ISSN: 2277-9655
Scientific Journal Impact Factor: 3.449
(ISRA), Impact Factor: 2.114

IJESRT

INTERNATIONAL JOURNAL OF ENGINEERING SCIENCES & RESEARCH TECHNOLOGY

DETAILED STUDY OF WEB MINING APPROACHES-A SURVEY

Khushbu Patel*, Anurag Punde, Kavita Namdev, Rudra Gupta, Mohit Vyas

ABSTRACT

This paper is a work on survey on the existing techniques of web mining and the issues which are related to it. The WWW i.e. World Wide Web acts as an interactive and trendy way to transfer information. The enormous and diverse information is available on the web, although the end users cannot make use of the information very effectively and easily. Data mining concentrates on non trivial extraction of implicit previously unknown and potential useful information from the very large quantity of data. Web mining is one of the applications of data mining which has become an important area of research due to vast amount of World Wide Web services in recent years. The aim of this paper is to provide the past and current techniques in Web Mining. This paper also reports the summary of various techniques of web mining approached from the following angles like Feature Extraction, Transformation and Representation and Data Mining Techniques in various application domains. In this paper, we will discuss the research work done by different users depicting the pros and cons are discussed. We will also discuss the overview of growth in research of web mining and some important research issues related to it.

KEYWORDS: Video mining, Audio mining, Text mining and Image mining, Association rule mining, Data pre-processing.

INTRODUCTION

Web is a collection of billion of documents. The use of World Wide Web is very massive, diverse, flexible, and vibrant. The World Wide Web continues to grow both in the huge volume of traffic and the size and complexity of Web sites. It is difficult to identify the relevant information present in the web. The most of the contents present on the web are unstructured in nature, but extremely tiny work deals with unstructured and mixed or heterogeneous information on the Web. The promising field of web mining aspires at finding and extracting relevant information that is hidden in Web-related data, in particular in text documents published on the Web. Data Mining involves the concept of extraction meaningful and valuable information from large volume of data. Web mining is an important area in data mining where we extract the interesting patterns from the contents. We will generally handle 3 kinds of information in web site namely 1. Content 2. Structure 3. Log data. Based on these kinds of information the Web Mining consists of 3 processes namely Web Content Mining, Web structure Mining and Web Usage Mining [8] as shown in figure 1. In this figure we have shown web mining problems and their different approaches. We basically use web structure mining mainly with the structure of the web sites [5]. Web Usage mining usage characteristics of the users of Web applications. It is in a semi-structured format so that it needs lots of pre-processing and parsing before the actual extraction of the required information. In this paper, we have given the survey of web mining techniques. Data mining process have several stages namely [9] Domain Understanding, Data selection, Data pre-processing and cleaning, Pattern discovery, Interpretation and Reporting. We provide the web mining techniques survey as shown in the figure 1.

ISSN: 2277-9655
Scientific Journal Impact Factor: 3.449
(ISRA), Impact Factor: 2.114

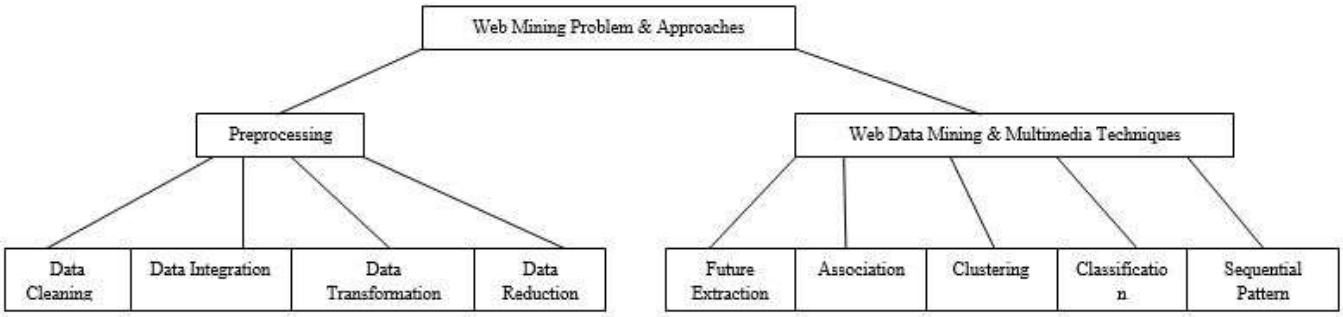

Figure 1: Web Mining techniques

The projected plan achieves the following goals:

- Discussion of existing techniques for web mining in text and multimedia.
- Identifying the issues during mining of data from Feature Extraction, Transformation and Representation and Data Mining Techniques in various application domains.
- Issues related to data pre-processing, pattern discovery, web usage mining, multimedia mining

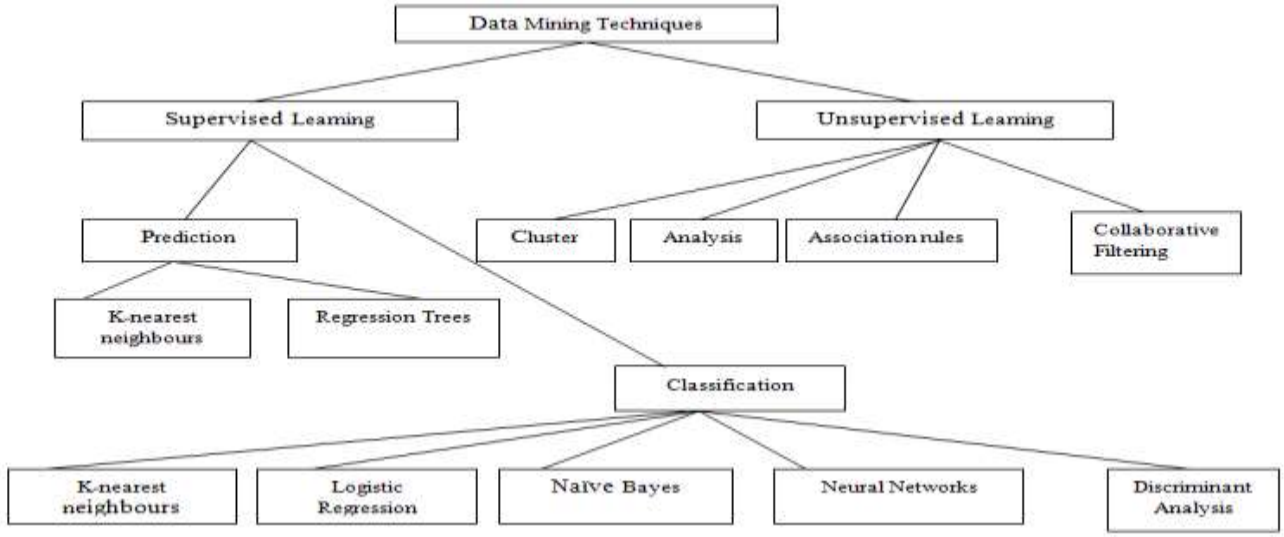

Figure 2: Data Mining Techniques outline

WEB MINING PROBLEMS AND APPROACHES

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.

Data Pre-processing Techniques

Web log pre- processing is the first step that is important to improve the efficiency and quality of the web data because almost 70% of the time is taken in pre-processing and these pre-processed data are given as an input to the next stages pattern discovery and pattern analysis. There are many techniques available for pre-processing since a long time. Web log file plays a significant role in pre-processing as the

ISSN: 2277-9655
Scientific Journal Impact Factor: 3.449
(ISRA), Impact Factor: 2.114

contents the user browse are recorded in these log files [1]. The data can be stored either at sever side, client side, on proxy servers and on operational database. Web Server Logs maintains a history of page requests. Information about the request, client IP address, request date/time, page requested, HTTP code, bytes served, user agent, are stored. Proxy Server Logs a caching mechanism which lies between client browsers and Web servers. It helps to reduce the load time of Web pages as well as the network traffic load at the server and client side and Browser Logs that can be modified or various JavaScript and Java applets can be used to collect client side data. Client-side collection scores over server-side collection because it reduces both the user and session identification problems. The advantages and disadvantages of log files and their behavior are shown in the table given below. To improve efficiency and quality of patterns mined and to avoid these noisy and dirty data various pre-processing techniques are available like Data cleaning, Data integration, Data transformations, and data reduction [7].

