Service Oriented Web Data Mining Using Xml

[1]Reetesh Rai, [2]Nitin Shukla
[1,2]Dept. of CSE, Shri Ram Institute of Technology, Jabalpur, India

Abstract
With the rapid development of science and technology, it will be competitive trends of modem society that a large number of the Internet information is analyzed in real time and multi-level. In view of the Web with the characteristics: openness, dynamic nature, heterogeneity and so on, Accurately finding the information you need from the scattered and unified management of massive amounts of data become a difficulty solve by Web mining. However, Web-oriented data mining is more complex than for a single data warehouse.
The WWW environment based on XML is the face of Web data. XML can be compatible with existing Web applications and Web information sharing and exchanging. Due to the emergence of XML technology, it provides a standard for data exchange on the Internet. At the same time, from the perspective of data, XML technology provides that can represent the means of the data content and meaning. So data mining based on XML technology provides new opportunities for data mining researching.
The algorithm being proposed is a dynamic and novice algorithm for the web data mining using XML. The proposed work shall provide an opportunity to the user to select the particular domain of web data as per his requirements and the implantation work shall apply the mining on the XML data converted from the web data.
Web is a major resource where data changes rapidly with time and therefore they are dynamic in nature. Web data is utilized by the users, advertisers, and search engines and therefore researchers have lot of responsibilities to provide fast and efficient mechanism for providing required information retrieval algorithms.
This work is offering to create clusters dynamically from web data and XML.

Keywords
Web Data Mining, Clustering, XML, XPATH, XSL, ANT Clustering

I. Introduction

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 Ttypes of Web Data Mining

The template is used to format your paper and style the text. All margins, colunm 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.

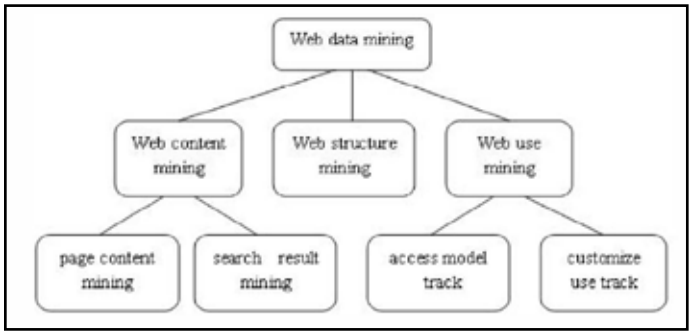

Fig. 1: Types of Web Data Mining

B. Application of XML in Web Data Mining

XML is well applied in the Web because it provides a good method for data processing in three-tier architecture. The advantage of three-layer model is you can upgrade. Making use of the upgraded three-tier model, XML can be generated from the data in the database. You can separate from the commercial norms and forms using the XML structured data.
The new-generation WWW environment based on XML is direct to Web data, It not only can be a well compatible with existing Web applications, but also can better achieve information sharing and exchange in the Web[5]. Developers can use XML formatting tags to exchange data. The XML document description can easily correspond to the attributes in the relational database to implement the exact query and model extraction.
The application of XML in Web mining mainly has the following four categories:

1. Data Exchange

In the Web data mining, XML provides a connection between relational database, object-oriented database and other database management systems. It is convenient that customers often deal with the different databases to exchange data. The Self-description and extensibility is enough to indicate the various types of data, it can naturally describe the record of data collected from the site of the WEB page. After clients receive the data, they can process and transfer them between different databases.

2. Integration of Heterogeneous Data

XML enables structured data of different sources to be easily combined. To a certain extent, the XML is considered as a semi-structured data model. With XML, it is easily correspond it to the attributes in the relational database to implement the exact query and model extraction. Because of this advantage, the unstructured data from different source is integrated together, and the problems of the incompatible background databases are resolved. So XML

resolves the problem of a variety of incompatible database, it makes the unstructured data of different sources can be easily combined.

3. Reducing the Information Content According to Individual User

In view of XML, it can cut and edit the information obtained to satisfy the needs of different users. XML separate the user interface and provide the same data with different browsers to different users. If your browser can display XML, you can directly send the XML document to the browser, or use XSL to translate XML into content witch your web browser can handle.

