" for web logs and discusses three pattern mining approaches from web usage mining point of view as Page sets, Page sequences and Page graphs. Atanasova et. al. [11] presents the solutions to some problems in the marketing subsystem through a proposed functional matrix for the application of data mining, text mining and web mining. Mei and Cheng [12] suggest the process of web mining techniques, its application towards electronic commerce and concludes the relationship between electronic and web data mining and gives the use of web mining technology in electronic commerce. Recently, Yu and Yang [13] discuss the methods and processes of data mining and their application in the field of electronic commerce. They offer the proper classification of web mining, thus differentiating it from data mining.

3. Web Mining Taxonomy

Web mining can be broadly divided into three distinct categories, according to the kinds of data to be mined. Fig. 1 shows the taxonomy.

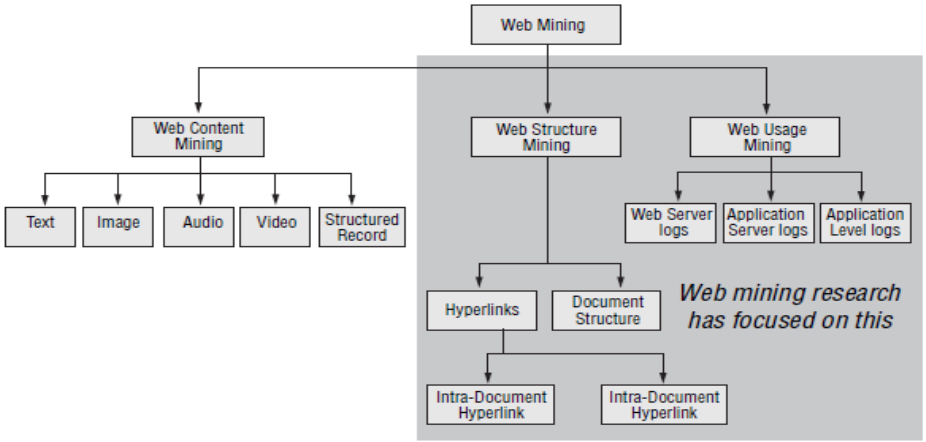

Fig. 1: Web Mining Taxonomy

3.1 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, or structured records such as 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. Research activities on this topic have drawn heavily on techniques developed in other disciplines such as Information Retrieval (IR) and Natural Language Processing (NLP). While there exists a significant body of work in extracting knowledge from images in the fields of image processing and computer vision, the application of these techniques to web content mining has been limited.

3.2 Web Structure Mining

The structure of a typical 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 two kinds based on the kind of structure information used.

A) Hyperlinks: A hyperlink is a structural unit that connects a location in a web page to a different location, either within the same web page or on a different web page. A hyperlink that connects to a different part of the same page is called an intra-document hyperlink, and a hyperlink that connects two different pages is called an inter-document hyperlink. There has been a significant body of work on hyperlink analysis, of which Desikan, Srivastava, Kumar, and Tan (2002) provide an up-to-date survey.

B) Document Structure: In addition, the content within a Web page can also be organized in a tree- structured format, based on the various HTML and XML tags within the page. Mining efforts here have focused on automatically extracting document object model (DOM) structures out of documents (Wang and Liu 1998; Moh, Lim, and Ng 2000).

3.3 Web Usage Mining

Web usage mining is the application of data mining techniques to discover interesting usage patterns from web usage data, in order to understand and better serve the needs of web-based applications (Srivastava, Cooley, Deshpande, and Tan 2000). Usage data captures the identity or origin of web users along with their browsing behavior at a web site. Web usage mining itself can be classified further depending on the kind of usage data considered:

A) Web Server Data: User logs are collected by the web server and typically include IP address, page reference and access time.

B) Application Server Data: Commercial application servers such as Weblogic, StoryServer, have significant features to enable E-commerce applications to be built on top of them with little effort. A key feature is the ability to track various kinds of business events and log them in application server logs.

C) Application Level Data: New kinds of events can be defined in an application, and logging can be turned on for them — generating histories of these events. It must be noted, however, that many end applications require a combination of one or more of the techniques applied in the above the categories.

