Use of Ant Colony Optimization for High Performance Web Usage Mining

[1]Rakhi Chourasia, [2]Brajesh Patel
[1,2]Dept. of Computer Science Engineering, SRIT Jabalpur, MP, India

Abstract
Web mining is important in current era for the ease of users of the web to achieve fast and accurate searches. Since web data is huge and requires online presence therefore it is difficult for the developers to process web data directly. Web Log mining is therefore applied to work offline to process web data. The process of web log mining is also involved as lot of information is collected in web logs and therefore mining and clustering techniques are applied for the same. This work proposes to apply Ant Clustering mechanism to extract relevant information from the web log dataset available for researchers over the Internet. The focus of this work is achieving highly efficient web log mining for usage of the web. For efficient web mining this work uses data cleaning and analyzing steps with the application of the Ant Clustering mechanism which applies parallel processing of the different characteristics of the web logs for mining of information as per the usage by the users. For usage mining, characteristics such as unique IP Addresses, URL, Time spent, Http Method used, status message generated and sent etc. have been used.

Keywords
Web Data Mining, ANT Clustering, Data Pre Processing, Data Post Processing, Characteristics of web usage, Web Usage Mining.

I. Introduction

Web Usage Mining (WUM) is all about identifying user browsing patterns over WWW, with the aid of knowledge acquired from web logs. The outcomes of the WUM can be used in web personalization, improving the performance of the system, modification of the site, business intelligence, usage characterization etc. The working of WUM has three steps –preprocessing of the data, pattern discovery and analysis of the patterns. Results of the pattern discovery directly influenced the quality of the data processing. Good data sources not only discover quality patterns but also improve the WUM algorithm. Hence, data preprocessing is an important activity for the complete web usage mining processes and vital in deciding the quality of patterns. In data preprocessing, the collection of various types of data differs not only on type of data available but also the data source site, the data source size and the way it is being implemented. The data preprocessing of WUM is focused research field nowadays. This research paper studies the preprocessing of data in Web usage mining.

Clustering techniques have been widely applied in the information retrieval (IR) and information filtering (IF) context. In the IR and IF contexts it is desirable that a document is assigned to more than one cluster to a distinct degree. Soft clustering techniques generate overlapping clusters.

Almost all clustering methods assume that each item must be assigned to exactly one cluster and are hence partitioned. However, in a variety of important applications, overlapping clustering is a technique where items are allowed to be members of two or more discovered clusters. The characteristic of soft clustering approaches with respect to the property of robustness i.e. the performance of a system should not be affected drastically due to outliers (bad observations).

Another issue when clustering textual documents is need to generate a hierarchy, i.e. a taxonomy-like structure of clusters, so as to provide a classification of documents organized into topics of distinct granularity. This allows analyzing the contents of a collection at distinct levels of specificity. Actual hierarchical clustering techniques are inadequate since they generate a dendrogram graph, in which at each node a cluster (or two clusters) is (are) split (merged) into two child cluster (one parent cluster). From the dendrograms several flat partitions can be obtained by specifying distinct thresholds on the minimum intra-cluster similarity. Therefore, from dendrograms one can derive cluster representing topics of distinct granularity, but at the cost of further processing.

A. Web Data Mining

Web data mining can be broadly defined as the discovery and analysis of useful information from the Internet. There are vast amounts of data information on the Web. It has become the research focus of the advanced database technology, the Internet and information retrieval field how to do complex applications of these data. Data mining is finding the implicit regularity information from large amounts of data to resolve the application of data quality problems. Taking full advantage of useful data and wasting useless data is the most important applications of data mining technology. Unlike a fully structured data in traditional databases, the top characteristic of data on the Web is semi-structured, which makes Web-oriented data mining to be more complex than a single data warehouse mining.

The data on the Web without a specific model description, the data of each site are independently designed, and the data itself has a readme and dynamic variability. Thus, the data on the Web has a certain structural levels of existence, but the readme, which is not fully structured data, which is also known as semi-structured data.

