Volume 4, Issue 7, July 2014 ISSN: 2277 128X

International Journal of Advanced Research in Computer Science and Software Engineering

Research Paper

Available online at: www.ijarcsse.com

Web Mining Research Direction and Open Source Tools

M. Vengateshwaran
BE UG Scholar
Department of CSE
Arasu Engineering College, Kumbakonam,
India

E. V. R. M Kalaimani
M. Tech, Ph. D
Professor &Head of Department-CSE
Arasu Engineering College, Kumbakonam,
India

Abstract- This paper aims to focus about web mining research direction,open source tools and their several applications to the commercially supported users.The World Wide Web is a huge, information center for a variety of applications. Web contains a dynamic and rich collection of hyperlink information. It allows Web page access, usage of information and provides numerous sources for data mining. Web mining is a research topic which combines two of the activated research areas: Data Mining and World Wide Web.It give the superficial knowledge and comparison about data mining. This paper describes the current, past and future direction of web mining.Here we introduce online resources for retrieval Information on the web i.e. web content mining, and the discovery of user access patterns from web servers, i.e. web usage mining that improve the data mining drawbacks.Further more cloud mining is a future of web mining.

Keyword- Web Mining, Web Content Mining, Web Structure Mining, Web Usage Mining,Data mining.

I. INTRODUCTION

The term Web Mining is a technique used to crawl through various web resources to collect required information, which enables an individual or a company to promote business, understanding marketing dynamics, new promotions floating on the Internet, etc. There is a growing trend among companies, organizations and individuals alike to gather information through web data mining to utilize that information in their best interest. The Web is changing fast over time and so is the users interaction in the Web suggesting the need to study and develop models for the evolving Web Content, Web Structure and Web Usage. World wide web is a enormous amount of widely dispersed, interconnected, beneficial and dynamic hypertext information. It has used in different needs of us in various stages like communication, business, entertainment and so on. Web data mining is not only focused to gain business information but is also used by various organizational departments to make the right predictions and decisions for things like business development, work flow, production processes and more by going through the business models derived from the data mining.Web data mining technology is opening avenues on not just gathering data but it is also raising a lot of concerns related to data security. There is loads of personal information available on the internet and web data mining had helped to keep the idea of the need to secure that information at the forefront.

II. WEB MINING

Web mining aims to discover useful information or knowledge from the Web hyperlink structure, page content, and usage data. Although Web mining uses many data mining techniques,it is not purely an application of traditional data mining due to the heterogeneity and semi structured or unstructured nature of the Web data. Many new mining tasks and algorithms were invented in the past decade. Based on the primary kinds of data used in the mining process, Web mining tasks can be categorized into three types: Web structure mining, Web content mining and Web usage mining.

Two different approaches were taken in the web mining.

i) **process-centric view**" -defined web mining as a sequence of tasks.
ii) "**data-centric view**" - types of web data that was being used in the mining process.

1. Resource finding: It is used to extract the data from online text resources available on web.
2. Information selection and pre-processing: This process transforms the original retrieved data into information.
3. Generalization:Individual web sites as well as across multiple sites.
4. Analysis: It involves the validation and interpretation of the mined patterns.