4. Web Mining process

The web mining process [14] consists of the following phases:

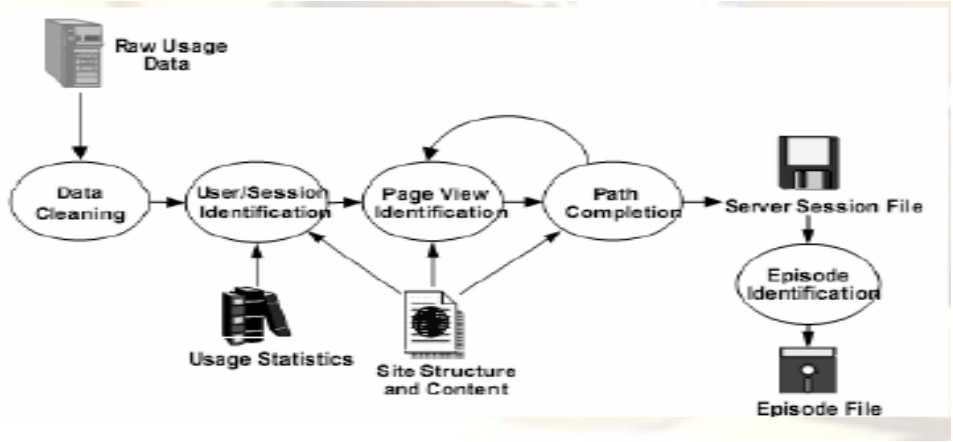

Fig. 2: Pre-processing of web usage data

A) Data accumulation: Data accumulation is the first step of web usage mining, the data authenticity and integrity will directly affect the following works smoothly carrying on and the final recommendation of characteristic service's quality.

B) Data pre-processing: Some databases are insufficient, inconsistent. The data pre-treatment is to carry on a unification transformation to those databases. The result is that the database will to become integrate and consistent, thus establish the database which may mine. In the data pre-treatment work, mainly include data cleaning, user identification, session identification and path completion.

i) Data Cleaning: The purpose of data cleaning is to eliminate irrelevant items, and these kinds of techniques are of importance for any type of web log analysis not only data mining. According to the purposes of different mining applications, irrelevant records in web access log will be eliminated during data cleaning.

a) The records of graphics, videos and the format information. The records have filename suffixes of GIF, JPEG, CSS, and so on, which can found in the URI field of the every record.
b) The records with the failed HTTP status code. By examining the Status field of every record in the web access log, the records with status codes over 299 or fewer than 200 are removed. It should be pointed out that different from most other researches, records having value of POST or HEAD in the Method field are reserved in present study for acquiring more accurate referrer information.

ii) User and Session Identification: The task of user and session identification is to find out the different user sessions from the original web access log. User's identification is, to identify who access web site and which pages are accessed. The goal of session identification is to divide the page accesses of each user at a time into individual sessions

iii) Path completion: Another critical step in data pre-processing is path completion. There are some reasons that result in path's in completion, for instance, local cache, agent cache, "post" technique and browser's "back" button can result in some important accesses not recorded in the access log file, and the number of Uniform Resource Locators(URL) recorded in log may be less than the real one. As a result, the user access paths are incompletely preserved in the web access log. To discover user's travel pattern, the missing pages in the user access path should be appended. The purpose of the path completion is to accomplish this task. The better results of data pre-processing, we will improve the mined patterns' quality and save algorithm's running time.

C) Knowledge Discovery Use statistical method to carry on the analysis and mine the pretreated data. We may discover the user or the user community's interests then construct interest model. At present the usually used machine learning methods mainly have clustering, classifying, the relation discovery and the order model discovery.

D) Pattern analysis Challenges of Pattern Analysis are to filter uninteresting information and to visualize and interpret the interesting patterns to the user. First delete the less significance rules or models from the interested model storehouse; Next use technology of OLAP and so on to carry on the comprehensive mining and analysis; Once more, let discovered data or knowledge be visible; Finally, provide the characteristic service to the electronic commerce website

5. WEB MINING APPLICATIONS

In this section, some of the most successful applications (with examples) of web mining are described.