B. The Types of Web Data Mining

The template is used to format your paper and style the text. All margins, column widths, line spaces, and text fonts are prescribed; please do not alter them. You may note peculiarities. For example, the head margin in this template measures proportionately more than is customary. This measurement and others are deliberate, using specifications that anticipate your paper as one part of the entire proceedings, and not as an independent document. Please do not revise any of the current designations.

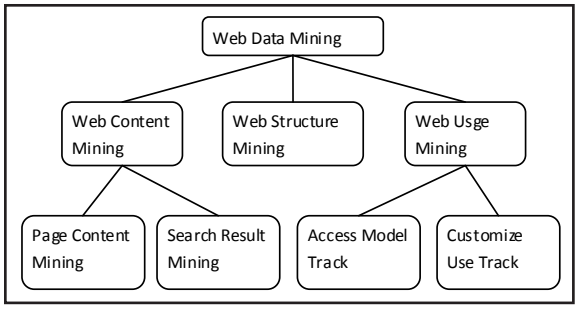

Fig. 1: Types of Web Data Mining

C. Ant Clustering

Ant-based techniques, in the computer sciences, are designed for those who take biological inspirations on the behavior of these social insects. Data-clustering techniques are classification algorithms that have a wide range of applications, from Biology to Image processing and Data presentation. Since real life ants do perform clustering and sorting of objects among their many activities, we expect that a study of ant colonies can provide new insights for clustering techniques [26].

Clustering is the classification of similar objects into different groups, or more precisely, the partitioning of a data set into subsets (clusters), so that the data in each subset share some common trait. Clustering is used as a data processing technique in many different areas of application, such as bioinformatics, data mining, image analysis, etc.

D. Overview of Web Log

Web log records all the information of user's activities from sending request to the server. The original data source is from the network access log. Originally, log files are produced for debugging, log files can be found from three different places: 1. network server, 2. network proxy server, 3. client browser.

Web log file is server recording information of user's requests for resources to a specific site each time. Most logs use common log format. The following is a log fragment taken from a web server:

65.52.109.26--
[18/Feb/2012:23:59:41+0800]"GET/index.php?do=rck&indus_id=6&slt_page_size=10&join=2&price=&style=&slt_order=2&p=Liaoning HTTP/1.1" 200 12311"-" "Mozilla/5.0 (compatible; bingbot/2.0; +http://www.bing.com/bingbot.htm)"

It shows the information as shown below:

Remote IP address or domain name: IP address is a 32-bit host address defined by the Internet Protocol; domain name is used to determine unique network address for any host online, an IP usually corresponds to a domain name.

Authorized user: username and password used when server requires user to be verified,

Date and time of login and logout

Mode requested: GET, POST, or HEAD method of CGI (Common Gateway Interface).

Status: HTTP status code returned to the user, such as 200 is "OK", 400 "not found".

Byte: length of text content transferred.

Remote log and agency-side log Remote URL (Uniform / Universal Resource Locator) "request": completely from request line of client URL requested

E. HTTP Overview

Hypertext transfer protocol is used on the World Wide Web in 1990, is an application layer protocol. The first version of HTTP is HTTP/0.9, is a simple protocol used in the whole Internet, for raw data transfer. HTTP/1.0 is defined by RFC 1945. Table 1 is frequently-used status code of data transmission error and success by Hypertext Transfer Protocol.

Table 1: State of Hypertext Transfer Protocol

Status code	Significance
101	Switching Protocols
200	OK
202	Accepted
305	Use Proxy
400	Bad Request
401	Unauthorized
403	Forbidden
404	Not Found
408	Request Time-Out
500	Server Error
502	Bad Gateway

II. Existing System

This paper first introduces the theory and knowledge related to Web logs, and then introduces a Web log mining process. As to this article is based on Web log processing, we draw valid user data according to a specific website users' access data through the preprocessing, and then research and analyze the users behaviors. The work can provide a theoretical basis for the management and optimization of the site for site managers. Experiments have showed that the work is effective [1].

