STRATEGY AND IMPLEMENTATION OF WEB MINING TOOLS

Danish Ahamad
College of Science and Arts, Sajir, Shaqra University, KSA
danish@su.edu.sa,

Md Mobin Akhtar
College of Computing and Information Technology, Shaqra, Shaqra University, KSA
jmi.mobin@su.edu.sa

Dr. Adil Mahmoud Mohamed Mahmoud
College of Science and Arts, Sajir, Shaqra University, KSA
adilmahmoud@su.edu.sa

Manuscript History
Number: **IJIRAE/RS/Vol.04/Issue12/DCAE10081**
DOI: 10.26562/IJIRAE.2017.DCAE10081
Received: 26, November 2017
Final Correction: 05, December 2017
Final Accepted: 15, December 2017
Published: **December 2017**
Citation: Ahamad, D., Mahmoud, D. A. M. M. & Akhtar, M. M. (2017). STRATEGY AND IMPLEMENTATION OF WEB MINING TOOLS, International Journal of Innovative Research in Advanced Engineering, Volume IV, 01-07. doi: 10.26562/IJIRAE.2017.DCAE10081
Editor: Dr.A.Arul L.S, Chief Editor, IJIRAE, AM Publications, India

Abstract— In the current development, millions of clients are accessing daily the internet and World Wide Web (WWW) to search the information and achieve their necessities. Web mining is a technique to automatic discovers and Extract information from www. Websites are a common stage to discussion the information between users. Web mining is one of the applications of Data mining techniques for extracting information from web data. The area of web mining is web content mining, web usage mining and web structure mining. These three category focus on Knowledge discovery from web. Web content mining involves technique for summarization, classification, clustering and the process of extracting or discovering useful information web pages, it includes image, audio, video and metadata. Web usage mining is the process of extracting information from web server logs. Web structure mining it is the process of using graph theory to analyse the node and connection structure of a website and deals with the hyperlink structure of web. Web mining is a part of data mining which relates to various research communities such as information retrieval, database management systems and Artificial intelligence.

Keywords—Web Mining; Mining Calcification; Web Structure; HITS Algorithm; Page Rank Algorithm; Vulnerability;

I. INTRODUCTION

Web mining is the applications of data mining techniques to discover pattern from the World Wide Web. It has quickly become one of the most important areas in Computer and Information Sciences because of its direct applications in e-commerce, e-CRM, Web analytics, information retrieval and filtering, and Web information systems. It is the application of data mining techniques to automatically discover and to extract knowledge from web data, including web documents, hyperlinks between documents, us-age logs of web sites, etc. Some of the data mining techniques applied in web mining are association rule mining, clustering, classification, frequent item set. Some of the sub tasks of web mining are resource finding, information selection and pre-processing, generalization and analysis.

Web mining is the expenditure of data mining methods to automatically discover and extract information from where documents and services. Although web mining practices many data mining methods, it is not purely an application of traditional data mining due to the heterogeneity and semi structured or unstructured nature of the web data. It has the collection of text, images, videos and other form of data. To handle these huge volumes of data and extract meaningful information and knowledge, there is a need to develop some new techniques and tools. Data mining is a process of extracting useful information from the large data set, when it is applied to the web content is called a web mining. The aim of resource finding is to extract the information from the web documents. During the second task, extract/select the relevant information and filter the irrelevant information from the actual data. Generalization is used to discovery the general patterns by applying machine learning or data mining techniques. During analysis, the patterns are analysed and verified. The main part information becomes very difficult for the users to find, extract, filter or evaluate the relevant information. This question raises the requirement of some technique that can solve these challenges. Web mining can be easily executed with the help of other areas like Database (DB), Information retrieval (IR), Natural Language Processing (NLP), and Machine Learning etc. The following challenges in Web Mining are: Information extracted. Web is huge. Web pages are semi structured. Conclusion of knowledge from information extracted.