TABLE-I
Log File Pros & Cons

Log File	Advantages	Disadvantages	Behavior Relation
Client Log File	Authentic & Accurate	Modification & Collaboration	One to many
Server Log File	Reliable & Accurate	Incomplete	Many to one
Proxy Log File	Control Efficiency of corporate access to the internet, log traffic.	Complex, unsecure	Many to Many
Operational Database	Simplicity, efficiency & Accurate	Vulnerability, Maintenance	--------

Data cleaning- It is needed to remove noise and correct inconsistencies in the data. The mains problems of data cleaning are missing values, noise, inconsistencies and duplicate elimination[3]. The techniques used in missing values are classification, regression, interference based tools using Bayesian formulation, Decision Tree Induction. Binning, Smoothing, Regression, Clustering is useful to remove the noisy data from the database. The Duplicate elimination uses sorted neighbourhood method developed to reduce the number of required comparisons. A number of commercial tools, e.g., IDCENTRIC (First Logic), PUREINTEGRATE (Oracle), QUICKADDRESS (QASSystems), REUNION (Pitney Bowes), and TRILLIUM (Trillium Software), focus on cleaning this kind of data. Duplicate elimination- Sample tools for duplicate identification and elimination include DATACLEANSER (EDD), MERGE/PURGELIBRARY (Sagent/QMSoftware), MATCHIT (HelpITSystems), and MASTERMERGE (Pitney Bowes).

Data integration - To merge data from multiple sources into a coherent data store, such as a data warehouse or a data cube we use this technique. There are a number of issues to consider during data integration. Schema integration can be tricky. This is referred to as the entity identification problem. Redundancy is another important issue. A third important issue in data integration is the detection and resolution of data value conflicts. Data transformation- Data transformation involves the techniques like Smoothing, Aggregation and Normalization.

Data Pre-processing Challenges

- Data cleaning seems to be difficult for semi structured data and unstructured data but most of the data seems to be structured. So more work has to be done in cleaning semi-structured data.
- Data transformation is an important phase that is done in pre-processing of data. But no exact tools are available.
- Research work should be done on implementing the best tool for data transformation[1].
- Limited interoperability.
- Though Duplicate elimination uses many methodologies and tool it still remains a tedious task to be performed.
- Query processing is difficult on heterogeneous data.

Survey on Pattern Extraction Techniques

Data mining techniques has two approaches that include descriptive mining and predictive mining. Descriptive mining concentrates on the general properties of data in the database and predictive mining concentrates on data to make predictions[5]. The data mining Techniques are illustrated in Fig 2. Tasks for performing preprocessing of Web Usage Mining involve data cleaning, user identification, session identification, path completion, session

ISSN: 2277-9655
Scientific Journal Impact Factor: 3.449
(ISRA), Impact Factor: 2.114

reconstruction, transaction identification and formatting [1]. However, in general, these tools provide little or no analysis of data relationships among the accessed files and directories within the Web space. Now more sophisticated techniques for discovery and analysis of patterns are emerging. These tools fall into two main categories: Pattern Discovery Tools and Pattern Analysis Tools. Pattern discovery draws upon methods and algorithms developed from several fields such as statistics, data mining, machine learning and pattern recognition. They are statistical analysis, association, rule mining [13], clustering, classification and sequential pattern mining. The works done by different author are categorized into Association rule mining Clustering, Classification and Sequential pattern mining [14].

Association rule mining: Association Rules find all sets of items that have support greater than the minimum support and then using the large item sets to generate the desired rules that have confidence greater than the minimum confidence. An algorithm for finding ass rule named as AIS was proposed by R.S.Agarwal et al. in 1993. The disadvantage of the AIS algorithm is that it results in unnecessary generation and many candidate item sets . The Apriori algorithm takes advantage of the fact that any subset of a frequent item set is also a frequent item set[5]. The disadvantages are that multiple scans have to be done on the database and it has complex time and memory consuming. The advantage of AprioriTid algorithm is that the number of entries may be smaller than the number of transactions in the database, especially in the later passes but the cost of switching should be taken into account. AprioriHybrid Algorithm Apriori does better than AprioriTid and AprioriTid does better than Apriori in the later passes. FP –Tree algorithm scans the database only twice but it seems to be difficult in incremental and interactive rule mining [13]. Custom built Apriori algorithm that is efficient and does effective pattern analysis. Another algo Bin Li Wang et al., 2010, proposed a new method to Improvement of Apriori Algorithm Based on Boolean Matrix. It scans transaction database only one time, thus reduces the system cost and increases efficiency of data mining[9].