A) Personalized Customer Experience in B2C E-commerce—Amazon.com

Early on in the life of Amazon.com, its visionary CEO Jeff Bezos observed, "In a traditional (brick-and-mortar) store, the main effort is in getting a customer to the store. Once a customer is in the store they are likely to make a purchase — since the cost of going to another store is high — and thus the marketing budget (focused on getting the customer to the store) is in general much higher than the in-store customer experience budget (which keeps the customer in the store). In the case of an on-line store, getting in or out requires exactly one click, and thus the main focus must be on customer experience in the store." This fundamental observation has been the driving force behind

Amazon's comprehensive approach to personalized customer experience, based on the mantra "a personalized store for every customer" (Morphy 2001). A host of web mining techniques, such as associations between pages visited and click-path analysis are used to improve the customer's experience during a "store visit." Knowledge gained from web mining is the key intelligence behind Amazon's features such as "instant recommendations," "purchase circles," "wish-lists," etc.

B) Web Search—Google

Google is one of the most popular and widely used search engines. It provides users access to information from over 2 billion web pages that it has indexed on its server. The quality and quickness of the search facility makes it the most successful search engine. Earlier search engines concentrated on web content alone to return the relevant pages to a query. Google was the first to introduce the importance of the link structure in mining information from the web.

i) PageRank, which measures the importance of a page, is the underlying technology in all Google search products, and uses structural information of the web graph to return high quality results. The key idea is that a page has a high rank if it is pointed to by many highly ranked pages. So, the rank of a page depends upon the ranks of the pages pointing to it. This process is done iteratively until the ranks of all pages are determined. The rank of a page p can be written as:

$$PR(p) = d/n + (1-d) \sum_{(q,p)\in G} \left(\frac{PR(q)}{Outdegree(q)}\right)$$

Here, n is the number of nodes in the graph and OutDegree (q) is the number of hyperlinks on page q . Intuitively, the approach can be viewed as a stochastic analysis of a random walk on the web graph. The first term in the right hand side of the equation is the probability that a random web surfer arrives at a page p by typing the URL or from a bookmark; or may have a particular page as his/her homepage. Here d is the probability that the surfer chooses a URL directly, rather than traversing a link and 1−d is the probability that a person arrives at a page by traversing a link. The second term in the right hand side of the equation is the probability of arriving at a page by traversing a link.

ii) The Google toolbar is another service provided by Google that seeks to make search easier and informative by providing additional features such as highlighting the query words on the returned web pages. The full version of the toolbar, if installed, also sends the click-stream information of the user to Google. The usage statistics thus obtained are used by Google to enhance the quality of its results.

iii) Google also provides advanced search capabilities to search images and find pages that have been updated within a specific date range.

iv) Built on top of Netscape's Open Directory project, Google's web directory provides a fast and easy way to search within a certain topic or related topics.

v) The advertising program introduced by Google targets users by providing advertisements that are relevant to a search query. This does not bother users with irrelevant ads and has increased the clicks for the advertising companies by four to five times. According to BtoB, a leading national marketing publication, Google was named a top 10 advertising property in the Media Power 50 that recognizes the most powerful and targeted business-to-business advertising outlets.

vi) One of the latest services offered by Google is Google News. It integrates news from the online versions of all newspapers and organizes them categorically to make it easier for users to read "the most relevant news." It seeks to provide latest information by constantly retrieving pages from news site worldwide that are being updated on a regular basis. The key feature of this news page, like any other Google service, is that it integrates information from various web news sources through purely algorithmic means, and thus does not introduce any human bias or effort. However, the publishing industry is not very convinced about a fully automated approach to news distillation (Springer 2002).