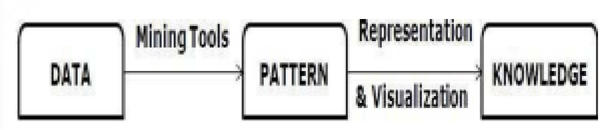

Fig.1 Route of Web Mining

Type	Algorithm
1.Web structure mining	▪ Link analysis algorithms ▪ Hits (hyper-link induced topic search) ▪ Page rank ▪ Weighted page rank ▪ Topic sensitive PageRank algorithm
2. Web usage mining	▪ Clustering ▪ k-mean algorithm ▪ Latent semantic analysis ▪ Prefix Span Algorithm ▪ One pass SI and one pass AIISI Algorithm
3. Web content mining	▪ Correlation algorithm for relevance ranking ▪ Cluster hierarchy construction algorithm ▪ Weighted Page Content Rank ▪ fuzzy c-mean algorithm

II. TAXONOMY OF WEB MINING

Web is a collection of inter-related files on one or more Web servers. Web mining is the application of data mining techniques to extract knowledge from Web data. Web mining is an iterative process for fetching the facts from web data.

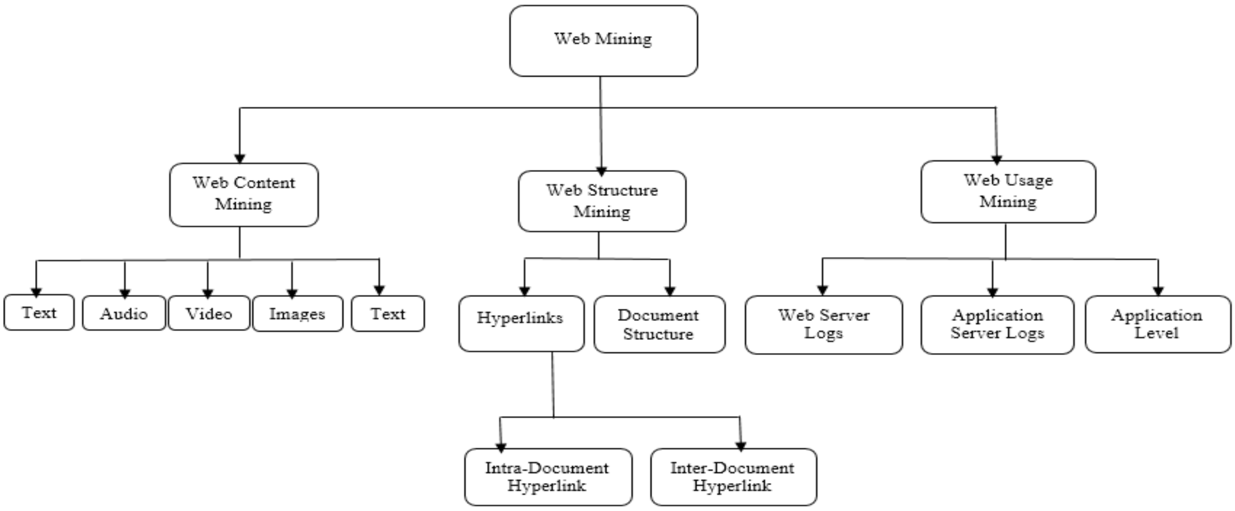

Fig.2 Taxonomy of web mining

III. WEB CONTENT MINING

Web Content Extract "snippets" from a Web document that represents the Web Document. Web Content Mining is the process of extracting useful information from the contents of Web documents Web Content mining refers to the discovery of useful information from the contents of the webpage using text mining techniques. It may consist of text, images, audio, video, or structured records such as lists and tables. It includes extraction of structured data/information from web pages, identification, similarity and integration of data's with similar meaning, view extraction from online sources, and concept hierarchy.

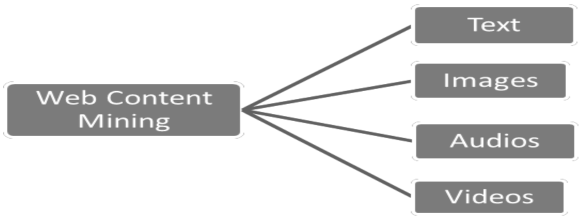

Fig.3 web Content mining

Web Content Mining Approaches: Two approaches used in web content mining are Agent based approach and database approach. The three types of agents are intelligent search agents, Information filtering/Categorizing agent, and personalized web agents. Intelligent Search agents automatically searches for information according to a particular query using domain characteristics and user profiles. Information agents used number of techniques to filter data according to the predefine information. Adapted web agents learn user preferences and discovers documents related to those user profiles.