TABLE II
ASSOCIATION RULE MINING LITERATURE SURVEY

Algorithms Used	Author	Advantages	Disadvantages	Year
AIS	R.S.Agarwat.et.al	Efficient.	Items below min support are eliminated. It Generates rules with single item set. Many candidates and low support value.	1993
Apriori	Q.Zhao.et.al	Reduce search space, computation, I/O and memory costs.	Multiple scans on database, complex, time and memory consuming.	1994
Apriori-TID	A.Ceglaret et. al.	Reduce multiple scans.	Cost of switching.	1995
FP-tree	Han&Pei	Scans are limited only twice. No candidate generation.	Difficult in incremental rule mining and iterative mining process.	2000
RARM	DAS,Ng& Woon	Fast, Efficient and scalable.	Difficult in Incremental rule mining and Iterative mining process.	2001
Improved Apriori	WANG Tong, et,al	Less complexity, time .	Memory space should be considered .	2005
Custom built Apriori	Sandeep sing et.al.	Effective pattern analysis.	Real world entry may be proposed work.	2010
Association rule mining from data with missing values	K.Rameshkumar	Outperforms when the ratio of missing values is low and high, and also when support is minimum and maximum level, and when representativity threshold level is low and high.	This work is implemented with real time domain like web and medical datasets.	2011
Association rule hiding algorithm	R.Natarajan, Dr.R.Sugumar, M.Mahendran, K.Anbazhagan.	Hide certain crucial information so they cannot be discovered through association rule.	Not specified	2012
Multiobjective Association Rule Mining with Evolutionary Algorithm	Jie Zhang, Yuping Wang, and Junhong Feng.	Reduces the number of comparisons and time consumption, and improves the performance .	To extend it to immediately use the categorical or numeric dataset rather than converting them into Boolean dataset.	2013

ISSN: 2277-9655
Scientific Journal Impact Factor: 3.449
(ISRA), Impact Factor: 2.114

Jain & dubes et.al in 1998 and Kaufmann et al. 1990 proposed the Agglomerative and Divisive algorithm to perform hierarchical clustering. It was flexible and easy to handle but vague and did not visit intermediate clusters. Partition Relocating Clustering and Density Based Clustering have an advantage of Interoperability and can be modified easier. The discovery of user navigation patterns using SOM is proposed by Etminai et al[5]. SOM is used to pre-process the web logs for extracting common patterns. Fuzzy clustering tech can be used to discover groups that share similar interest by examining data gathered in web servers. Mehrdad et al. gave an approach based on graph portioning for mining navigation patterns. Kobra et al. used Ant Based Clustering algorithm to extract frequent patterns for pattern discovery and the result was displayed in an interpretable format. N.Sujata has proposed a new framework to improve web session cluster quality from k-means with genetic programming. The k means was used for clustering and GA to improve the cluster quality.

TABLE III
CLUSTERING LITERATURE SURVEY

Algorithms Used	Author	Advantages	Disadvantages	Year
Algomerative Divisive	Jam&dubes 1998 Kaufman 1990	Flexibility, ease of handling.	Vagueness, do not visit intermediate cluster.	1998
Relocating Probalistic k-means, k-mediods		Interpretable, modifiable and Easier.	Lack of scalability.	1996
SOM	Pola britos et.al, Teuvo Kohonen	Simple, Effective.	Missing data, computationally expensive, time consuming.	2007
Ant based	Kobra etminami.et.al	Flexible, Robust.	Performance is high.	2009
K means with genetic algorithm	N.sujata.et.al	Minimise objective function.	Not the fastest algorithm but performance is comparable.	2010
EB-DBSCAN (Entropy-Based DBSCAN)	Quan Qian, Tianhong Wang, and Rui Zhan	1.EB-DBSCAN algorithm has a great prospect in clustering of high-speed and massive data stream of arbitrary shape. 2. EB - DBSCAN algorithm uses batch data Processing so size of the data processing is effectively reduced and it can greatly reduce the time complexity. 3. quick-building detecting models for high-speed, huge amount of stream data.	1. The size of window is a direct factor affecting the average clustering precision. 2. EB-DBSCAN algorithm has a bit lower average purity than DBSCAN algorithm, but almost Equivalent. 3. Parameter selection is very critical for the EB-DBSCAN algorithm.	2013
Hierarchical Agglomerative Clustering Algorithm	Kriti Srivastava, R. Shah, D. Valia, and H. Swaminarayan	Can handle large dataset, increase efficiency of algorithm and parallelism have reduced time required for execution.	There is an increase only in linearly grow of execution time.	2013

Chen et al. [1] introduced the concept of "maximal forward references" which can be described as the sequences of user's request documents up to the last one before backtracking. Pie et al. [2] introduced a WAP-mine algorithm. They also proposed a WAP-tree. This method is faster than conventional methods. WAP-mine is efficient than GSP-based solution in a wide margin. Mortazaviasl et al. [3] introduced a novel projection based algorithm Prefix Span, which support sequential patterns mining. They basically worked on partitioning the database of user sequences into smaller databases for mining sequential patterns. Bestavros et al. [13] presented a Markov modelling application for web data. To predict the subsequent link within a certain period of time that a user might follow, the first-order Markov model.