Web engineering practices are used to build web based software developments. To successfully develop web based application developers need a systematic model. To manage the changes in requirements is a challenging task for web application development. We present a model for improved version of web engineering practice using Data Mining technique along with the effective usage of requirement change management. Structured and semi structured research methods are used to interview as reputed organization experts. Control experiment is also used for evaluating the Data Mining technique with requirement change management. This model improves the existing web development practices using AI technique. Expert's reviews are used to evaluate this study. The model provides a guideline to the team of developers for developing any web based application [2].

Our objective is to develop a data preprocessing method applied to Moodle logs based on SCORM content structure. In earlier works [1-2], we proposed a preprocessing tool to implement these new methods and present the first discovered knowledge.

In this research, we define new static variables according to the SCORM content tree and we apply more statistics and visualization techniques. In addition, we present multidimensional graphics in order to understand users' accesses. These aggregated variables provide teachers and tutors with interesting knowledge about students' learning process according to different levels of content accessed. [3]

Web data mining is a key tool for e-commerce in such an age of Internet. Due to previous studies, there is no such a mining technology superior to others. Therefore, this paper can give a simplified comprehension of web data mining and indicate the improvement direction of each kind of mining algorithm based on their existing defects. This paper first introduces the main process of data mining, including collect data, preprocess data, store the data, apply data mining technologies and evaluate the result. Then this paper tries to analyze different kinds of data mining technologies and how they are used in e-commerce. Finally, we predict the development trend of web data mining in ecommerce domain [4].

This paper presents an efficient approach for Document Classification based on FDCKE. The paper introduces a new Framework for Document Classification and Knowledge Extraction (FDCKE).The FDCKE approach is an integration of document classification phases like document collection from

heterogeneous sources, Text Pre-processing of the documents, Feature Selection, Indexing, Classification Process, Results Analysis and Performance Measures. The proposed FDCKE is a unified interface that can be used for Classification, Association and Clustering. Twenty News group data sets [23] are used in the Experiments. The performance evaluation of Experimental Results has been done by SAS Software. The Experimental Results show that the proposed approach out performs [5].

Web Usage Mining (WUM) is one of the categories of data mining technique that identifies usage patterns of the web data, so as to perceive and better serve the requirements of the web applications. The working of WUM involves three steps - preprocessing, pattern discovery and analysis. The first step in WUM - Preprocessing of data is an essential activity which will help to improve the quality of the data and successively the mining results. This research paper studies and presents several data preparation techniques of access stream even before the mining process can be started and these are used to improve the performance of the data preprocessing to identify the unique sessions and unique users. The methods proposed will help to discover meaningful pattern and relationships from the access stream of the user and these are proved to be valid and useful by various research tests. The paper is concluded by proposing the future research directions in this space [6].

Recommender system based on web data mining is widely used in e-commerce for it generates more accurate and objective recommendation results and provides personalized service for web users. This paper makes analysis on some major recommendation methods based on web data mining such as Collaborative Filtering and Association Rules mining, and discusses the practical application of these methods in the tourism e-commerce, and then presents a design of web mining based tourism e-commerce recommender system with offline and online modules [7].

Knowledge bases, which consist of concepts, entities, attributes and relations, are increasingly important in a wide range of applications. We argue that knowledge about attributes (of concepts or entities) plays a critical role in inference. In this paper, we propose methods to derive attributes for millions of concepts and we quantify the typicality of the attributes with regard to their corresponding concepts. We employ multiple data sources such as web documents, search logs, and existing knowledge bases, and we derive typicality scores for attributes by aggregating different distributions derived from different sources using different methods. To the best of our knowledge, ours is the first approach to integrate concept- and instance-based patterns into probabilistic typicality scores that scale to broad concept space. We have conducted extensive experiments to show the effectiveness of our approach [8].