III. WEB MINING TAXONONMY

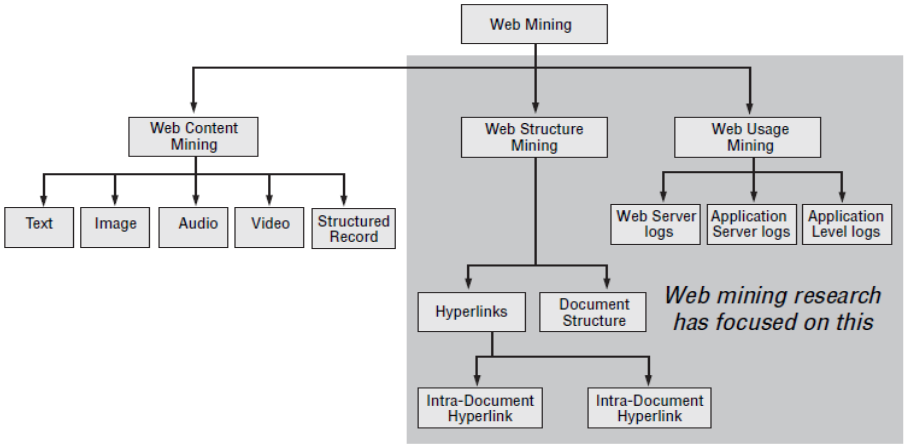

Fig .1 Diagram for Taxonomy

A) Web Content Mining

Web content mining is also known as text mining is generally the second step in Web data 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. content mining provides the results lists to search engines in order of highest relevance to the keywords in the query. Web text mining is very effective when used in relation to a content database dealing with specific topics. For example online universities use a library system to recall articles related to their general areas of study. The ability to conduct Web content mining allows results of search engines to maximize the flow of customer clicks to a Web site, or particular Web pages of the site, to be accessed numerous times in relevance to search queries.

B) Web Structure Mining

Web structure mining, one of three categories of web mining for data, is a tool used to identify the relationship between Web pages linked by information or direct link connection. This structure data is discoverable by the provision of web structure schema through database techniques for Web pages. Structure mining uses minimize two main problems of the World Wide Web due to its vast amount of information. The first problems is irrelevant search results and another one is inability to index the vast amount of information provided on the Web. 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.

C) Web Usage Mining

This type of web mining allows the collection of Web access information for Web pages. This usage data provides the paths leading to accessed Web pages. This information is often gathered automatically into access logs via the Web server. CGI scripts offer other useful information such as referrer logs, user subscription information and survey logs. This category is important to the overall use of data mining for companies and their internet/ intranet based applications and information access.

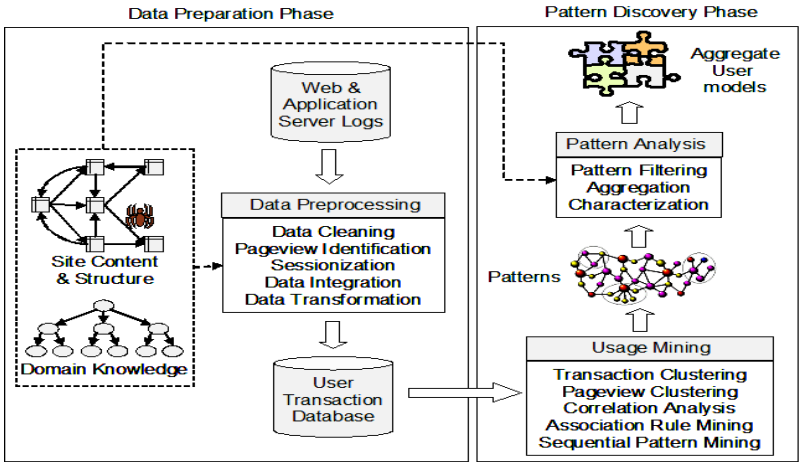

Fig.2 Process of usage mining

Analysis of this pertinent information will help companies to develop promotions that are more effective, internet accessibility, inter-company communication and structure, and productive marketing skills through web usage mining.

	Web Mining			
	Web Content Mining		Web Structure Mining	Web Usage Mining
	IR view	DB View		
View of Data	-Unstructured -Structured	-Semi Structured -Web Site as DB	-Link Structure	-Interactivity
Main Data	- Text documents -Hypertext documents	-Hypertext documents	-Link Structure	-Server Logs -Browser Logs
Representation	-Bag of words, n-gram Terms, -phrases, Concepts or ontology -Relational	-Edge labeled Graph, -Relational	-Graph	-Relational Table -Graph
Method	-Machine Learning -Statistical (including NLP)	-Proprietary algorithms -Association rules	-Proprietary algorithms	-Machine Learning -Statistical -Association rules
Application Categories	-Categorization -Clustering -Finding extract rules -Finding patterns in text	-Finding frequent sub structures -Web site schema discovery	-Categorization -Clustering	-Site Construction -adaptation and management -Marketing, -User Modeling

Fig.3 categories of web mining

IV. WEB MINING RESEARCH COMMUNITY

A) Ranking Metrics-for Page Quality and Relevance

Searching the web involves two main steps are there Extracting the pages relevant to a query and ranking them according to their quality. Ranking is important as it helps the user look for "quality" pages that are relevant to the query. Different metrics have been proposed to rank web pages according to their quality. We have two type of prominent ones Pagerank and Hubs and Authorities.

