GENERAL WEB KNOWLEDGE MINING FRAMEWORK

B. Madasamy
Research Scholar & Assistant Professor, Department of Computer Applications,
Agni College of Technology, Anna University
Chennai, India.
bmadasamy@gmail.com

Dr. J. Jebmalar Tamilselvi
Assistant Professor, Department of Computer Applications,
Jaya Engineering College, Anna University
Chennai, India
jjebamalar@gmail.com

***Abstract* — Mining the web is defined as discovering knowledge from hypertext and World Wide Web. The World Wide Web is one of the longest rising areas of intelligence gathering. Now a day there are billions of web pages, HTML archive accessible via the internet, and the number is still increasing. However, considering the inspiring diversity of the web, retrieving of interestingness web based content has become a very complex task. The large amount of data heterogeneity, complex format, high dimensional data and lack of structure of web, knowledge mining is a challenging task.**

In this paper, it is proposed to introduce a new framework generated to handle unstructured complex data. This web knowledge mining expertise brings forward a kind of XML-based distributed data mining architecture. Based on the research of web knowledge mining, XML is used to create well structured data. Web knowledge mining framework attempts to determine useful knowledge from derived data, complex format, and high dimensional data obtained from the interactions of the users through the Web.

Keywords- High dimensional data, Knowledge mining, XML.

I Introduction

Data mining is the process which automates the extraction of predictive information, discovers the interesting knowledge from large amounts of data stored in databases, data warehouses or other information repositories.

The WWW continues to grow at a wonderful rate as an information gateway and as a medium for conducting business. Web knowledge mining has been widely used in the past for analyzing huge collections of data, and is currently being applied to a variety of domains. Based on several research studies web mining can be broadly classified into three domains: content, structure and usage mining. Web content mining is the process of extracting knowledge from the content of the actual web documents (text, content, multimedia, etc.). Web structure mining is targeting knowledge from the Web structure, hyperlink references and so on. Web usage mining attempts to discover useful knowledge from the secondary data obtained from the interactions of the users with the Web. Brief classifications of web mining are shown in fig.1 [2, 3, and 4]

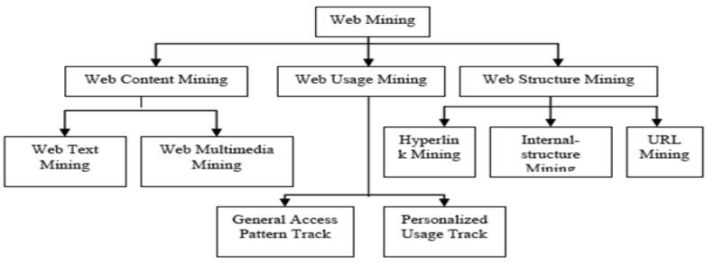

Figure 1 Classification of web knowledge mining framework

Web Knowledge Mining is a Multidisciplinary field which is used for data mining, machine learning, natural language processing, statistics, databases, information retrieval and multimedia, etc. Web offers an unprecedented opportunity and challenges to knowledge mining. The amount of information on the Web is huge, and easily accessible. The coverage of Web information is very wide and diverse. Information/data of almost all types exist on the Web, e.g., structured tables, texts, multimedia data; etc.

A. Web Structure Mining

It is the technique to analyze and explain the links between different web pages and web sites. It mainly focuses on developing web crawlers. It works on hyperlinks and mines the topology of their arrangement.

B. Web Content Mining

It focuses on extracting knowledge from the contents or their descriptions. It involves techniques for summarizing, classification and clustering of the web contents. It can provide useful and interesting patterns about user needs and contribution behavior.

C. Web Usage Mining

It focuses on digging the usage of web contents from the logs maintained on web servers, cookies logs, application server logs etc. It works on how and when user moves from one type of content to other. Thus, it can provide association between different contents.

II. Literature Review

In this paper, it is proposed to describe a framework that aims at a solution to mine unstructured complex web data. The framework proposes to acquire the outcome of the web knowledge mining process as input, and convert these results into actionable knowledge, by enriching them with information that can be extracted from unstructured complex data format.