III. Proposed Work

This paper first introduces the theory and knowledge related to Web logs, and then introduces a Web log mining process. As to this article is based on Web log processing, we draw valid user data according to a specific website users' access data through the pre-processing, and then research and analyze the users behaviours. The work can provide a theoretical basis for the management and optimization of the site for site managers. Experiments have showed that the work is effective.

From the above discussion and work in base paper it is found that the web usage mining has been a basic requirement for the attracting extensive traffic for the websites. A proper analysis of the web usage mining leads to multiple advantages:

1. Finding pages on the current website with errors
2. Finding pages which are visited most
3. Time of highest traffic
4. For incorporating the proper and required contents for the users of the website

From the work in the base paper it is found that they have focused their research on finding the number of users visited the website day wise, number of independent users of the site and click rate of the pages of the website day wise.

I am extending this work to provide more analyzed data with high efficiency and accuracy. The proposed work shall be having the following steps:

1. Sample data of web usage will be loaded in the software
2. Data Pre-processing: Data Cleaning will be performed to remove various unwanted data for which following three steps will be applied:
   - Data Stemming
   - Data Stopping
   - Data Scrubbing
3. Data Mining: Users web access shall be analysed on the basis of the following parameters
   - IP Address
   - URL
   - Time Spent
   - Http Method Used
   - Status Message Sent
   - Bytes Returned
4. Data Post Processing: The analysed data shall be used to formulate the results and graphs for showing the performance of the proposed algorithm and enhancements achieved in respect of the work in base paper and other researches.

IV. Results & Discussion

The implementation of the proposed work has been done using Microsoft Visual Studio and results have been obtained for different characteristics stated in proposed work as follows:

A. Cluster Frequency Counts

Table 2: Cluster Frequency Based on IP Address

Day 01	76
Day 02	90
Day 03	99
Day 04	86
Day 05	92
Day 06	75
Day 06	89

Inference: The frequency of the cluster items based on IP Address is an indicator of repetitive access of server from a particular IP. Access frequency based on IP Address can lead the administrators to check for the possible denial of service attacks or information theft.

Table 3: Cluster frequency based on Page Status

Day 01	2
Day 02	3
Day 03	2
Day 04	2
Day 05	2
Day 06	3
Day 06	2

Page Status is to check the different web pages working properly or throwing errors. Such clusters can be analyzed to find out the frequency of the erroneous page hits on the server.

Table 4: Cluster frequency based on Size of Response

Day 01	27
Day 02	26
Day 03	32
Day 04	32
Day 05	29
Day 06	23
Day 06	34

Size of response shows the traffic being delivered to the clients. It is a good measure for finding the busy hours on the site and possible clients who are trying to steal lot of information from the server.

B. Time Taken in Processing of Various Steps:

Table 5: Time taken in different steps

Processing Step	Time Taken in Milli Seconds
Data Loading Time	1744
Data Cleaning Time	818
Ant Clustering Time	1495

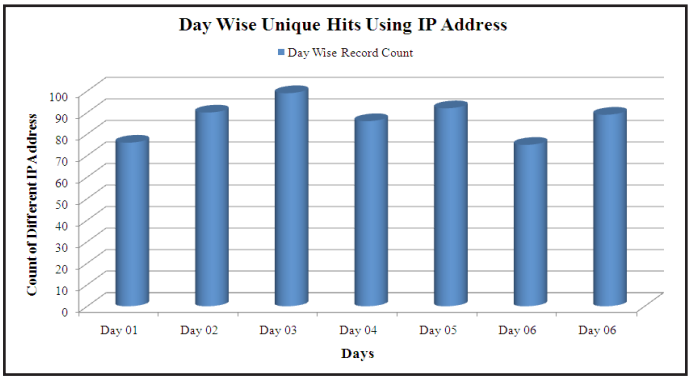

Fig. 2: Day Wise Unique Hits Using IP Address

The graph drawn from the data fetched shows that for the given dataset average days wise hit of the server using different IP Address is not varying too much and hence the there are no possible attacks related with denial of service.