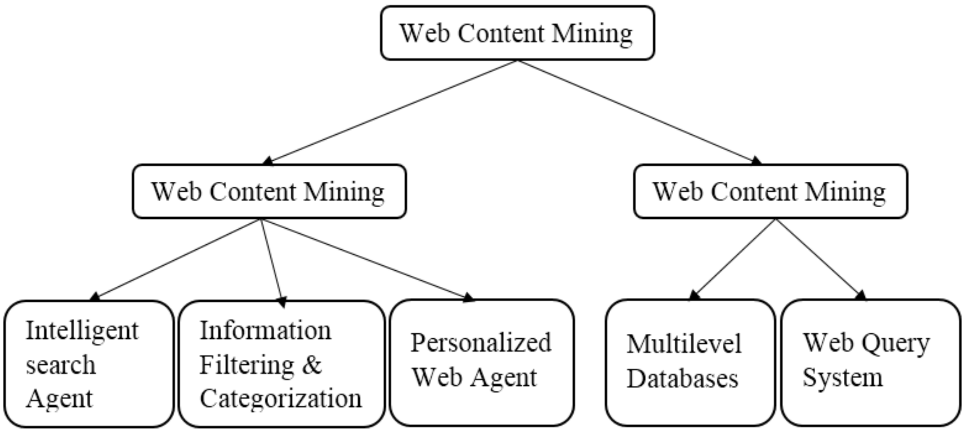

Fig.4 Type of Web Content Mining

IV. WEB CONTENT MINING TOOLS

The tools related to web content mining is available which can extract useful information from web pages. There are different type's tools available for web content mining are

A. **Web Info Extractor:** These are the resources for data mining, extracting Web content, and Web content analysis. Extract structured or unstructured data from Web page, reform into local file or save to database Features:
    - No need to learn boring and complex template rules and it is easy to define extract tool.
    - Extract tabular as well as unstructured data to file or database.
    - Monitor Web pages and extract new content when update.
    - Can deal with text, image and other link file
    - Can deal with Web page in all language

B. **Screen-scraper:** Screen-scraping is a tool for extracting information from web sites. It can be used for searching a database, SQL server or SQL database, which Interfaces with the software, to achieve the content mining requirements. The programming languages like Java, .NET, PHP, Visual Basic and Active Server Pages (ASP) can also be used to access screen scraper.

C. **Octoparse:** Octoparse is a powerful web scraping tools that can grab open data from almost all the websites. The features of Octoparse enable you to work with dynamic unstructured data by just clicking on single data points and Octoparse will generate efficient code to extract data automatically.

Octoparse is client-side software written in .NET for extracting information from websites. It is cloud based web crawling and web scraping software that helps to extract any web data without coding in real time. It can collect data from websites and sort the data into database.

D. **Web Content Extractor:** web content extractor (WCE) in order to extract data from any given website. It is a powerful and easy to use data extraction tool for Web scraping, data mining or data extraction from the Internet. This tool extracts the product data from online shopping, stock market, financial, song or movie information, helpful for extracting news from different news sites for reporter.

E. **Scrapy:** Web scraping is a very powerful tool to learn for any data professional. With web scraping the entire internet becomes your database. Scrapy is a free and open source software written in Python for extracting data from websites. It is application framework for extracting structured and crawling data used for applications like data mining and information processing.

V. WEB STRUCTURE MINING

Web structure mining is otherwise called as link mining. It is to deal with structure of hyperlink within web pages itself. Based on the hyperlinks, web structure mining will classify the web pages and generate the information. The structure of a typical wave graph consists of web pages as nodes and hyperlink as edges connecting between two related graphs. Web structure mining is the process of extracting knowledge from the interconnection of hypertext document in the World Wide Web (WWW).