4. Mitigating the Burden on the Server

XML can conveniently deliver the most handlers from the Web server to the Web client, so clients choose and produce suitable data processing and display program according to their needs, and the server only has to provide the complete and correct XML data file.

C. Web Data Mining System Architecture Based on XML

System can generally be divided into three layers. The underlying layer is XML data integration layer, it integrate and extract the relevant data using XML as a tool to form the original XML data sets of a structural information and transfer data to the middle layer. The data selection, cleaning and standardization, resulting in the structure degree higher with rich semantic XML data set is done in the middle layer, and these data will be as the top data mining application layer data source. There are some specific data mining applications need to show us the results through the form of reports, charts, etc in the data mining application layer. It is shown in figure

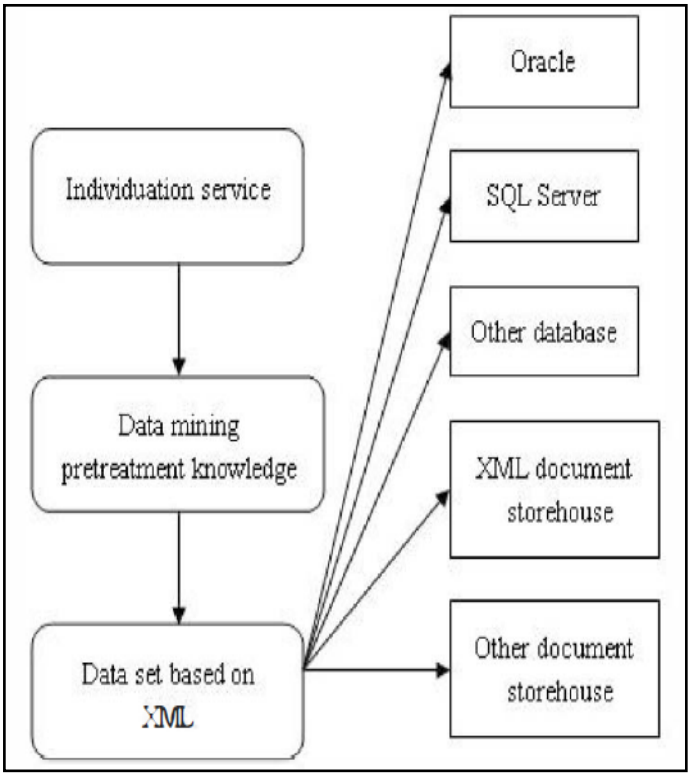

Fig. 2: Web Data Mining System Architecture Based on XML Existing System

With the continuous development of science and technology, more and more database and information systems connect with the network. That is to say how to discover the required information from complex network data has become a more and more important concern problem. Data mining refers to the practical application of data from which implicit information and knowledge are extracted to analysis and use. Data mining from Web-Oriented is more complex than that from a single data warehouse, XML technology provides a convenient way for the Web data mining. Web-oriented data mining and it's types, XML technology, the transforming from HTML into a model of XML and the application of XML in Web Data Mining are described in this paper. And then an example is adopted to illustrate the application of the model [1].

With the continuous development of science and technology, more and more database and information systems continue to join the network. In the face of such complex Web data, we should excavate required information from the complex network data, and it has become an important issue which we concerned. XML can transfer the structured data to query and extract Web information. Web Oriented data mining is a complex technology, Web data mining is more complex than a single data warehouse mining. XML technology provides convenience to Web data mining. XML can facilitate structured data binding of the different sources. It is possible to search a variety of incompatible database. With XML as the emergence of a standard way for data exchange on the Web, Web-Oriented data mining has become very convenient [1].

In this paper, the basic characteristic of web access information puts forward the improved data mining algorithm in the application of e-commerce industry. Deep analysis of web access to information, it comes from solving real problem from the design of web based access information mining application model. It puts forward several key technical problem methods. for solving the problems [2].