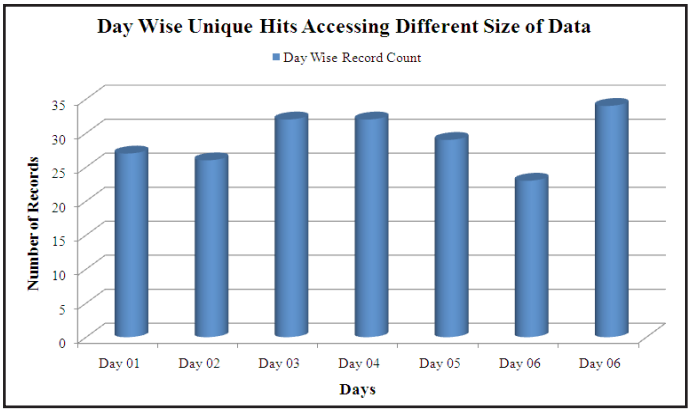

Fig. 3: Day Wise Unique hits using Size of Data

Size of data based day wise hits shows that the largest data fetched from the server is on 7th day but the hit of the data is also not beyond limits.

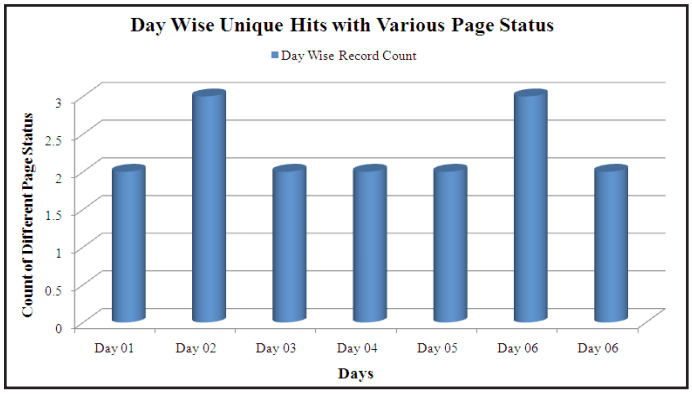

Fig. 4: Day wise Unique hits using Page Status

Day wise unique hits are a measure which shows how good a website is hit every day. From the graph it is clear that on day 2 and 6 maximum number of unique hits have been recorded.

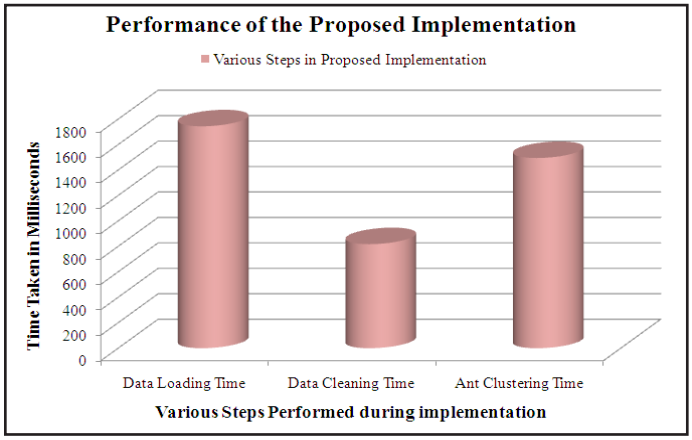

Fig. 5: Performance of the Proposed System

The above graph show that the maximum time taken in execution of the proposed implementation is in loading of the data in system. Time taken in performing Ant Clustering is also high but it is justifiable as the processing is also quite complex for each Ant and still the time taken is less in respect of the linear or manual clustering.