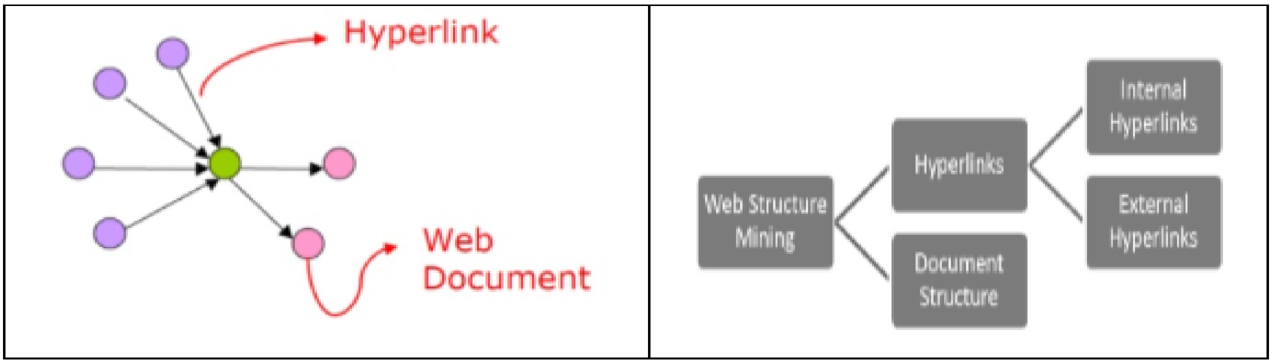

Fig.5 Web Structure Mining

The goal of web structure mining is to generate structural summary about web pages and web sites. It shows the relationship between the user and the web. It discovers the link structure of hyperlinks at the inter document level. It also helps in discovering the structure of document which is used in revealing the structure the structure of web pages and it's possible to compare the web page schemes. Web Structure Mining is also Known as link Mining.

VI. WEB STRUCTURE MINING TOOLS

It is a process to discover the relationship between webpages linked by information or direct link connection. The tools related to Web structure mining is a process to discover the relationship between web pages linked by information. There are different types of tools available for web structure mining are:

- HITS is applied on a subgraph after a search is done on the complete graph
- Uses hubs and authorities to define a recursive relationship between web pages.
- An Authority is a page that many hubs link to
- A hub is a page that links to many authorities

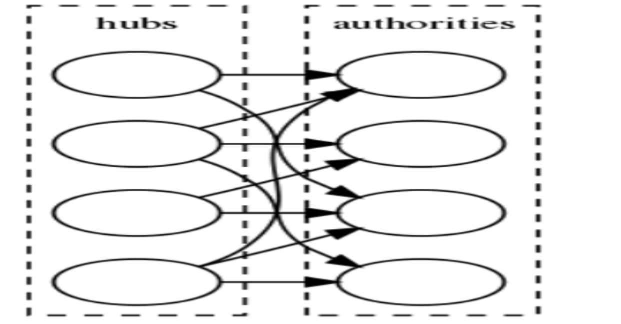

Fig.6 Hyperlink induced topic search

A. **HITS Algorithm:** Hyperlink induced topic search (HITS: also known as bus and authority's) is a link analysis algorithm that rates web pages. The step in HITS algorithm is to retrieve the most relevant pages. This set is called as root set can be obtained by taking top pages and base set is generated by supplementing the root set with all web pages.

- A good hub represented a page that pointed to many other pages.
- A good authority represented a page that was linked by many different hub.

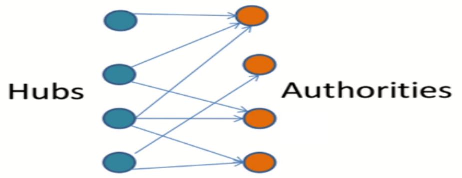

Fig.7 HITS: Hubs and Authorities

Hyperlink induced topic search (HITS)

For Web graph G is = (V, E) to become a defined for nodes p, q belongs to V where Xp is the authority score and Yq is the hub Score.

G=(V,E) Define Nodes q,q ∈ V

$$X_p = \sum y_p$$

$$Y_p = \overset{(p,q)\in E}{\underset{(p,q)\in E}{\sum}} X_q$$

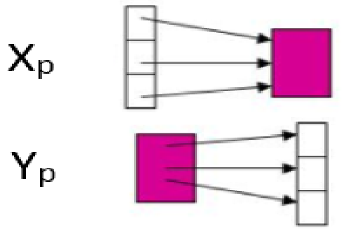

Fig.8 Hyperlink induced topic search (HITS) Algorithm

B. PageRank Algorithm: PageRank is a function that assigns a real number to each page in the Web (or at least to that portion of the Web that has been crawled and its links discovered).It is a link analysis algorithm and works by counting the number and quality of links to a web page. Think of the Web as a directed graph, where pages are the nodes, and there is an arc from page p1 to page p2 if there are one or more links from p1 to p2.Figure below is an example of a tiny version of the Web, where there are only four pages.