B) Robot Detection and Filtering

It provide a human and non human web behavior .Web robots are software programs that automatically traverse the hyperlink structure of the web to locate and retrieve information. The importance of separating robot behavior from human behavior prior to building user behavior models has been illustrated. First, e-commerce retailers are particularly concerned about the unauthorized deployment of robots for gathering business intelligence at their web sites. Another,web robots tend to consume considerable network bandwidth at the expense of other users. Sessions due to web robots also make it difficult to perform click-stream analysis effectively on the web data. Conventional techniques for detecting web robots are based on identifying the IP address and user agent of the web clients. While these techniques are applicable to many well-known robots, they are not sufficient to detect camouflaged and previously unknown robots.classification based approach that uses the navigational patterns in click-stream data to determine if it is due to a robot. Experimental results have shown that highly accurate classification models can be built using this approach. Furthermore, these models are able to discover many camouflaged and previously unidentified robots.

C) Information Scent-Applying Foraging Theory to Browsing Behavior

It is used for snippets of information present around the links in a page as a "scent" to evaluate the quality of content of the page it points to, and the cost of accessing such a page.The key idea is to model a user at a given page as "foraging" for information,and following a link with a stronger "scent." The "scent" of a path depends on how likely it is to lead the user to relevant information, and is determined by a network flow algorithm called spreading activation. The snippets,graphics, and other information around a link are called "proximal cues." The user's desired information need is expressed as a weighted keyword vector.The similarity between the proximal cues and the user's information need is computed as "proximal scent." With the proximal cues from all the links and the user's information need vector, a "proximal scent matrix" is generated.Each element in the matrix reflects the extent of similarity between
the link's proximal cues and the user's information need. If enough information is not available around the link, a "distal scent" is computed with the information about the link described by the contents of the pages it points to. The proximal scent and the distal scent are then combined to give the scent matrix.The probability that a user would follow a link is then decided by the scent or the value of the element in the scent matrix.

D) User Profiles- Understanding How Users Behave

The web has taken user profiling to new levels. For example, in a "brick-and mortar" store, data collection happens only at the checkout counter, usually called the "point-of-sale." This provides information only about the final outcome of a complex human decision making process, with no direct information about the process itself. In an on-line store, the complete click-stream is recorded, which provides a detailed record of every action taken by the user, providing a much more detailed insight into the decision making process.Adding such behavioral information to other kinds of information about users, for example demographic, psychographic, and so on, allows a comprehensive user profile to be built, which can be used for many different purposes.While most organizations build profiles of user behavior limited to visits to their own sites, there are successful examples of building web-wide behavioral profiles such as Alexa Research6 and DoubleClick7. These approaches require browser cookies of some sort, and can provide a fairly detailed view of a user's browsing behavior across the web.

E) Interestingness Measures- When Multiple Sources Provide Conflicting Evidence

One of the significant impacts of publishing on the web has been the close interaction now possible between authors and their readers. In the preweb era,a reader's level of interest in published material had to be inferred from indirect measures such as buying and borrowing, library checkout and renewal, opinion surveys, and in rare cases feedback on the content. For material published on the web it is possible to track the click-stream of a reader to observe the exact path taken through on-line published material.We can measure times spent on each page, the specific link taken to arrive at a page and to leave it, etc. Much more accurate inferences about readers' interest in content can be drawn from these observations. Mining the user click-stream for user behavior, and using it to adapt the "look-and-feel" of a site to a reader's needs was first proposed.While the usage data of any portion of a web site can be analyzed, the most significant, and thus "interesting," is the one where the usage pattern differs significantly from the link structure. This is so because the readers' behavior,reflected by web usage, is very different from what the author would like it to be, reflected by the structure created by the author. Treating knowledge extracted from structure data and usage data as evidence from independent sources, and combining them in an evidential reasoning framework to develop measures for interestingness.