ISSN: 2277-9655
Scientific Journal Impact Factor: 3.449
(ISRA), Impact Factor: 2.114

TABLE IV
SEQUENTIAL PATTERN MINING LITERATURE SURVEY

Technique	Algorithm	Advantage	Function
Association rule, Hasing, Pruning	Hasing and pruning based algorithm for mining association rules.	Scalability	Traversal in distributed info providing environment.
WAP tree association rule algorithm	Conditional search strategy.	Scalability	Use to access patterns from web log.
Sequential patterns	General Algorithm.	-do-	Prediction of web logs for improving hypertext structure.
Markov chains clustering	Transaction matrix comparison algorithm.	Scalability	Clusters web pages with similar transaction behavior and probability.
Classification	Prefix span.	Scalability	Traversed pattern in OLAP.
Sequential pattern mining	High utility sequential Patterns.	Scalability	USpan can efficiently identify high utility sequences in large-scale data with low minimum utility.
Pattern Growth based	C-PrefixSpan (constraint PrefixSpan).	Efficient and scalable.	The C-Prefixspan approach produces sequential patterns which satisfies Length, Aggregate, Frequency and Gap constraints also minimum confidence.
Pattern-growth methodology	Prefix-Suffix-Binding Span	Efficient and effective	Generating the complete set of 2-piece correlated patterns k-Piece Correlated Pattern and k-Piece Maximum Cutting Probability.
CSW-BV (Customer Sliding Window with Bit-Vectors), lexicographic tree-based data structure, called LexSeq-Tree (Lexicographic Sequence Tree	IncSpam (Incremental Sequential pattern mining.	Handles memory requirement efficiently and Scalable.	One-pass approach for mining sequential patterns from streaming Itemset-sequences.

Classification is a data mining technique used to predict group membership for data instances. In this paper, we present the basic classification techniques [5]. Several major kinds of classification method including decision tree induction, Bayesian networks, k-nearest neighbour classifier, case-based reasoning, genetic algorithm and fuzzy logic techniques.

ISSN: 2277-9655
Scientific Journal Impact Factor: 3.449
(ISRA), Impact Factor: 2.114

TABLE V
CLASSIFICATION LITERATURE SURVEY

Tech	Advantages	Disadvantages	Algorithms used	Year
Classification	Requires a small amount of training data to estimate the parameters, simple and efficient.	Not capable of solving more complex problems.	Naive Bayesian Algorithm.	2010

Pattern Analysis [1]: This is the final step in the Web Usage Mining process. After the pre-processing and pattern discovery, the obtained usage patterns are analyzed to filter uninteresting information and extract the useful information. The methods like SQL processing and OLAP[11] can be used.

Pattern Discovery Challenges

- Classifications of the documents are done using many techniques. But maintaining Accuracy in document classification is not to the expectation.
- Research work has to be done to Improve cluster quality
- There is no efficient algorithm for pattern extraction.
- Personalization
- Identification of exact user is not possible for mining purpose
- The exact sequence of pages user visit is difficult to uncover from server site.
- Security , privacy issues

Multimedia Data Mining

The data present in the web contains several types of data, such as text, image, audio, video, metadata and hyperlinks. Some of them are semi-structured or more structured data but most of the data are unstructured [5] Multimedia data mining is part of the content mining, which is engaged to mine the high-level information and knowledge from large online multimedia sources. All together the web content mining consists of mining of text, multimedia and the semantic web.

Image Data Mining

The classification module present in Multimedia Miner classifies multimedia data including image based on some class labels. Vailaya et al. uses binary classifier to perform hierarchical classification of images in indoor and outdoor categories[5]. The SVM based approaches uses SVMs to maximize the margins between positive and negative images. The A-priori based growth association is used to find the relation between the structures and functions of the human brain. The main issue is scalability in terms of candidate generation [13].