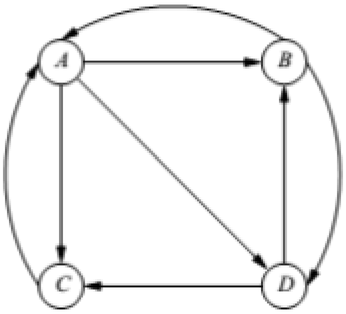

Fig.9 A Hypnotical example of the Web

Page A has links to each of the other three pages; page B has links to A and D only; page C has a link only to A, and page D has links to B and C only. Suppose a random surfer starts at page A in Fig.9. There are links to B, C, and D, so this surfer will next be at each of those pages with probability 1/3, and has zero probability of being at A. A random surfer at B has, at the next step, probability 1/2 of being at A, 1/2 of being at D, and 0 of being at B or C. PageRank of

$$\text{Site} = \sum \frac{\textit{PageRank of inbound link}}{\textit{Number of link on that page}}$$

PageRank is a model of user behaviour where surfer clicks on the link at random with no regard towards content.

$$PR(A)= (1-d) +d\left[\frac{PR(Ti)}{C(Ti)} + \ldots\ldots\ldots\ldots\ldots\ldots + \frac{PR(Tn)}{C(Tn)}\right]$$

Where PR(A) is a PageRank of page A, PR(Ti) is a PageRank of pages Ti which link to page A, C(Ti) is a Number of Outbound link on page Ti and **d** is the damping factor can be set between 0 ,1.

VII. WEB USAGE MINING

Web usage mining is based on the techniques that could predict the pattern of the user while the user interacts with web. It is otherwise called as web log mining. It collects data from web log records to discover the patterns of web pages. Web log records are unformatted text file which contain data like User name, date, time, IP address, status code etc. whenever the user interacts with website, the information are recorded and maintained in web servers. Web logs maintain data like user browsing history. Application logs business transaction and are stored in application server.

Web usage mining consists of three phases

1. Pre-processing, 2. Pattern discovery 3. Pattern analysis

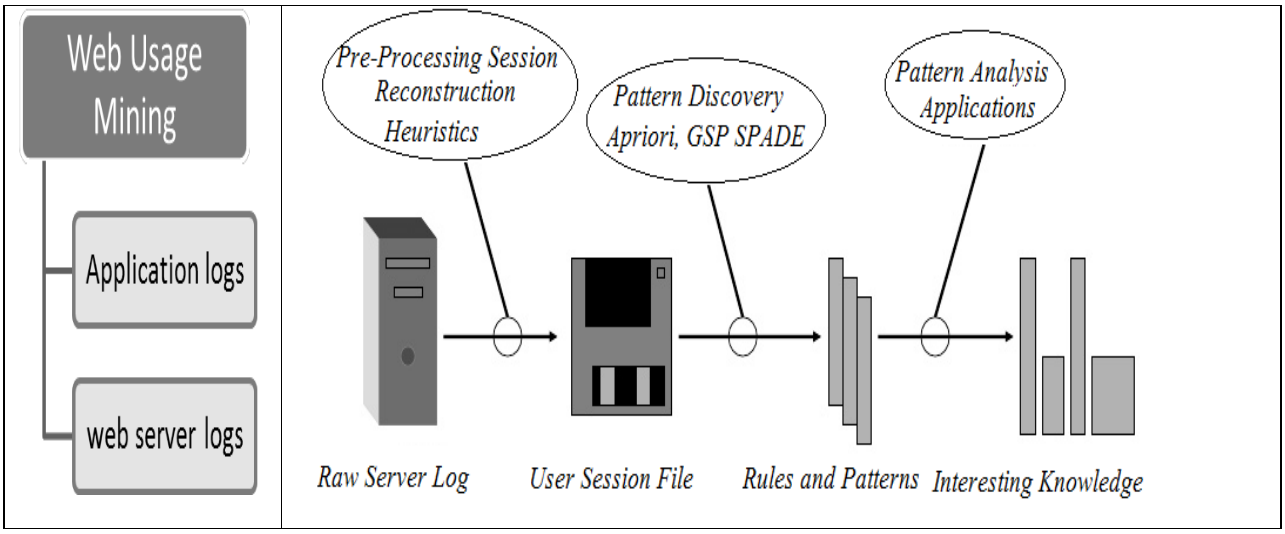

Fig.10 Phases of Web Usage Mining

In web usage mining the first step is pre-processing, the noisy and useless data in web usage log file is cleaned and transformed so that the size can be reduced. Second step, pattern discovery the cleaned and transformed log file is used to discover patterns. Third step, Pattern Analysis in which discovered patterns are further analysed to generate more useful and related information to the user.