F) Preprocessing-MakingWeb Data Suitable for Mining

The Preprocessing of web data to make it suitable for mining was identified as one of the key issues for web mining. A significant amount of work has been done in this area for web usage data, including user identification and session creation, robot detection and filtering and extracting usage path patterns.dissertation provides a comprehensive overview of the work in web usage data preprocessing.Preprocessing of web structure data, especially link information, has been carried out for some applications, the most notable being Google style web search.

G) Online Bibiliometrics

Web having become the fastest growing and most up to date source of information, the research community has found it extremely useful to have online repositories of publications. online makes them more easily accessible than offline. Such online repositories not only keep the researchers updated on work carried out at different centers, but also makes the interaction and exchange of information much easier.With such information stored in the web, it becomes easier to point to the most frequent papers that are cited for a topic and also related papers that have been published earlier or later than a given paper.it helping researchers to explore new areas. Fundamental web mining techniques are applied to improve the search and categorization of research papers, and citing related articles.

H) Visualization of the World Wide Web

Mining web data provides a lot of information which can be better understood with visualization tools.It is possible with pure textual representation. Hence, there is a need to develop tools that provide a graphical interface that aids in visualizing results of web mining.Analyzing the web log data with visualization tools has evoked a lot of interest in the research community.An interactive web log visualization tool that is designed to display the user browsing pattern on the web site at a global level, and then display each browsing path on the pattern displayed earlier in an incremental manner. The support of each traversal is represented by the thickness of the edge between the pages. Such a tool is very useful in analyzing user behavior and improving web sites.

IV. PROMINENT APPLICATIONS

Past few years has led to the web applications being developed at a much faster rate in the industry than research in web related technologies. Many of these are based on the use of web mining concepts, even though the organizations that developed these applications.

A) 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.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 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.Google's web directory provides a fast and easy way to search within a certain topic or related topics.The advertising program introduced by Google targets users by providing advertisements that are relevant to a search query.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.

B) Web-Wide Tracking

"Web-wide tracking," is 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. Example- DoubleClick Inc.

C) Understanding Web Communities-AOL

It is 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. AOL provides them with useful information and services. Over time these communities have grown to be well-visited waterholes 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 TwentySomethings."

D) EBay

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. E-bay has detailed data on bid history,participant rating,bid data,usage data. 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. eBay is now using web mining techniques to analyze bidding behavior to determine if a bid is fraudulent .Recent efforts are geared towards understanding participants' bidding behaviors/patterns to create a more efficient auction market.

E) Personalized Portal for the Web—MyYahoo

Yahoo is an one of the search engine.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. 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.

F) V-TAG Web Mining server-cannotate Technologies:

The web mining server supports information agents that monitor,extract and summarize information from web sources.It is easily to set up graphical user interface.Automation of tracking and summarizing helps businesses and enterprises to analyse the various processes easily

V. FURTHER RESEARCH DIRECTION

Web and its usage grows, it will continue to generate ever more content,structure, and usage data, and the value of web mining will keep increasing.To develop a web mining technologies that will enable this value to be realized.

A) Process Mining

Mining of market basket data, collected at the point-of-sale in any store, has been one of the visible successes of data mining. However, this data provides only the end result of the process, and that too decisions that ended up in product purchase. Click-stream data provides the opportunity for a detailed look at the decision making process itself, and knowledge extracted from it can be used for optimizing, influencing the process, etc.it has conclusively proven the value of process information in understanding users' behavior in traditional shops. Research needs to be carried out in (1) extracting process models from usage data, (2) understanding how different parts of the process model impact various web metrics of interest, and (3) how the process models change in response to various changes that are made, i.e. changing stimuli to the user. Example-online shopping.

B) Web Mining and Privacy

There are many benefits to be gained from web mining, a clear drawback is the potential for severe violations of privacy. Public attitude towards privacy seems to be almost schizophrenic, i.e. people say one thing and do quite the opposite. For example, famous cases like those involving Amazon and Doubleclick seem to indicate that people value their privacy, while experience at major e-commerce portals shows that over 97% of all people accept cookies with no problems, and most of them actually like the personalization features that are provided based on it.The research issue generated by this attitude is the need to develop approaches, methodologies and tools that can be used to verify and validate that a web service is indeed using user's information in a manner consistent with its stated policies.