Web access information is large data information. It is various and updating, and remember the visitors, the visited web, the visiting time and so on. Some meaningful data are applied in the electronic commerce environment according to mining algorithm. And get valuable commercial intelligence information about electronic business operation manage. Such as identifying the user's speciality and forecasting the interest of potential customers by remembering the web visitors' information. It explores the visiting data in the particular period. At last it finds the commonality of colony visiting' behaviors and the potential customers' information. So then update the web structure, offer some relational base for the strategies of web electronic business operator [2].

Many data mining techniques have been proposed for mining useful patterns in text documents. However, how to effectively use and update discovered patterns is still an open research issue, especially in the domain of text mining. Since most existing text mining methods adopted term-based approaches, they all suffer from the problems of polysemy and synonymy. Over the years, people have often held the hypothesis that pattern (or phrase)-based approaches should perform better than the term-based ones, but many experiments do not support this hypothesis. This paper presents an innovative and effective pattern discovery technique which includes the processes of pattern deploying and pattern evolving, to improve the effectiveness of using and updating discovered patterns for finding relevant and interesting information. Substantial experiments on RCV1 data collection and TREC topics demonstrate that the proposed solution achieves encouraging performance [3].

Many data mining techniques have been proposed in the last decade. These techniques include association rule mining, frequent itemset mining, sequential pattern mining, maximum pattern mining, and closed pattern mining. However, using these discovered knowledge (or patterns) in the field of text mining is difficult and ineffective. The reason is that some useful long patterns with high specificity lack in support (i.e., the low-frequency problem). We argue that not all frequent short patterns are useful. Hence, misinterpretations of patterns derived from data mining techniques lead to the ineffective performance. In this research work, an effective pattern discovery technique has been proposed to overcome the low-frequency and misinterpretation problems for text mining. The proposed technique uses two processes, pattern deploying and pattern evolving, to refine the discovered patterns in text documents. The experimental results show that the proposed model outperforms not only other pure data mining-based methods and the concept based model, but also term-based state-of-the-art models, such as BM25 and SVM-based models [3].

With the increasing of the information on Internet, more and more electronic data are appearing. Then, how should we immediately discover useful knowledge and improve the utilization rate of information without being confused in the sea of information? Data mining come up with a new way of dealing with such problem. This paper sets force web data mining sources in e-commerce, the flow process and some techniques in dealing with web data mining. Finally, analyses the functions of web data mining used in e-commerce [4].

The application of web data mining technology in ecommerce plays an important role. This article introduces the data source, data mining technology and its applications of data mining in e-commerce. The web data mining technology focused on finding the valuable knowledge in the mass network heterogeneous information resources. In recent years, with the rapid development of e-commerce, web data mining is becoming more useful. It can automatically predict the customer's expenditure trends and the market trends, also help businessman obtain and retain customers, adjust the marketing strategy, make the right decisions, to promote the development of electronic commerce. The combination of the Web data mining technology and electronic commerce will help the enterprises identify target markets more effectively and improve decision-making to gain competitive advantage. It has a very broad application prospects [4].

With the rapid development of Internet, web data mining, especially weblog mining plays an important role in many fields, including personalized information service, improving designs, services of websites and so on. This paper introduces web data mining firstly, and then discusses the process of weblog mining. Based on these studies, the experiment research-a web log mining tool is presented in detail. The conclusion of the research and the direction of further study are pointed out in the last part of this paper [5].

Web log mining is an important branch of web data mining. In this paper, we design a web log mining tool to help website analysis experts to understand user behaviors by mining web log data and help them find usability problems. Based on web log mining theory and web usability engineering, we preprocess web log files to get click stream data firstly, and then choose a sequential pattern mining algorithm for mining visitor access patterns. The result of pattern analysis is represented in visual spatial forms. Finally we try to use the tool, and choose some real websites to see the effect. On one hand this tool can help website analyze web log files, and make the analysis work much easier, but on the other hand there are some limitations in it, for example the data about users come from log files, in which only the information about links that users access are kept orderly, so the tool can't simulate the actions of users exactly. So much work remains to do to improve the tool, for example, to remove the distortion of the users' browsing data map, we can try to get more data from the clients' side [5].