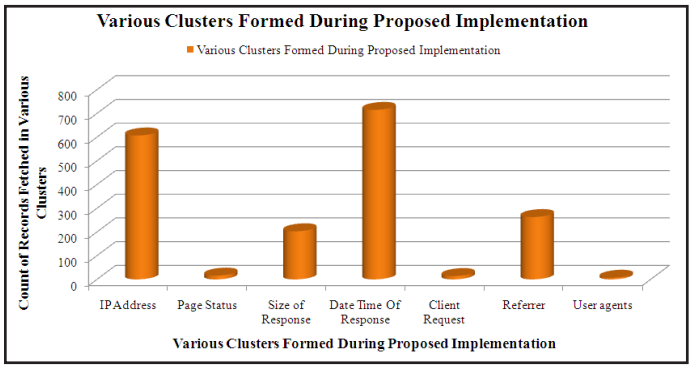

Fig. 6: Various Clusters Formed during Proposed Implementation

The graph above shows the comparison of the unique data clusters formed during the processing and it is found that the clusters formed from the date time of response is very high in respect of others as the time of hits on server varies continuously. Some of the clusters are very less indicating that the system is performing well in those cases such as in case of page status based clusters.

VII. Comparison With Existing System

Since the work in base paper is not evaluating the performance of the system, therefore comparison is beind based on number of users identified based on IP Address. Number of unique visitors identified in access log gives an idea that how frequently a website is used by the different users around the world. For comparison purpose data normalization has been applied to compare the results of the existing system in base paper and work done in this paper.

Table 6: Comparison of the Work

WORK	Total Records	Number of unique visitors
Existing Work	64576	7597
Proposed Work	64576	39197

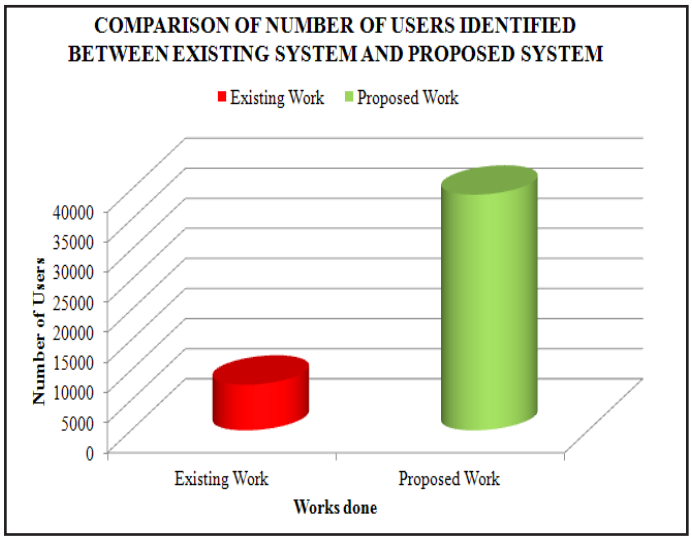

Fig. 7: Comparison of Number of Users Identified Between Existing System And Proposed System

VIII. Conclusion

In this paper, we propose a new perspective of Web Usage Mining. This work proposes to use Ant Clustering mechanism to retrieve the web usage by the various users, IP Addresses, page information accessed and others.

As per the expectation, to provide the high efficiency and least time consumption as Ant Clustering is applied using various processes working in parallel are found to be generating the output quickly. This algorithm shall be more accurate as each process works on exactly one characteristic and will only retrieve details related with it, therefore the overall accuracy provided will be better and will have high throughput as well. The work is found to be providing better performance and accuracy from the existing works.

In future work can be applied on other different data sets and on real time systems to test the performance and accuracy. The work also can be applied over cloud systems to test accuracy and efficiency.

References

[1] Jia Li,"Research of Analysis of User Behavior Based on Web Log", 2013 International Conference on Computational and Information Sciences, 2013 IEEE.

[2] Hummera Naz, Yasir Hafeez Motla, Sohail Asghar, Mehmood Ahmed, M.Shabbir Hassan,Mehwish Mukhtar,Abida Javed, "A Systematic Approach for Web Engineering Practices by Integrating Data Mining Technique with Requirement Change Management", 2013 IEEE.

