International Journal on Emerging Technologies **2**(1): 83-86(2011) **ISSN : 0975-8364**

Survey on Mining Diverse Patterns from Web Data

Mamta K. Varma* and Rajesh Nigam*

**Department of Computer Science and Engineering, Technocrats Institute of Technology, Bhopal, (MP)*

(Received 11 March, 2011 Accepted 11 April, 2011)

ABSTRACT : Outliers are the data objects with different characteristics compared to other data objects. Outliers are also referred to as diverse patterns or noise. Outlier detection and analysis is an interesting data mining task, referred to as Outlier Mining. Web outliers are data objects that show significantly different characteristics than other web data. This paper has focus on how the study from researchers grows from web mining to various techniques for finding diverse patterns(outlier mining) from the web including some applications. This paper reviews about 24 papers and presents the detailed discussion on the topic.

Keywords : Outliers, Web Outliers, Web Mining, Outlier Mining.

I. INTRODUCTION

An Outlier is an observation that deviates so much from other observations so that it arouses suspicion that it is generated by a different mechanism. Data objects that show significantly different characteristics from remaining data are declared outliers. Outlier detection and analysis is an interesting data mining task, referred to as Outlier Mining that has a lot of practical applications in many different domains. A web content outlier is defined as web document(s) having different contents from similar web documents taken from the same category. Many data mining algorithms consider outliers as noise that must be eliminated since it degrades their predictive accuracy. However, as pointed out in, "one person's noise could be another person's signal", thus outliers themselves can be of great interest. The outlier mining can be used in Telecom or Credit Card frauds to detect the atypical usage of telecom services or credit cards. In contrast to traditional data mining task that aims to find general pattern applicable to the majority of data, outlier detection targets the finding of the rare data (nuisance, noise, or outliers) whose behavior is very exceptional when compared with rest large amount of data. Formally Outlier mining can be described as:

Given a set of *n* data points or objects, and *k*, the expected number of outliers, find the top *k* objects that are considerably dissimilar, exceptional, or inconsistent with respect to the remaining data.

Studying the extra ordinary behavior of outliers helps uncovering the valuable knowledge hidden behind them and aiding the decision makers to make profit or improve the service quality.

II. RESEARCH SURVEY

Research initiates from the introductory concepts of web and web mining in the period of 1999. Web mining is a new research issue which draws great interest from many communities. There is no agreement about Web mining yet. Wang *et. al.* [1] presents a preliminary discussion about Web mining, including the definition, the relationship between information mining and retrieval on the Web, the taxonomy and the function of Web mining. In the year 2001, Masaru *et. al.* [2] discussed that the Web mining can be classified into two categories, Web access log mining and Web structure mining. They performed association rule mining and sequence pattern mining against the access log which was accumulated at NTT Software Mobile Info Search portal site. Detail web log mining process and the rules derived are reported in this paper.

Dongkwon and Songchun [3] discussed that for web mining, the biggest problem is the scarcity of data. To overcome the problem and prepare as many needed data as possible for business intelligent information, they propose backward induction in web mining. Web mining itself is an iterative process where data mining techniques are used back and forth and iteratively. To support backward induction and web mining characteristics, the scalable web mining architecture in data warehouse environment is proposed. The proposed web mining architecture has three kinds of scalabilities. These are the scalabilities of operational database, the scalabilities of data model and the scalabilities of data mining engines. By implementing this scalable web mining architecture having three kinds of scalabilities in data warehouse environment to support backward induction procedures, we can extract the business intelligent information from web mining.

Charu *et. al.* [4] discussed the Outlier Detection for High Dimensional Data. In high dimensional space, the data is sparse and the notion of proximity fails to retain its meaningfulness. In fact, the sparsity of high dimensional data implies that every point is an almost equally good outlier from the perspective of proximity based definitions. Consequently, for high dimensional data, the notion of finding meaningful outliers becomes substantially more complex and non-obvious. This paper gave new techniques for outlier detection which find the outliers by studying the behavior of projections from the data set.

Anny *et. al.* [8] discussed the enhancements on Local Outlier Detection. Outliers, or commonly referred to as exceptional cases, exist in many real-world databases. Detection of such outliers is important for many applications. This paper focuses on the density-based notion that discovers local outliers by means of the Local Outlier Factor (LOF) formulation. Three enhancement schemes over LOF are introduced, namely LOF', LOF''and GridLOF.