A. Web Usage Mining Tools

1) **Oracle Data Mining:** Oracle Data Mining is implemented in the Oracle Database kernel, and mining models are first class database objects. Oracle Data Mining processes use built-in features of Oracle Database to maximize scalability and make efficient use of system resources. The functions of Oracle Data mining can mine data tables, schema, transactional data, structured and unstructured data.
2) **Tableau:** Tableau is one of the commercial intellect tool for exploring the data. It permits user to create and transform data into interactive and variations called dashboards. Data will be represented by graphs and charts. Tableau is used by businesses, researches and many government organizations for visually analysing the data.
3) **Speed Tracer:** Speed Tracer is one of the web usage mining and analysing tool. Speed Tracer tool helps to analyse and debug critical issues in web applications. It is a part of Google web Toolkit. It uses the information like IP address, Timestamp, URL address and session identification.

VIII. CONCLUSION

This paper characterize about web structure mining, web content mining, and web usage mining including its Structure. Internet and websites provide opulent platform for searching the information. Most of the websites are complex and larger in their size and structure. Therefore, it is mandatory to develop tools for websites. The importance of web mining continues to increase due to the increasing tendency of web documents. The mining of web data still be present as a challenging research problem in the future. Because the web documents possess numerous file formats along with its knowledge discovery process. There are many concepts available in Web Mining but this paper tried to expose the Web content mining strategy and explore some of the techniques, tools in Web Content mining.

REFERENCES

1. J. M. Kleinberg. Authoritative sources in a hyperlinked environment. In Proc. of ACM- SIAM Symposium on Discrete Algorithms, pages 668–677, 1998.
2. Cooley, R.; Mobasher, B.; Srivastava, J.; “Web mining: information and pattern discovery on the World Wide Web”. In Proceedings of Ninth IEEE International Conference. pp. 558 – 567, 3-8 Nov. 1997.
3. J. Srivastava, R. Cooley, M. Deshpande, Pag-Ning Tan, “Web Usage Mining: Discovery and Applications of Usage Patterns from WebData” in proceedings of ACMSIGKDD Explorations NewsletterVol.1Issue2, January 2000.
4. Johnson, F., Gupta, S.K., Web Content Minings Techniques: A Survey, International Journal of Computer Application. Volume 47 – No.11, p44, June (2012).
5. Weiming Yang, 2016, “An Improved HITS Algorithm Based on Anal ysis of Web Page Links and Web Content Similarity”, International Conference on Cyberworlds, IEEE, pp.147-150.

6. Ashish Jain, Rajeev Sharma, Gireesh Dixit and Varsha Tomar,2013,"Page Ranking Algorithms in Web Mining, Limitations of Existing methods and a New Method for Indexing Web Pages" IEEE International Conference on Communication Systems and Network Technologies.
7. Lya Hulliyyatus Suadaa, 2014, "A Survey on Web Usage Mining Techniques and Applications", IEEE International Conference on Information Technology Systems and Innovation, PP 24-27,ISBN: 978-1-4799-6526-7.
8. P. Ravi Kumar and Ashutosh Kumar Singh —Web Structure Mining: Exploring Hyperlinks and Algorithms for Information Retrieval in American Journal of Applied Sciences.
9. Theint Theint Aye, 2011, "Web Log Cleaning for Mining of Web Usage Patterns", Proceedings of 3rd International Conference on Computer Research and Development, Volume 2, pp. 490 - 494.
10. A. Bhargav and M. Bhargav, 2014, "Pattern discovery and users classification through web usage mining," in IEEE International Conference on Control, Instrumentation, Communication and Computational Technologies (ICCICCT), pp. 632–636.
11. Dilip Singh Sisodia, Shrish Verma, 2012, "Web Usage Pattern Anal ysis through Web Logs: A Review", Proceedings of 2012 International Joint Conference on Computer Science and Software Engineering, pp. 49 - 53.
12. Arvind Kumar Sharma, P.C. Gupta, 2012, "Study & Analysis of Web Content Mining Tools to Improve Techniques of Web Data Mining", in International Journal of Advanced Research in Computer Engineering &Technology (IJARCET) Volume 1, Issue 8.
13. Chhavi, R 2012, 'A Study of Web Usage Mining Research Tool', International Journal of Advanced Networking andApplications, vol. 3, no.6, pp.1422-1429.
14. [14] Aggarwal C, Wolf JL, Yu PS. caching on the World Wide Web. IEEE Trans Knowledge Data Engg 1999;11(1): 94–107.