C) Web-Wide Tracking—DoubleClick

"Web-wide tracking," i.e. tracking an individual across all sites he visits, is an intriguing and controversial technology. It can provide an understanding of an individual's lifestyle and habits to a level that is unprecedented, which is clearly of tremendous interest to marketers. A successful example of this is DoubleClick Inc.'s DART ad management technology. DoubleClick serves advertisements, which can be targeted on demographic or behavioral attributes, to the end-user on behalf of the client, i.e. the web site using DoubleClick's service. Sites that use

DoubleClick's service are part of The DoubleClick Network and the browsing behavior of a user can be tracked across all sites in the network, using a cookie. This makes DoubleClick's ad targeting to be based on very sophisticated criteria. Alexa Research has recruited a panel of more than 500,000 users, who have voluntarily agreed to have their every click tracked, in return for some freebies. This is achieved through having a browser bar that can be downloaded by the panelist from Alexa's website, which gets attached to the browser and sends Alexa a complete click-stream of the panelist's web usage. Alexa was purchased by Amazon for its tracking technology. Clearly web-wide tracking is a very powerful idea. However, the invasion of privacy it causes has not gone unnoticed, and both Alexa/Amazon and DoubleClick have faced very visible lawsuits. Microsoft's Passport technology also falls into this category. The value of this technology in applications such as cyber-threat analysis and homeland defense is quite clear, and it might be only a matter of time before these organizations are asked to provide information to law enforcement agencies.

D) Understanding Web Communities—AOL

One of the biggest successes of America Online (AOL) has been its sizeable and loyal customer base. A large portion of this customer base participates in various AOL communities, which are collections of users with similar interests. In addition to providing a forum for each such community to interact amongst themselves, AOL provides them with useful information and services. Over time these communities have grown to be well-visited water- holes for AOL users with shared interests. Applying web mining to the data collected from community interactions provides AOL with a very good understanding of its communities, which it has used for targeted marketing through advertisements and e-mail solicitation. Recently, it has started the concept of "community sponsorship," whereby an organization, say Nike, may sponsor a community called "Young Athletic Twenty Somethings." In return, consumer survey and new product development experts of the sponsoring organization get to participate in the community, perhaps without the knowledge of other participants. The idea is to treat the community as a highly specialized focus group, understand its needs and opinions on new and existing products, and also test strategies for influencing opinions.

E) Understanding Auction Behaviour—eBay

As individuals in a society where we have many more things than we need, the allure of exchanging our useless stuff for some cash, no matter how small, is quite powerful. This is evident from the success of flea markets, garage sales and estate sales. The genius of eBay's founders was to create an infrastructure that gave this urge a global reach, with the convenience of doing it from one's home PC. In addition, it popularized auctions as a product selling and buying mechanism and provides the thrill of gambling without the trouble of having to go to Las Vegas. All of this has made eBay as one of the most successful businesses of the internet era. Unfortunately, the anonymity of the web has also created a significant problem for eBay auctions, as it is impossible to distinguish real bids from fake ones. eBay is now using web mining techniques to analyze bidding behaviour to determine if a bid is fraudulent (Colet 2002).

F) Personalized Portal for the Web—MyYahoo

Yahoo was the first to introduce the concept of a "personalized portal," i.e. a web site designed to have the look-and-feel and content personalized to the needs of an individual end-user. This has been an extremely popular concept and has led to the creation of other personalized portals such as Yodlee for private information like bank and brokerage accounts. Mining MyYahoo usage logs provides Yahoo valuable insight into an individual's web usage habits, enabling Yahoo to provide personalized content, which in turn has led to the tremendous popularity of the Yahoo web site.

G) CiteSeer—Digital Library and Autonomous Citation Indexing

NEC Research Index, also known as CiteSeer (Bollacker, Lawrence, and Giles 1998) is one of the most popular online bibiliographic indices related to computer science. The key contribution of the CiteSeer repository is its "Autonomous Citation Indexing" (ACI) (Lawrence, Giles, and Bollacker 1999). Citation indexing makes it possible to extract information about related articles. Automating such a process reduces a lot of human effort, and makes it more effective and faster. CiteSeer works by crawling the web and downloading research related papers. Information about citations and the related context is stored for each of these documents. The entire text and information about the document is stored in different formats. Information about documents that are similar at a sentence level (percentage of sentences that match between the documents), or at a text level is also given. Citation statistics for documents are computed that enable the user to look at the most cited or popular documents in the

related field. They also maintain a directory for computer science related papers, to make search based on categories easier. These documents are ordered by the number of citations.