"A framework for mining evolving trends in Web data streams using dynamic learning and retrospective validation" It study the effect of similarity measures on the mining process and on the interpretation of the mined patterns in the harsh single pass requirement scenario. To propose a simple similarity measure that has the advantage of explicitly coupling the precision and coverage criteria to the early learning stages. It effectively handles high level stream data. But does not concentrate t noisy data. [8]

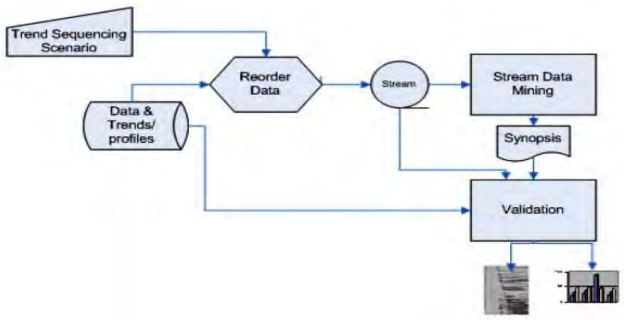

Figure 2 Web Data Stream Frameworks

"Web Service Framework Research of Data Mining in E-business" It propose a service composition framework to support a Web services-based approach for developing e-business data mining applications. It can support only e- business data. It concentrates to mine domain based data. It doesn't support complex format of the data. [9]

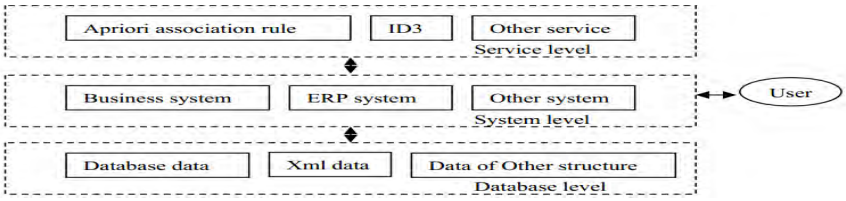

Figure 3 Web Service Framework Research of Data Mining in E-business

"Agent Based Framework for Semantic Web Content Mining " This work focuses on proving agent-based framework for mining semantic web contents employing clustering techniques. Clustering will help provide user with query relevant cluster of web contents, which will better satisfy user requirement and will provide optimal utilization of web surfing time. It can extract semantic web contents using clustering methods. It doesn't handle complex knowledge oriented data. [5]

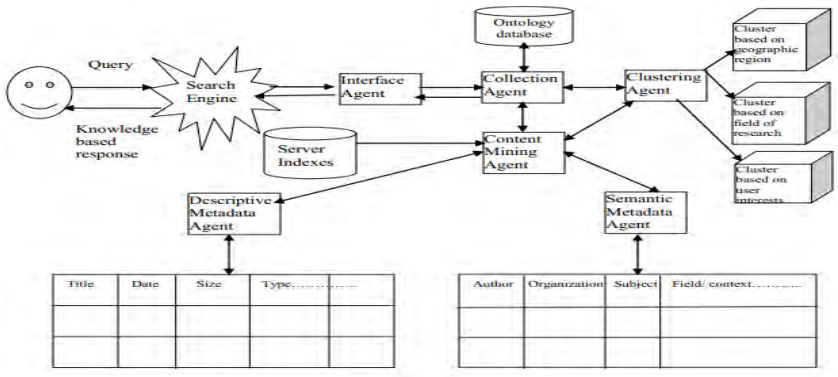

Figure 4 Semantic Web Mining

"Intelligent web traffic mining and analysis" It proposes a concurrent neuro-fuzzy model to discover and analyze useful knowledge from the available Web log data. To made use of the cluster information generated by a self organizing map for pattern analysis and a fuzzy inference system to capture the chaotic trend to provide short-term (hourly) and long-term (daily) Web traffic trend predictions. Empirical results clearly demonstrate that the proposed hybrid approach is efficient for mining and predicting Web server traffic and could be extended to other Web environments as well. This framework can mine web traffic data, but doesn't filter high dimensional volume, complex data format and unstructured data. [7]

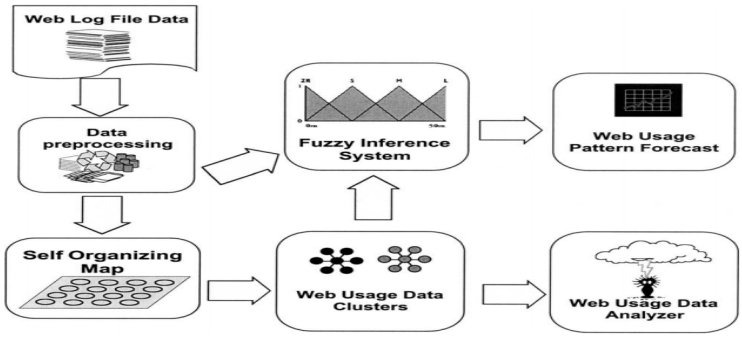

Figure 5 Web Traffic Mining Architecture

"An Evaluation of Techniques for Adaptive Search Web Mining Framework" this system deals with web personalization for individual viewers according to their taste & flavors. Provide a video on demand scheme, where viewers can choose their own video programs from the internet & customized web pages. For example for sports video – cricket, a web page can be personalized for cricket & can be faster. This framework can issue only multidimensional data. Doesn't touch complex formats. [11]