[3] Nawal Sael, Abdelaziz Marzak, Hicham Behja,"Web Usage Mining Data Preprocessing and Multi Level Analysis on Moodle", 2013 IEEE.

[4] Yanduo Zhao,"The Review of Web Mining in E-commerce", 2013 International Conference on Computational and Information Sciences, 2013 IEEE. 2013.158

[5] Neeraj Sahu, Krishna Kumar Mohbey, G. S. Thakur, "Document Clustering using message passing between data points", 2013 International Conference on Communication Systems and Network Technologies, 2013 IEEE.

[6] Sudheer Reddy, K.; Kantha Reddy, M.; Sitaramulu, V.,"An effective data preprocessing method for Web Usage Mining," Information Communication and Embedded Systems (ICICES), 2013 International Conference on , vol., no., pp. 7, 10, 21-22 Feb. 2013.

[7] Xuesong Zhao; Kaifan Ji,"Tourism e-commerce recommender system based on web data mining," Computer Science & Education (ICCSE), 2013 8th International Conference on, pp. 1485-1488, 26-28 April 2013.

[8] Taesung Lee; Zhongyuan Wang; Haixun Wang; Seung-won Hwang,"Attribute extraction and scoring: A probabilistic approach," Data Engineering (ICDE), 2013 IEEE 29th International Conference on, pp. 194-205, 8-12 April 2013.

[9] Arun K Pujari: Data Mining Techniques, Universities Press (India) Private Limited 2001.

[10] Jhon A. Hartigan: Clustering Techniques, Willey Publications 2005.

[11] Jaiwai Han, Micheline Kamber,"Data Mining concepts and Techniques, Elsevier 2006.

[12] Gloria Bordongna and Gabriella Pasi, National Research Council- IDPA, Dalmine (BG), Italy and University degli studi di Milono Bicocca, Italy. "Hierarchical Data Divisive Soft Clustering algorithm"

[13] Liang Feng, Ming- Hui Qiu, Yu- Xuan Wang, Qiao- Liang Xiang, Yin- Fei Yang, Kai Liu, ECHO Laboratory, School of Communication and Information Engineering, Nanjing University of Posts and Telecommunication, Nanjing, Jiangsu, China and School of Computing, National University of Singapore,Singapur. "A fast divisive clustering algorithm"

[14] Arindam Banerjee Chase Krumpelman Joydeep Ghosh , Dept. of Electrical and Computer Engineering University of Texas, Austin, USA and Sugato Basu Raymond J. Mooney, Dept Of Computer Sciences University of Texas at Austin, USA." Model based Overlapping Clustering"

[15] Yanfei Zhao, "Study on Web Data Mining Based on XML", 2012 International Conference on Computer Science and Information Processing (CSIP), 2012 IEEE.

[16] Xingyuan LI, Ningbo,China, Yanyan Wu, PING CHENG, "Research of Business Intelligence based on Web Accessing Data Mining", The 7th International Conference on Computer Science & Education (ICCSE 2012) July 14-17, 2012. Melbourne, Australia, 2012 IEEE.

[17] Ning Zhong, Yuefeng Li, and Sheng-Tang Wu, "Effective Pattern Discovery for Text Mining", Published by the IEEE Computer Society, IEEE Transactions on Knowledge and Data Engineering, Vol. 24, No. 1, January 2012, IEEE.

[18] Weigang Zuo,.Qingyi Hua, Weigang Zuo,"The application of Web data mining in the electronic commerce", 2012 Fifth International Conference on Intelligent Computation Technology and Automation, 2012 IEEE.

[19] Jianli Duan, Shuxia Liu,"Research on web log mining analysis", 2012 International Symposium on Instrumentation and Measurements, Sensor Network and Automation (IMSNA), 2012 IEEE.
[20] Jinyue Yang.Lin Yang,"Customers's intelligence: Kernel of CRM[J] Modernization of Management, 2002-07