C) Fraud and Threat Analysis

The anonymity provided by the web has led to a significant increase in attempted fraud, from unauthorized use of individual credit cards to hacking into credit card databases for blackmail purposes . Example is auction fraud, which has been increasing on popular sites like eBay.Since all these frauds are being perpetrated through the internet, web mining is the perfect analysis technique for detecting and preventing them. Research issues include developing techniques to recognize known frauds, characterize them and recognize emerging frauds. The issues in cyber threat analysis and intrusion detection are quite similar in nature.

D) Web Services Performance Optimization

These services over the web continue to grow , there will be a continuing need to make them robust, scalable and efficient. Web mining can be applied to better understand the behavior of these services, and the knowledge extracted can be useful for various kinds of optimizations. The successful application of web mining for predictive prefetching of pages by a browser has been demonstrated. It is necessary to do analysis of the web logs for web services performance optimization. Research is needed in developing web mining techniques to improve various other aspects of web services.

VII. WEB MINING PROS AND CONS

A) PROS

Web usage mining essentially has many advantages which makes this technology attractive to corporations including the government agencies. This technology has enabled e-commerce to do personalized marketing, which eventually results in higher trade volume.

B) CONS

This technology when used on data of personal nature might cause concerns. The most criticized ethical issue involving web usage mining is the invasion of privacy. Privacy is considered lost when information concerning an individual is obtained, used, or disseminated, especially if this occurs without their knowledge or consent.

VIII. OPEN SOURCE TOOLS

Web data mining tools vendors **DMOZ**,**Kdnuggets** also has a list is categorized commercial or free .Two very well known industry vendors are SAS and **COGNOS**. Bixo is an open source web mining toolkit that runs as a series of Cascading pipes on top of Hadoop. By building a customized Cascading pipe assembly, you can quickly create specialized web mining applications that are optimized for a particular use case Bixo is an open source project released under the Apache License, Version 2.0.

Bixo Tool

Bixo is a major tool of web mining.Bixo consists of a number of Cascading Operations and Subassemblies, which can be combined to form a data processing workflow that (typically) starts with a set of URLs to be fetched, and ends with some results extracted from parsed HTML pages.

Fetch Subassembly	component where the heavy lifting is done.
Parse Subassembly	used to process the fetched content. It uses Tika to extracting text from various formats.

IX. WEB SIZE

- Some 80% of Web pages are in English
- About 30% of domains are in .com domain

  GG=Google, ATW=AllTheWeb, INK=Inktomi, TMA=Teoma, AV=AltaVista

  Number of pages

  -Technically ,infinite and much duplication(30-40%)

  -Best estimate of "unique" static HTML pages comes from search engine claims

  -Google= 8 billions and yahoo= 10 billions
- In Net craft survey says that 72 million of sites are there.