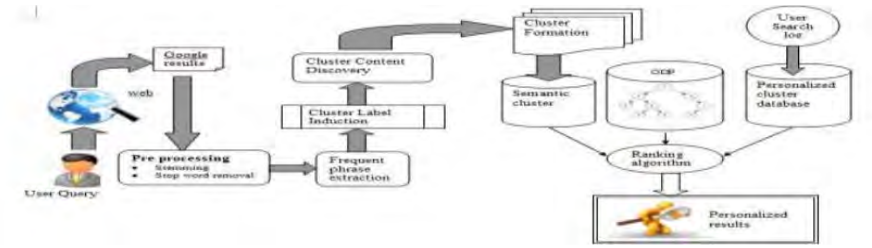

Figure 6 Multidimensional Data Framework

"A Knowledge Base for the maintenance of knowledge extracted from web data Knowledge-Based Systems" To introduce a Knowledge Base (KB), which consists of a database type repository for maintaining the patterns, and rules, as an independent program that consults the pattern repository. Using the proposed architecture, an artificial system or a human user can consult the KB in order to improve the relation between the web site and its visitors. The proposed structure was tested using data from a Chilean virtual bank, which proved the effectiveness of our approach. This framework can extract only particular domain based data. It doesn't handle complex and noisy types of patterns. [10]

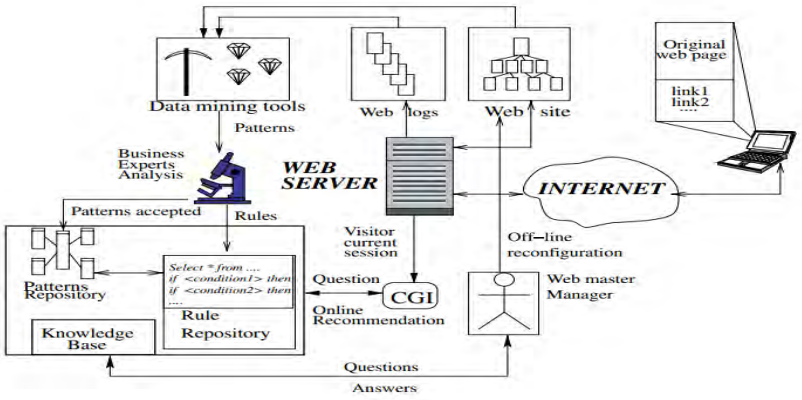

Figure 7 Web Knowledge Extraction Frame work

"Comprehensive Survey of Framework for Web Personalization using Web Mining" web mining coupled with recommendation techniques provides personalized contents at the disposal of users. It is customizing the content and structure of a web site to the specific and individual needs of each user taking advantage of the users navigational behavior. The methods employed to analyze the collected data includes Content-based filtering, Collaborative filtering, Rule-based filtering and Web Usage Mining.[12]

"A Dynamic Web Mining Framework for E-Learning Recommendations using Rough Sets and Association Rule Mining" This framework attempts to engage e-learners at an early stage of learning by providing navigation recommendations. E-learning personalization is done by mining the web usage data like recent browsing histories of learners of similar interest. [13]

"A Framework for Semantic Web Mining Model" This framework is a seven step process to find services but this model finds web services statically. It is proposed which is a semantic web service mining model that allows finding services from existing services dynamically by using OWL-S ontologies. This work detects the services based on semantic relevance and this semantic relation can be identified by the ontology analysis. [14]

"A Web Usage Mining Framework for Mining Evolving User Profiles in Dynamic Web Sites " This framework mines web usage patterns from web log files of a real web site that has all the challenging aspects of real-life web usage mining, including evolving user profiles and external data describing an ontology of the Web content.[15]

" Multithreaded Fuzzy Logic based Web Services Mining Framework" Comparative study of web services compositions with mining concepts is presented in this paper. This framework helps in finding valuable services and composing those services into composite web services. Framework is tested with different UDDI registries of large sizes and the results are compared with existing techniques.[16]

III. Proposed Framework

This section, presents a "Visual Web Knowledge Mining Architecture" for web data systems, which relies on mining and on visualization of Web Services log data captured in web environment. Web Knowledge Miner is a multi-phase architecture capable of dealing with Web services. XML based logs, and traditional Web server logs as input data. Currently, Web usage mining finds patterns in Web server logs. The logs are preprocessed to group requests from the same user into sessions during the preprocessing, irrelevant information for Web usage mining such as background images and unsuccessful requests are ignored. Databases are used instead of simple log files to store information to improve querying of massive log repositories and extract unstructured complex data.