6. Web Mining Issues

Web mining is a technique in data mining that automatically retrieves extracts and analyzes the information from web. Yang and Wu et al, (2006) discuss about the various issues to be addressed in data mining. The major issues include Automated Data Cleaning, Over Fitting, Under Fitting and Oversampling of data, Scaling up for high dimensional data, Mining sequence and time series data. A poll was conducted and given by K.D. nuggets and many of the researchers suggested the important work for research as Scaling up Data Mining algorithms for huge data, mining text and automated data cleansing as the major issues discussed with highest priorities [13]. Other issues include dealing with unbalanced data, mining data streams, link and networks. Security in mining and distributed data mining also caught the significance but not to as greater extent. A hotly debated technical issue is whether it is better to set up a relational database structure or a multidimensional one. Finally, there is the issue of price.

Major issues in Web Mining are:

- Web data sets can be very large; it takes ten to hundreds of terabytes to store on the database
- It cannot mine on a single server so it needs large number of server
- Proper organization of hardware and software to mine multi-terabyte data sets
- Limited customization, limited coverage, and limited query interface to individual users
- Automated data cleaning
- Over fitting and Under fitting of data
- Over sampling of data
- Scaling up for high dimensional data
- Mining sequence and time series data
- Difficulty in finding relevant information
- Extracting new knowledge from the web

7. Conclusions and Future Scope

As the web and its usage continue to grow, there is a need to grow the opportunity to analyze web data and extract all manner of useful knowledge from it. In the past few years, we 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. This paper has attempted to provide an up-to-date survey of the rapidly growing area of Web mining. With the growth of Web-based applications, especially electronic commerce, there is significant interest in analyzing Web contents, its structure and usage of data to better understand and apply the knowledge to better serve users. This paper has also discussed about the research issues and challenges in web mining. Several open research issues and drawbacks which exist in the current techniques are also discussed. This study and review would be helpful for the researchers in the domain of web mining.

References:

[1]Wang Jicheng et al.. "Web Mining: Knowledge Discovery on the Web ".
[2]Usama Fayyad et al., "The KDD Process for Extracting Useful Knowledge from Volumes of Data", Communications of the ACM, Vol. 39, No. 11, Nov. 1996, pp. 27-34.
[3]D. Jayalatchumy, Dr. P.Thambidurai, "Web Mining Research Issues and Future Directions –A Survey", OSR Journal of Computer Engineering (IOSR-JCE), Volume 14, Issue 3 (Sep. -Oct. 2013), PP 20-27.
[4] Jaideep Srivastava et al. "Web Mining — Concepts, Applications, and Research Directions".
[5]S. Yadav et al., "Analysis of Web mining applications and beneficial areas", IIUM Engineering Journal, Volume 12, No. 2, 2011.
[6]Soumen Chakraborti, "Data mining for hypertext: A tutorial survey", 2000.
[7]Jaideep Srivastava et al., "Web usage mining: Discovery and applications of usage patterns from web data" published in ACM SIGKDD Explorations, Volume 1, Issue 2, Jan 2000, pp. 12-23.
[8] Shankar et. al., "Web mining in soft computing Framework: Relevance, State of the Art and Future Directions", 2002.
[9]Arijit Abraham, "Business Intelligence from Web Usage Mining", 2003.
[10]Renata Ivancsy, Istvan Vajk, "Frequent Pattern Mining in Web Log Data",2006.
[11] T. Atanasova et. al., "Analysis of the possible application of Data Mining, Text Mining and Web Mining in Business Intelligent System", published in MIPRP, 2010, pp. 1294-1297.

[12] Li Mei and Feng Cheng, "Overview of WEB Mining Technology and Its Application in E-commerce", preceding in 2nd International Conference on Computer Engineering and Technology, Volume 7, IEEE, 2010, pp. 277-280.
[13] HongKui et. al., "Research of Data Mining in Electronic Commerce", in IEEE, 2011, pp. 4323-4326.
[14]Pradnyesh Bhisikar,Prof. Amit Sahu, "Overview on Web Mining and Different Technique for Web Personalisation", International Journal of Engineering Research and Applications (IJERA) Vol. 3, Issue 2, March - April 2013, pp.543-545.