Malik *et. al.* [9] developed the Framework for Mining Web Content Outliers. This paper refers to outliers present on the web as web outliers to distinguish them from traditional outliers. Traditional outlier mining algorithms designed solely for numeric data sets are inappropriate for mining web outliers. This paper establishes the presence of web outliers and discusses some practical applications of web outlier mining. Finally, this paper present taxonomy for web outliers and propose a general framework for mining web content outliers. Malik *et. al.* [10] took advantage of the HTML structure of web and N-gram technique for partial matching of strings and propose an N-gram based algorithm for mining web content outliers. Classifying text into predefined categories is a fundamental task in information retrieval (IR). IR and web mining techniques have been applied to categorize web pages to enable users to manage and use the huge amount of information available on the web. Thus, developing user-friendly and automated tools for managing web information has been on a higher demand in web mining and information retrieval communities. Text categorization, information routing, identification of junk materials, topic identification and structured search are some of the hot spots in web information management. A great deal of techniques exists for classifying web documents into categories. Interestingly, almost none of the existing algorithms consider documents having 'varying contents' from the rest of the documents taken from the same domain (category) called web content outliers. Experimental results indicate the proposed N-gram-based algorithm is capable of finding web content outliers.

Fabrizio and Pizzuti, [13] described the outlier mining in Large High-Dimensional Data Sets. In this paper, a new definition of distance-based outlier and an algorithm, called HilOut, designed to efficiently detect the top n outliers of a large and high-dimensional data set are proposed. Zhi-Wei *et. al.* [14] described a paper which mainly discusses how to maintain case bases in CBR system by adopting outlier data mining and case sieving techniques.

Malik *et. al.* [15] developed the algorithm for Detecting Web Content Outliers from Web Documents. Outlier mining is dedicated to finding data objects which differ significantly from the rest of the data. Outlier mining has been extensively studied in statistics and recently data mining. However, exploring the web for outliers has received very little attention in the mining community. Web content outliers are documents with 'varying contents' compared to similar web documents taken from the same domain. Mining web content outliers may lead to the identification of competitors and emerging business patterns in electronic commerce. They proposed WCONDMine algorithm for mining web content outliers using N-grams without a domain dictionary. Experimental results with embedded motifs show that WCOND-Mine is capable of finding web content outliers from web datasets.

Jianghui *et. al.* [16] seeking the unknown celestial body is one of the profound goals for the mankind explores pursued universe. Outlier mining is a kind of effective way of finding the spectrum data of unknown celestial body. Using outlier mining as the way of analyzing star spectrum data and VC++, Oracle9i as development tools, the outlier mining system on star spectrum data is designed and realized, its software architecture and function modules are given. At the same time, the paper elaborates the following key techniques: the preprocessing to star spectrum data, clustering techniques, outlier mining and visual simulation. The running results of the system show that it is feasible and valuable to apply this method to mining the outlier in spectrum data.

Nannan *et. al.* [19] gave an outlier mining-based method for anomaly detection. In this paper, a new technology is proposed to solve anomaly detection problems of the high false positive rate or hard to build the model of normal behavior, etc. What this technology based on is the similarity between outliers and intrusions. So a new outlier mining algorithm is proposed based on index tree to detect intrusions. The algorithm improves on the HilOut algorithm to avoid the complex generation of hilbert value. It calculates the upper and lower bound of the weight of each record with r-region and index tree to avoid unnecessary distance calculation. The algorithm is easy to implement, and more suitable to detect intrusions in the audit data. Many experiments have been performed on the KDDCup99 dataset to validate the effect of TreeOut and obtain good results.

Elio [20] developed the Framework for Outlier Mining in RFID data. Radio Frequency Identification (RFID) applications are emerging as key components in object tracking and supply chain management systems. In next future almost every major retailer will use RFID systems to track the shipment of products from suppliers to warehouses. Due to RFID readings features this will result in a huge amount of information generated by such systems when costs will be at a level such that each individual item could be tagged thus leaving a trail of data as it moves through different locations. They define a technique for efficiently detecting anomalous data in order to prevent problems related to inefficient shipment or fraudulent actions. Since items usually move together in large groups through distribution centers and only in stores do they

move in smaller groups we exploit such a feature in order to design our technique. The preliminary experiments show the effectiveness of our approach.