III. Proposed Work

Ant clustering technique in data mining is a novel approach which uses Ants technique to find the relevant information and put them in various clusters. Since several Ants work in parallel therefore the processing speed of the system is high and in case of large data sets it is worth using Ant clustering to apply.

The algorithm being proposed is a dynamic and novice algorithm for the web data mining using XML. The proposed work shall provide an opportunity to the user to select the particular domain of web data as per his requirements and the implantation work shall apply the mining on the XML data converted from the web data.

Web is a major resource where data changes rapidly with time and therefore they are dynamic in nature. Web data is utilized by the users, advertisers, and search engines and therefore researchers have lot of responsibilities to provide fast and efficient mechanism for providing required information retrieval algorithms.

This work is offering to create clusters dynamically from web data and XML as follows:

Step 1: A user Interface shall be created to select the websites from which data is to be extracted for web data mining.

Step 2: From the links provided by the user, other links will be tracked and data will be retrieved to perform mining.

Step 3: Data Filtering: Extracted data shall be filtered before converting it to XML data on the basis of the XML attributes selected for the particular data.

Step 4: Data Conversion: Filter data is in HTML format and will be converted into XML format using characteristics provided by the user / retrieved from headings in the web page.

Step 5: Data Mining: Users will be providing the clustering characteristics for the data which will be used for data clustering. ANT Clustering Algorithm shall be used for providing the clusters to the users.

Step 6: whole algorithm will be as follows:

- Data Cleaning
- Data Preprocessing Using Stopping & Stemming
- Data Mining Using ANT Clustering
- Data Extraction from online web sites.
- Data Filtering and Conversion into XML format.

Step 7: The proposed Algorithm will be useful for all types of users for the web mining and will be compared with the existing systems using:

- Time Taken
- Size of data used for simulation

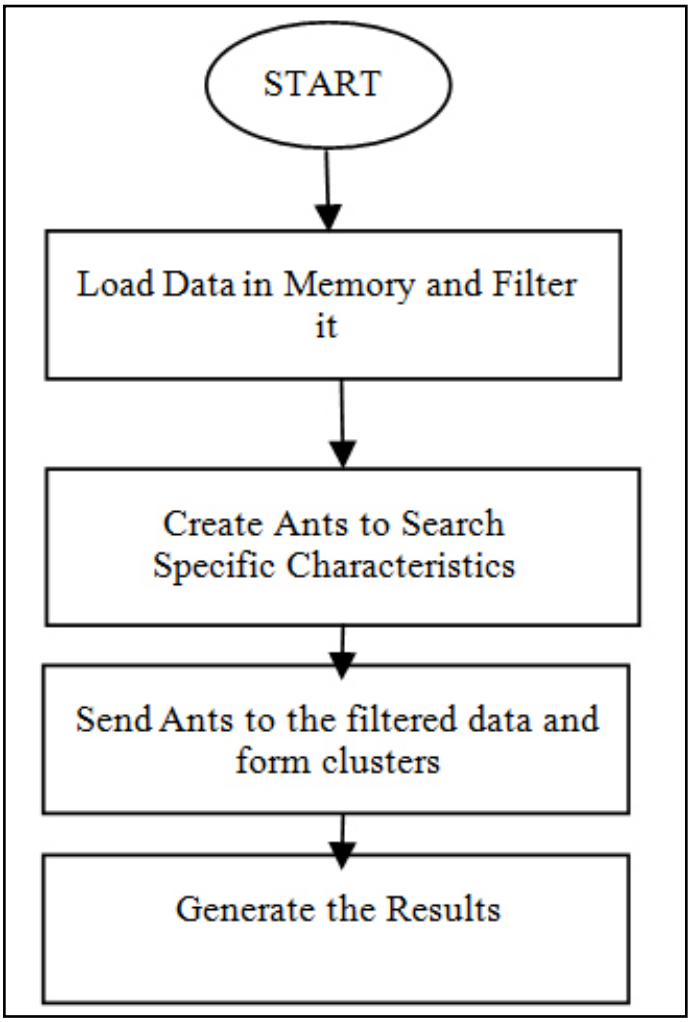

Fig. 3: Flow Chart of Proposed Work

Flow Chart above shows the working of the proposed algorithm, which includes the step by step procedure as follows:

- Dataset is required to be loaded using dataset of web knowledge base data set.
- Data Preprocessing step is to applied on loaded data which is done in second step
- Ant Clustering Technique is applied for all the characteristics i.e. for each characteristic, there is one Ant created which iterates over the preprocessed data set.
- Data filtering is done for creating the clusters.

VI. Results & Discussion

Readings have been obtained with the sample dataset of web->kb by proposed algorithm implementation as follows:

Table 2: Reading Taken from the Implemented Work

Timing Types	Time in milliseconds
Data Loading Time	949 ms
Data Filtering Time	295 ms
Characteristics Retrival Time	395 ms
Time Taken in XML Conversion	916 ms

Table 3: Reading for Comparison of the Proposed Work With Existing Algorithm

Cluster Name	Frequency of Characteristics
Project	5396
Department	2147
Staff	1253
Faculty	1726
Student	1230
Other	1182
Course	2178

Various graphs are being drawn using the results obtained in table 1 which includes time efficiency of ANT Clustering, distribution of data in created clusters.

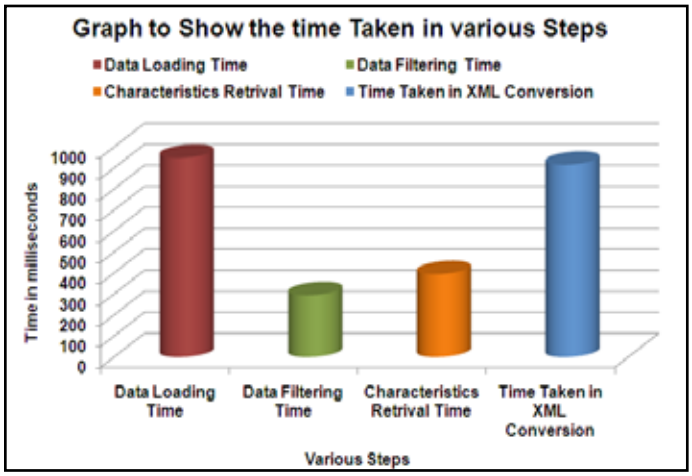

Fig. 10: Graph Showing Time Taken in Processing in Various Phases of the Proposed Algorithm Implementation

Inference: The above readings & Graphs show the time in milliseconds required for various steps of proposed work. It is seen from the graph that the major time is required in loading data and converting in clustered data in XML format. Time required in filtering and retrieving characteristics is less. This is an indicator of fast processing provided by the Ant Clustering algorithm.

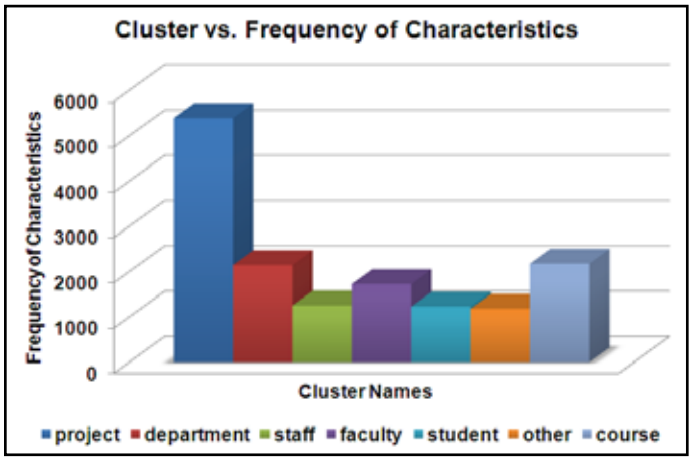

Fig. 10: Graph Showing Time Taken in Processing in Various Phases of the Proposed Algorithm Implementation

Inference: Various clusters have been formed from the various directory names where the files are grouped for particular clusters and the characteristics are retrieved from all of the files as one bunch. Frequency of the words in characteristics retrieved are

indicated in the above table and graph and shows that even when the words found in project category are more still the clustering is performed exactly as per the cluster provided to the system. This means that even if the files of the various clusters are mixed into other clusters, they are grouped in the same cluster to which they exactly belong to.