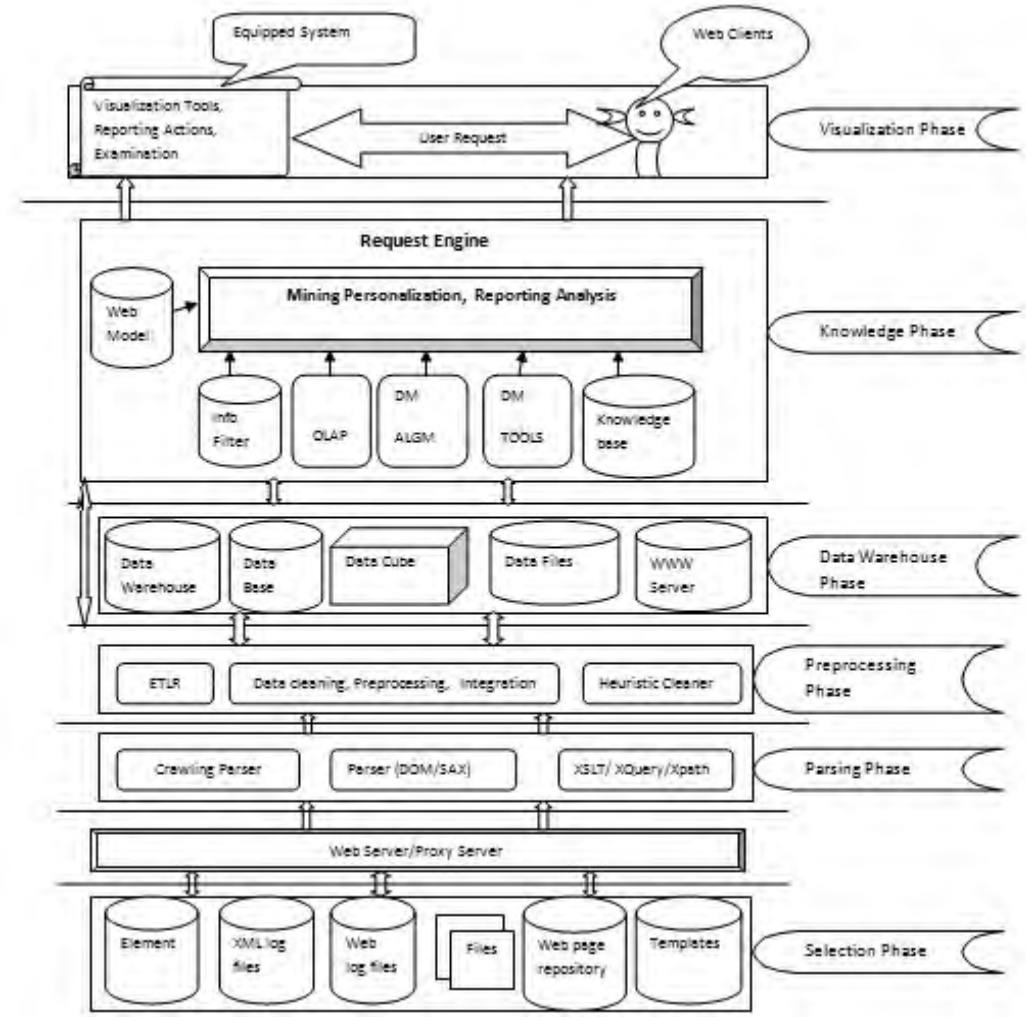

Figure 8 Web Knowledge Mining Framework

A. **Data selection Phase**

This phase proposes the foundation for a flexible data input management system as a vital part of a generic solution for quick-response decision making. It contains a collection of Web repositories such as mark up files, web log files, XML files, log files, elements; templates and Web page repositories. Data may be collected from (a) Web servers, (b) proxy servers, and (c) Web clients. Web servers collect large amounts of information in their log files. It is a data source for the framework.