Bo Yu *et. al.* [22] described that distance-based outlier detection is an important data mining technique that finds abnormal data objects according to some distance function. However, when this technique is applied to datasets whose density distribution is different, usually the detection efficiency and result are not perfect. With analysis of features of outliers in datasets, as the improvement of Local Sparsity Coefficient-Based (LSC) Mining of Outliers, we rank each point on the basis of its distance to its kth nearest neighbor and the distribution of its k nearest neighborhood. A novel outlier detecting algorithm based Local Isolation Coefficient (LIC) is presented in this paper, which is shown better outlier mining results through the experiments.

Donghui-Shi, [24] gave the application of Outlier Mining in Power Load Forecasting. According to the theory of power load forecasting, data mining based on historical data of power load is used in load predicting. During the practical application, there are some errors in the data collection, and a load forecasting curve often contains big jagged edges. This paper presents a new outlier data mining approach. It finds sharp angle points between two lines, which correspond to outliers of power load. They smooth the curve at same time outliers are handled. Experiments show that after the new outlier mining approach was applied, load forecast results were improved significantly.

III. ANALYSIS OF COMPARATIVE STUDIES ON MINING DIVERSE PATTERNS

Initial researches discusses the concepts of web mining [1] and the classification [2] of web mining as Web access log mining and Web structure mining. But the problem of scarcity of data was ignored by initial researches, that was overcome by a technique backward induction developed by Dongkwon Joo and Songchun Moon [3] and to support it, the scalable web mining architecture in data warehouse environment is proposed by him.

Till this time, the outlier detection on high dimension data was not discussed. The outlier detection problem has important applications in the field of fraud detection, network robustness analysis, and intrusion detection. Most such applications are high dimensional domains in which the data can contain hundreds of dimensions. Study on this topic was suggested by Charu and Philip [4] and Fabrizio and Pizzuti [13]. Fabrizio Angiulli [13] proposes a distance bases outlier and an algorithm to efficiently detect the top *n* outliers of a large and high-dimensional data set are proposed.

Then the concept of Local Outlier Detection was introduced by Anny Lai-mei Chiu, [8] in the year 2003.She discussed the enhancements on Local Outlier Detection focuses on the density-based notion that discovers local outliers by means of the Local Outlier Factor (LOF) formulation.

Then the big achievement on finding the web content outliers was done by Malik Agyemang et.al. [9]. This paper shows the basic information of web mining and discusses the taxonomy as Web Structure Mining, Web Usage Mining and Web Content Mining along with practical applications and proposes a framework for mining web content outliers. The disadvantage of full word match, results in more outlier pages, was overcome by him in [10,15]. He took advantage of the HTML structure of web and N-gram technique for partial matching of strings and proposes an N-gram based algorithm for mining web content outliers. The study shows the effectiveness of applying N-grams on web content data, by minimizing the number of outlier pages, in contrast, without applying n-gram technique on web content data.

Zhi-Wei *et. al.* [14] that in case-based reasoning (CBR) system, as the scale of case base is enlarging, the system performance is gradually dropping pays attention to the case-based reasoning (CBR) system, how to maintain case bases in CBR system by adopting outlier data mining and case sieving techniques. Jianghui *et. al.* [16] demonstrates a method to mining the outlier in spectrum data.

Outlier mining-based method are used for anomaly detection [19], Radio Frequency Identification (RFID) applications [20], in Power Load Forecasting applications [24].

IV. CONCLUSION

In this paper, 24 research papers are reviewed for survey on detecting diverse patterns from web data. Various approaches for outlier detection are discussed in this paper along with applications. New trends are developed for mining diverse patterns. The paper shows the comparison between N-gram technique and without N-gram technique *i.e.* full word matching for finding web content outliers as explained in [9] and [10]. Results of these techniques, after implementing shows less weights to the web pages when it is used only full word matching technique resulting in more number of outlier pages. If algorithm uses N-gram technique, it assigns more weights to the web pages resulting in less number of outlier pages. So, accuracy of outlier detection system is improved by using N-gram technique.

REFERENCES

[1] Wang Jicheng, Huang Yuan, Wu Gangshan and Zhang Fuyan, "Web Mining: Knowledge Discovery on the Web", *IEEE international conference,* (1999).

[2] Masaru Kitsuregawa, Takahiko Shintani, Iko Pramudiono, "Web Mining and its SQL based Parallel Execution", *IEEE international conference,* (2001).

[3] Dongkwon Joo and Songchun Moon, "Scalable Web Mining Architecture for Backward Induction in Data Warehouse Environment", *IEEE Catalogue* No. 01 CH37239, (2001).