VII. Conclusion

In this paper, we propose a new perspective of Web Mining through XML. Based on this, a novel WWW-oriented web recommendation system is proposed and shall be implemented. With the explosive growth of the World Wide Web, the amount of information available on-line is increasing rapidly. This certainly provides users with more options, but also makes it difficult to find the "right" or "interesting" information today. Web mining discovers user preference from the available data automatically and makes recommendations based on the extracted knowledge. More recently, a combination of web content, web structure and web usage mining has been studied and shows superior results in web recommendations.

From the comparison of the work done using various data mining algorithm for intrusion detection and the work done using proposed work is shown in the above results and it is found that the proposed work is providing high detection rate in comparison with other algorithms. Ant clustering technique has been found to be more accurate and provides high performance. The reason for the improved result is due to its parallel processing mechanism.

In one of the base papers work, an effective pattern discovery technique has been proposed to overcome the low-frequency and misinterpretation problems for text mining. The proposed technique uses two processes, pattern deploying and pattern evolving, to refine the discovered patterns in text documents. The experimental results show that the proposed model outperforms not only other pure data mining-based methods and the conceptbased model, but also term-based state-of-the-art models, such as BM25 and SVM-based models.

In this proposed research work it has been found that the effective recovery of patterns is optimum and also efficiency of the clustering has been emphasized to produce the clusters within the optimum time period. Frequency of the words as shown in figure two above is not relevant to define the clusters.

References

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

[2] 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

[3] Ning Zhong, Yuefeng Li, 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, 1041-4347, 2012 IEEE.

[4] Jinyue Yang.Lin Yang,"Customers's intelligence: Kernel of CRM[J] Modernization of Management", 2002-07

[5] Lin Yang,"Basic knowledge of customer intelligence[J]", ChinaComputer Users, 2003-12

[6] Gaofeng Zhang,"Applicational research about getting customer, intelligence through data mining[J]", China Doctor Dissertation Full-text Database.

[7] Xiaoping Zheng,"NET essence—web service principle and explore, [M]", China Pub, 2002.

[8] S. Branson, A. Greenberg, "Clustering Web Search Results Using Suffix Tree Methods", Stanford University, unpublished.

[9] Adetokunbo Makanju, A. Nur Zincir-Heywood, and Evangelos E. Milios,"A Lightweight Algorithm for Message Type Extraction in System Application Logs", IEEE transactions on knowledge and data engineering, Vol. 24, No. 11, November 2012, IEEE.

[10] Kuang Yu Huang,"A hybrid particle swarm optimization approach for clustering and classification of datasets", Department of Information Management, Ling Tung University, Ling Tung Road, Taichung City 408, Taiwan Knowledge-Based Systems 24, 2011.

[11] Rui Xu, Jie Xu, Donald C. Wunsch, II,"A Comparison Study of Validity Indices on Swarm-Intelligence-Based Clustering", IEEE transactions on systems, man, and cybernetics—part b: cybernetics, Vol. 42, No. 4, August 2012, 1083-4419/ 2012 IEEE

[12] [Online] Available: http://www.en.wikipedia.org/wiki/User-generated_content.

[13] Yongli Liu, Qianqian Guo, Lishen Yang, Yingying Li, "Research on Incremental Clustering", 2012 IEEE.

[14] Y. Liu, Y. Ouyang, Z. Xiong,"Incremental Clustering using Information Bottleneck Theory", International Journal of Pattern Recognition and Artificial Intelligence, 2011, 25(5), pp. 695-712.

[15] X. Wan,"A novel document similarity measure based on earth mover's distance", Information Sciences, 2007, 177(18), pp. 3718-3730.