B. **Parsing Phase**

The parsing phase is a set of programs and parsers are used to prepare data for further processing. This phase uses X Query, X path, DTD, XSL-fo, XSLT and XML Schemas to feed the data repository, i.e., relational or XML native database. The Web logs parser (DOM/SAX) component is used to parse and transform plain ASCII files produced by a web server to a standard database format. This component is significant to create the architecture independent from the web server supplier.

C. **Preprocessing Phase**

The preprocessing phase is used to bind the illustrations of web services and web pages to sessions and to users. For instance, extraction, cleaning, transformation and loading operations are being done. This phase is important to investigate the usage of the Web services composition used through user sessions.

D. **Data warehouse Phase**

The warehouse phase is a knowledge mining engine and is responsible for bulk loading XML data from the

database, executing SQL commands beside it and executes the mining algorithms. This phase integrates BI tools, e.g. OLAP and data mining, etc. It contains collection of website, data streams, data bases, data cubes, data files and web servers.

E. **Knowledge Phase**

The knowledge phase is a repository of input/output for complex unstructured data. It also stores pre-processed logs, sessions, and information about the web services execution. Several data mining algorithms, open source tools, OLAP, filters are used to mine the complex unstructured data.

F. **Visualization Phase**

Visualization phase should be used to present implicit and useful knowledge from request engine. Data can be viewed at typical levels such as reporting, analysis and operational system environment. This visual representation easily shows the inter relationships and dependencies between distinct components. Interactively, the model can be used to discover sensitivities and to do approximate optimization, etc.

IV. **Summary**

In this paper, a new framework is proposed based on knowledge mining techniques to improve the efficiency and effectiveness of the traditional information retrieval process. The proposed framework is designed to enhance this interaction by analyzing user access behaviors within the system. In addition to the complex content analysis (i.e., content-based filtering), information is also retrieved according to everyone's preferences and by request from other users. In contrast to the search engine, in which the users need to formulate a query composing of keywords, the request engine system automatically retrieves the unstructured and complex format information by analyzing both content and user access behavior.

V. Conclusion & Future Work

The research work existing in this paper makes several offerings to the framework of request engine systems linked research. It is proposed to introduce a new Web knowledge mining framework used to structuring a Web-page. It attempts to determine useful knowledge from derived data and high dimensional data obtained from the interactions of the users through the Web. Additionally, this XML based distributed framework demonstrates on how it handles unstructured complex data formats. Research for more discoveries will be appropriate. In near future this framework can be extended to support a specific data set, algorithm oriented implementation.

REFERENCES

[1] Subramonian, R. and Parthasarathy, S. "A Framework for Distributed Data Mining." In Proceedings of KDD 98 Workshop on Distributed Data Mining, 1998.

[2] Bhatia C.S. & Jain S., "Semantic Web Mining: Using Ontology Learning and Grammatical Rule Interface Technique". In IEEE 2011.

[3] Kosala R. & Blockeel H., "Web Mining Research: A Survey", ACM SIGKDD, Vol. 2, Issue 1,July 2000.

[4] Jicheng W., Yuan H., Gangshan W. & Fuyan Z., "Web Mining: Knowledge Discovery on the Web". In Proceedings of IEEE International Conference on System, Man and Cybernetics 1999 (IEEE SMC'99), Vol. 2 , pp. 137-141.

[5] Aarti Singh," Agent Based Framework for Semantic Web Content Mining", International Journal of Advancements in Technology, ISSN 0976-4860.

[6] Zhan L. & Zhijing L., "Web Mining based on Multi-Agents". Published in proceedings of Fifth International Conference on Computational Intelligence and Multimedia Applications (ICCIMA'03), 2003.

[7] Xiaozhe Wang, Ajith Abraham, Kate A. Smith, "Intelligent web traffic mining and analysis", Journal of Network and Computer Applications 28 (2005) 147–165.

[8] Olfa Nasraoui, Carlos Rojas, Cesar Cardona, "A framework for mining evolving trends in Web data streams using dynamic learning and retrospective validation", Computer Networks 50 (2006) 1488–1512 Science direct.

[9] HongLiu, JinHuaXu, "Web Service Framework Research of Data Mining in E-business", Advanced in Control Engineeringand Information Science, Procedia Engineering 15 (2011) 1968 19721877-7058 © 2011 .Published by Elsevier Ltd.

[10] Juan D. Vela´ squez, Vasile Palade, "A Knowledge Base for the maintenance of knowledge extracted from web data Knowledge-Based Systems" 20 (2007) 238–248.