[4] Charu C. Aggarwal, Philip S. Yu, "Outlier Detection for High Dimensional Data", International conference, ACM SIGMOD 2001 May 21-24, Santa Barbara, California USA.

[5] Lizhen Liu, Junjie Chen, Hantao Song, "The Research of Web Mining", *Proceedings of the 4th World Congress on Intelligent Control and Automation*, Shanghai, P.R. China, June 10-14, (2002).

[6] Sankar K. Pal, Varun Talwar and Pabitra Mitra, "Web Mining in Soft Computing Framework: Relevance, State of the Art and Future Directions", *IEEE Transactions On Neural Networks*, Vol. **13**, No. 5, September, (2002).

[7] Hiroyuki Kawano, "Web Archiving Strategies by using Web Mining Techniques", *IEEE international conference*, (2003).

[8] Anny Lai-mei Chiu, Ada Wai-chee Fu, "Enhancements on Local Outlier Detection", *Proceedings of the Seventh International Database Engineering and Applications Symposium* (IDEAS'03), IEEE, (2003).

[9] Malik Agyemang, Ken Barker, Reda Alhajj, "Framework for Mining Web Content Outliers", *ACM Symposium on Applied Computing,* (2004).

[10] Malik Agyemang, Ken Barker, Reda Alhajj "Mining Web Content Outliers using Structure Oriented Weighting Techniques and N-Grams", *ACM Symposium on Applied Computing,* (2005).

[11] Yue Xu, "Hybrid Clustering with Application to Web Mining", *IEEE international conference,* (2005).

[12] Steffen Bickel, Peter Haider, and Tobias Scheffer, "Predicting Sentences using N-Gram Language Models", *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language, Vancouver,* October, (2005).

[13] Fabrizio Angiulli and Clara Pizzuti, "Outlier Mining in Large High-Dimensional Data Sets", *IEEE Transactions on knowledge and data engineering,* vol. **17**, no. 2, February, (2005).

[14] Zhi-Wei Ni, Yu Liu, Feng-Gang Li, Shan-Lin Yang, "Case Base Maintenance Based On Outlier Data Mining", *Proceedings of the Fourth International Conference on Machine Learning and Cybernetics*, Guangzhou, 18-21 August, (2005).

[15] Malik Agyemang, Ken Barker, Rada S. Alhajj, "WCOND-Mine: Algorithm for Detecting Web Content Outliers from Web Documents", *Proceedings of the 10th IEEE Symposium on Computers and Communications,* (2005).

[16] Jianghui Cai, Jifu Zhang, Xujun Zhao, "Design and Implement of Star Spectrum Outliers Mining System", *Proceedings of the 6th World Congress on Intelligent Control and Automation,* Dalian, China June 21 - 23, (2006).

[17] Juliana Lucas de Rezendel, Vinicios Batista Pereira, Geraldo Xexeo, Jano Moreira de Souza, "Building a Personal Knowledge Recommendation System using Agents, Learning Ontologies and Web Mining", *Proceedings of the 10th International Conference on Computer Supported Cooperative Work in Design,* (2006).

[18] Jiang Yiyong , Zhang Jifu, Cai Jianghui, Zhang Sulan, Hu Lihua, "The Outliers Mining Algorithm Based On Constrained Concept Lattice", *First International Symposium on Data, Privacy and E-Commerce,* (2007).

[19] Nannan Wu, Liang Shi, Qingshan Jiang, and Fangfei Weng, "An Outlier Mining-Based Method for Anomaly Detection", *IEEE international conference,* (2007).

[20] Elio Masciari, "A Framework for Outlier Mining in RFID data", *11th International Database Engineering and Applications Symposium,* (2007).

[21] Vue Zhang, lie Liu, "An Outlier Mining Algorithm Based on Probability", *2nd International Conference on Power Electronics and Intelligent Transportation System,* (2009).

[22] Bo Yu, Mingqiu Song, Leilei Wang "Local Isolation Coefficient-Based Outlier Mining Algorithm", *International Conference on Information Technology and Computer Science,* (2009).

[23] Cheng Ping-guang, "Research on Outlier Data Mining Algorithms Based on Subspace", *3rd International Conference on Advanced Computer Theory and Engineering,* (2010).

[24] Donghui-Shi, "Application of Outlier Mining in Power Load Forecasting", *International Conference on Computer Application and System Modeling,* (2010).

[25] Arun K. Pujari "Data Mining Techniques", Universities Press (India), Eleventh impression 2007.