[16] K.M. Hammouda, M. S. Kamel,"Incremental document clustering using cluster similarity histograms", In Proc. of Int. Conf. on Web Intelligence, 2003, pp. 597-601.

[17] O. Zamir, O. Etzioni,"Web document clustering: A feasibility demonstration", In Proc. of the 21st Annual Int. ACM SIGIR Conf., 1998, pp. 46-54.

[18] W. Wong and A. Fu,"Incremental document clustering for Web page classification", In Proc. 2000 Int. Conf. Information Soc. In the 21st Century: Emerging Technologies and New Challenges (IS2000), 2000.

[19] S. Noam, T. Naftali,"Document clustering using word clusters via the information bottleneck method", In Proc. 23rd Ann. Int. ACM SIGIR Conf. Research and Development in Information Retrieval, 2000, pp. 208-215.

[20] [Online] Available: http://www.en.wikipedia.org/wiki/K-nearest_neighbor_ algorithm.

[21] S. Branson, A. Greenberg,"Clustering Web Search Results Using Suffix Tree Methods", Stanford University, unpublished.

[22] M. Ester, H. Kriegel, J. Sander, X. Xu ,"A density-based algorithm for discovering clusters in large spatial databases with noise", In Proc. of the 2nd Int. Conf. on Knowledge Discovery and Data Mining, 1996, pp. 226-231.

[23] Tu. Nguyen-Hoang, K. Hoang, D. Bui-Thi, A. Nguyen, "Incremental Document Clustering Based on Graph Model", Advanced Data Mining and Applications, 2009, pp. 569-576.

[24] S. Noam, F. Nir, T. Naftali,"Unsupervised document classification using sequential information maximization", In Proc. of the 25th Ann. Int. ACM SIGIR Conf. Research

and Development in Information Retrieval, 2002, pp. 129-136.

[25] Karel Dejaeger, Wouter Verbeke, David Martens, Bart Baesens,"Data Mining Techniques for Software Effort Estimation: A Comparative Study", IEEE TRANSACTIONS ON SOFTWARE ENGINEERING, Vol. 38, No. 2, March/ April 2012, IEEE

[26] Markou, M., Singh, S.,"Novelty Detection: A review", Part 1: Statistical Approaches, Signal Processing, 8(12), 2003, pp. 2481- 2497.

[27] M. Basavaraju, Dr. R. Prabhakar of Research Scholar, Dept. of CSE, CIT, Anna University, Coimbatore, Tamilnadu,"A Novel Method of Spam Mail Detection using Text Based Clustering Approach", International Journal of Computer Applications, Volume 5– No.4, August 2010

[28] X Z Wang, S A Yang, S H Yang et ai,"The Application of Fuzzy Qualitative Simulation in Safety and Operability Assessment of ProcessPlants[J].Computers Chern Engng, 1996, pp. 671-676.

[29] Hoffer J A, Prescott M B,McFadden F R,"Modem Database Management [M]", 8th ed.New Jersey:Pearson Prentice Hall, 2007, pp. 501-502.

[30] Yang Liu, Tianshuang Qiu, Fuquan Ren, Xianyao Yu, "Robust Optimal Filtering Method for Cyclostationary Signals", Procedia Engineering, Vol. 29, pp. I889-1896, February, 2012.

[31] ZHOU Xiao-mei, WANG Qian-ping, SU Lin,"Design of web mining model based on XML", Computer Engineering and Design, Vol. 28, 2007, pp. 272-274.

[32] CUI Jianqun, HE Yanxiang, ZHENG Shijue, WU Libing, "Research on Key Technologies of Web Mining Based on XML", Computer Engineering, Vol. 20, 2006, pp. 43-44.

[33] Wang Yuzhen,"Web data Mining Technology and XML", Information Technology, Vol. l0, 2005, pp. 142-143.

[34] 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.

[35] Jianli Duan, Shuxia Liu,"Research on web log mining analysis", International Symposium on Instrumentation and Measurements, Sensor Network and Automation (IMSNA), 2012 IEEE.