FUTURE DIRECTION

The web usage mining algorithms are more efficient and accurate. But there is a challenge that has to be taken into consideration. Web cleaning is the most important process as researchers say 70% of the time is spent on data pre-processing. But data cleaning becomes difficult when it comes to heterogeneous data. Maintaining accuracy in classifying the data needs to be concentrated. Though many classification techniques exist the quality of clustering is still a question to be answered. The database is huge and it contains large dataset so mining interesting rules adds on to uninterested rules that are huge. These are due to large item set which naturally decrease the efficiency of the mining methodologies. Moreover mining rules from semi structure and unstructured as in the semantic web becomes a great challenge. This leads to time and memory consumption[5][11][15]. Research work has to be concentrated on these issues as web data rule the Web. Maintain privacy of the user also peeps in as it is misused in data pre-processing.

CONCLUSION

In this paper we have discussed about the research issues and the drawbacks of the existing techniques. More research work need to be done on the web mining domain as it will rule the web in the near future. Web mining along with semantic web known as semantic web mining is to be concentrated that is evolving which helps us to overcome the cons of web mining. Though various algorithms and techniques have been proposed still work has to be done in discovering new tools to mine the web.

REFERENCES

1. Ms. Dipa Dixit, Fr.CRIT, Vashi, M Kiruthika," PREPROCESSING OF WEB LOGS", (IJCSE) International Journal on Computer Science And Engineering, Vol. 02, No. 07, 2010, 2447-2452.
2. Dr. Sohail Asghar, Dr. Nayyer Masood," Web Usage Mining: A Survey On

ISSN: 2277-9655
Scientific Journal Impact Factor: 3.449
(ISRA), Impact Factor: 2.114

Preprocessing Of Web Log File Tasawar Hussain", 978-1-4244-8003-6/10@2010.
3. Theint Theint Aye "Web Log Cleaning For Mining Of Web Usage Patterns".
4. S. K. Pani, et.al L "Web Usage Mining: A Survey On Pattern Extraction From Web Logs", International Journal Of Instrumentation, Control & Automation (IJICA), Volume 1, Issue 1, 2011.
5. Chidansh Amitkumar Bhatt • Mohan S. Kankanhalli, "Multimedia Data Mining: State Of The Art And Challenges" Published Online: 16 November 2010© Springer Science+Business Media, LLC 2010.
6. Margaret H. Dunham, Yongqiao Xiao Le Gruenwald, Zahid Hossain," A SURVEY OF ASSOCIATION RULES Web Usage Mining".
7. Brijendra Singh1, Hemant Kumar Singh2,"WEB DATA MINING RESEARCH: A SURVEY", 978-1-4244-5967-4/10/$26.00 ©2010 IEEE.
8. Rajni Pamnani, Pramila Chawan 1 Qingtian Han, Xiaoyan Gao, "Web Usage Mining: A Research Area In Web Mining".
9. Wenguo Wu, "Study On Web Mining Algorithm Based On Usage Mining", Computer- Aided Industrial Design And Conceptual Design, 2008. CAID/CD 2008. 9th International Conference On 22-25 Nov.2008.
10. R. Kosala, H. Blockeel. "Web Mining Research: A Survey," In SIGKDD Explorations, ACM Press, 2(1): 2000, Pp.1-15.
11. http://www.kdnuggets.com
12. http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.91.1602
13. J Vellingiri, S.Chenthur Pandian, "A Survey on Web Usage Mining", Global Journal of Computer Science and Technology .Volume 11 Issue 4 Version 1.0 March 2011.
14. Chen L, Mao X,Wei P, Xue Y, Ishizuka M (2012) Mandarin emotion recognition combining acoustic and emotional point information. Appl Intell 37(4):602–612.
15. Shang F, Jiao LC, Shi J, Wang F, Gong M (2012) Fast affinity propagation clustering: a multilevel approach. Pattern Recognition 45(1):474–486.
16. J. Shao, X. He, C. Bohm, Q. Yang, C. Plant, "Synchronization-Inspired Partitioning and Hierarchical Clustering," IEEE Transactions on Knowledge and Data Engineering, 2012.
17. Ta¸sdemir K (2012) Vector quantization based approximate spectral clustering of large datasets. Pattern Recognition 45(8):3034–3044.
18. Abdelhakim Herrouz, Chabane Khentout, Mahieddine Djoudi,''Overview of Web Content Mining Tools", The International Journal of Engineering And Science (IJES), Volume 2, Issue 6, 2013.