[11] H K Sawant, Shah Ashwini V. "An Evaluation of Techniques for Adaptive Search Web Mining Framework" SSN 2249-6343 International Journal of Computer Technology and Electronics Engineering (IJCTEE) Volume 1 , Issue 2

[12] A.K.Verma , S.S.Bhatia, Vikas Verma , "Comprehensive Survey of Framework for Web Personalization using Web Mining" International Journal of Computer Applications (0975 8887) Volume 35 No.3, December 2011

[13] A.Anitha, Dr.N.Krishnan "A Dynamic Web Mining Framework for E-Learning Recommendations using Rough Sets and Association Rule Mining" International Journal of Computer Applications (09758887) Volume 12 No.11, January 2011

[14] G Ramu, Dr B Eswara Reddy, "A Framework for Semantic Web Mining Model" International Journal of Internet Computing, Volume-I, Issue-1, 2011

[15] Amsaveni.K, Vydehi.S, "A Web Usage Mining Framework for Mining Evolving User Profiles in Dynamic Web Sites" International Journal of Computer Trends and Technology- volume3Issue4- 2012

[16] Khurram Shehzad, Muhammad Younus Javed, " Multithreaded Fuzzy Logic based Web Services Mining Framework", European Journal of Scientific Research ISSN 1450-216X Vol.41 No.4 (2010), pp.632-644

[17] Cathartica,J ., Darlington, J, "An Architecture for Distributed Enterprise Data Mining," 7th Intl. Conf. on High-Performance Computing and Networking, 1999.

[18] CHEN Yu-ru,HUNG Ming-chuan Don-Iin YANG. "Us|ng data mining to construct an intelligent web search system" [J].International Journal of Computer Processing of Oriental Languages,2003,16(2)

[19] C. Haruechaiyasak, M.-L. Shyu, and S.-C. Chen, "A Web-page recommended system via a data mining framework and the semantic web concept," accepted for publication, International Journal of Computer Applications in Technology, Special Issue on Data Mining Applications, 2004.
[20] Chen ting, Niu xiao, Yang Weiping, "The Application of Web Data Mining Technique in Competitive Intelligence System of Enterprise based on XML" Third International Symposium on Intelligent Information Technology Application, 2009, pp. 396 – 399.
[21] M. Spiliopoulou, "The laborious way from data mining to Web log mining" International Journal of Computer Systems Science and Engineering, Vol. 14, No. 2, 113-125, 1999.
[22] R. Cooley, B. Mobasher, and J. Srivastava, "Data Preparation for Mining World Wide Web Browsing Patterns", Journal of Knowledge and Information Systems, 1(1), 1999.
[23] Ingram, A. (1999). "Using web server logs in evaluating instructional web sites". Journal of Educational Technology Systems, 28(2), 137–157.
[24] G. Piatetsky-Shapiro, C.J. Matheus, "Knowledge discovery workbench for exploring business databases", International Ž. Journal of Intelligent Systems 7 1992 675–686.
[25] [25]Vuda Sreenivasa Rao, "Multi Agent-Based Distributed Data Mining": An Over View, International Journal of Reviews in Computing, pp 83 – 92, ISSN: 2076-3328, E-ISSN: 2076-3336.
[26] D.J.H and, H.Mannila, and P.Smyth. "Principles of Data Mining".MIT Press, 2000.f Computer Science, 2000.
[27] F. Masseglia, P. Poncelet, and R. Cicchetti, "An Efficient Algorithm for Web Usage Mining", Networking and Information Systems Journal (NIS), vol.2, no. 5-6, pp. 571603, 1999.
[28] K.P. Joshi, A. Joshi and Y. Yesha, "On using a warehouse to analyze web logs, Distributed and Parallel Databases", 13 (2), pp. 161180, 2003.
[29] 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. 1, no. 2, pp. 12-23,2000.
[30] F.M. Facca and P.L. Lanzi, "Mining interesting knowledge from weblogs: a survey, Data & Knowledge Engineering", Volume 53, Issue 3, pp.225-241, 2005.
[31] Sarwar, B., Karypis, G., Konstan, J.A., & Reidl, J. "Item-based Collaborative Filtering Recommendation Algorithms". Proceedings of the Tenth International --Conference on World Wide Web, pp. 285 -295, 2001.
[32] Cheng Z., Yong F.Y.S. : "The Implementation of the Web Mining based on XML" technology, International Conference on Computational Intelligence and Security,2009 Page(s):84-87