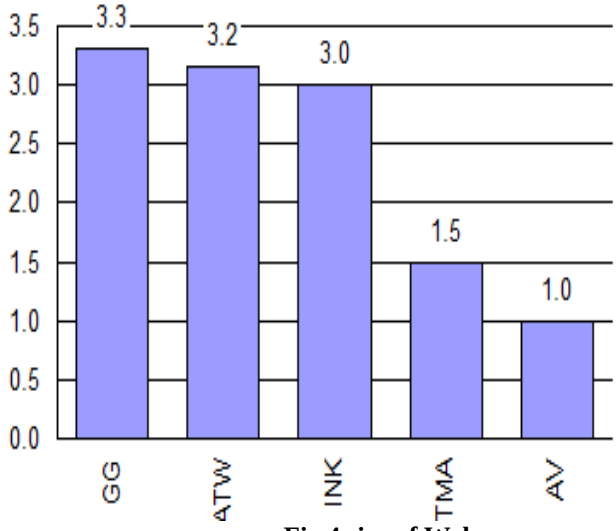

Fig.4 size of Web

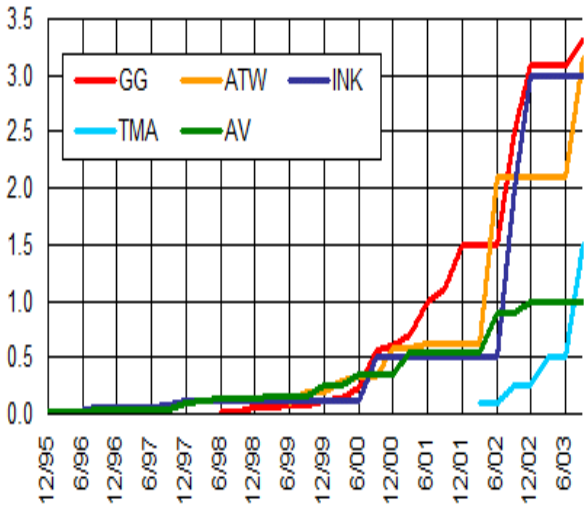

Fig.5 Size Trends

X. BENIFTS OF WEB MINING

- personalization
- collaborative filtering,
- enhanced customer support,
- product and service strategy in marketing
- marketing and fraud detection

XI. APPLICATION OF WEB MINING

- E-commerce
- Information retrieval search on the web
- Network management

XII. ISSUES IN WEB MINING

- It contain very large data set on Web
- It cannot be mine on single server
- How to organize hardware and software to mine in multi-tera byte data sets?

XIII. CONCLUSION

Today Web mining growth is continuously to increase.Web mining is one of the most important applications of data mining. It is having its own benefits and successful applications with which we can overcome the problems or difficulties faced in data mining. usage of the internet in the present day is growing in faster rate, the personalization process of the web mining provides us a great opportunity of maximizing the efficient usage of the internet.web mining in future growing online shopping activities, e-services industry and e-commerce are important domains. For Counter terrorism, many challenges are needed yet to be addressed to make data mining to become a useful tool. Research is to be carried out is to explore the semantic Web structure in the Web for getting benefits in many areas of the industries. Web mining enables us to screen specific data through content mining, to discover the structural summary of web sites through structure mining and to predict the behaviour and interaction of the surfers' with the web through usage mining and encompass a broad range of issues. Towards this goal, in this paper, we proposed a definition of web mining, research direction and benefits of web mining, and web mining taxonomies. We identified some of the issues and problems in this area that may require further research and development.Web mining is applied to various fields E-Commerce, Information filtering , Fraud detection Education and research.

REFERENCE

[1] Jaideep Srivastava, Prasanna Desikan, Vipin Kumar, ―Web Mining - Concepts, Applications.

[2] Yan Li, Boqin Feng, Qinjiao Mao, "Research on Path Completion Technique in Web Usage Mining", Computer Science and Computational Technology, 2008. ISCSCT '08.

[3] Ricardo Baeza-Yates. Web usage mining in search engines. In Web Mining: Applications and Techniques, Anthony Scime, editor.Idea Group, 2004.

[4] N. Barsagade, Web usage mining and pattern discovery: A survey paper, Computer Science and Engineering Dept., CSE Tech Report 8331 (Southern Methodist University,Dallas, Texas, USA, 2003).

[5] R. Chau, C. Yeh and K. Smith, Personalized multilingual web content mining, KES (2004)

[6] P. Kolari and A. Joshi, Web mining: Research and practice, Comput. Sci. Eng.July/August (2004) [7]. B. Liu and K. Chang, Editorial: Special issue on web content mining, SIGKDD Explorations (2004)

[8] Semantic Web Mining:State of the art and future directions‖ Web Semantics: Science, Services and Agents on the World Wide Web, Volume 4, Issue 2, June 2006

[9] Gerd Stumme ,Andreas Hotho , Bettina Berendt, "Semantic Web Mining", Journal of Web Semantics: Science, Services and Agents on the World Wide Web, Vol.4 Issue 2, June, 2006.

[10] J. Srivastava, R. Cooley, M. Deshpande, and P.N. Tan, Web Usage Mining: Discovery and Applications of Usage Patterns from Web Data. SIGKDD Explorations, vol. I, no. 2, pp. 12-23, 2000.

