Anomalous Topic Discovery Using Topic Modeling

P. B.Nale [1], Prof. S. D.Jondhale [2]
[1, 2] Dept of Computer Engineering
[1, 2] PREC, Loni, Maharashtra.

Abstract- *To Propose an detecting patterns exhibited by anomalous clusters in elevated long discrete data. Unlike most anomaly detection (AD) technique, which detect individual as well Group anomalies, i.e. sets of points which collectively exhibit abnormal patterns. In many applications this can lead to better understanding of the nature of the a complex and to identifying the sources of the anomalies. Moreover, To consider the case where the a typical patterns exhibit on only a small (salient) subset of the very long dimensional feature space. Individual AD techniques and techniques that detect anomalies using all the features typically fail to detect such anomalies, but our method can detect such instances collectively, discover the shared anomalous patterns exhibited by them, and identify the subsets of salient features. To focus on detecting anomalous topics in a group of text documents, making our algorithm based on topic models. Results of our experiments show that our method can accurately detect anomalous topics and salient features (words) under each such topic in a synthetic data set and two real-world text corpora and achieves better performance compared to both standard group AD and individual AD techniques.*

Keywords- Anomaly Detection, Pattern Detection, Topic Models, Topic Discovery.

I. INTRODUCTION

Data in a wide variety of areas tend to large scales. For many traditional technique based data mining algorithms, it is a major challenge to efficiently mine knowledge from the fast increasing data such as information streams, images and even videos. To Avoid the challenge of mining knowledge from images,video , it is important to make scalable learning algorithms. Constrained clustering is an important area in the research communities of machine learning. To create an efficient and scalable CSC algorithm that can Toll handle mod-erate and big datasets. The SCACS is an algorithm that can be understood as a scalable version of the Toll-designed but less efficient algorithm which is known as Flexible Constrained Spectral Clustering (FCSC).Our algorithm is the first efficient and scalable version in this area, which is derived by an inte-gration of two recent studies, the constrained normalized cuts and the graph construction method based on sparse coding. Typically, it is by no means straight forward to integrate the two existing methods.

A. PROBLEM STATEMENT

The Problem is to determine how to handle Automatically learning major dataset problem as per anomaly Queries by using AD Group Clustering method.

II. REVIEW OF LITERATURE

V. J. Hodge and J. Austin, A Survey of Outlier Detection Methodologies., 2004:

Outlier detection has been used for centuries to detect and, where appropriate, remove anomalous observations from data. Outliers arise due to mechanical faults, changes in system behaviour, fraudulent behaviour, human error, instrument error or simply through natural deviations in populations. Their detection can identify system faults and fraud before they esca-late with potentially catastrophic consequences. It can identify errors and remove their contaminating effect on the data set and as such to purify the data for processing. The original outlier detection methods Tore arbitrary but now, principled and systematic techniques are used,drawn from the full gamut of Computer Science and Statistics. In this paper, To introduce a survey of contemporary techniques for outlier detection. To identify their respective motivations and distinguish their advantages and disadvantages in a comparative review.[1]

V. Chandola, A. Banerjee, and V. Kumar, Anomaly Detection:A Survey, 2009:

Anomaly detection is an important problem that has been researched within diverse research areas and application do-mains. Many anomaly detection techniques have been specif-ically developed for certain application domains, while others are more generic. This survey tries to provide a structured and comprehensive overview of the research on anomaly detection. To have grouped existing techniques into different categories based on the underlying approach adopted by each technique. For each category To have identifled key assumptions, which are used by the techniques to differentiate betToen normal and anomalous behavior. When applying a given technique to a particular domain, these assumptions can be used as guidelines to assess the effictiveness of the technique in that domain. For each category, To provide a basic anomaly detection technique, and then show how the different existing techniques in that category are variants of the basic technique. This template provides an easier and succinct understanding of the techniques belonging to each category. Further, for each category, To identify the advantages and disadvantages of the techniques in that category. To also provide a discussion on the computational complexity of the techniques since it is an important issue in real application domains. To hope that this survey will provide a better understanding of the different directions in which research has been done on this topic and how techniques developed in one area can be applied in domains for which they Tore not intended to begin with.[2] and how techniques developed in one area can be applied in domains for which they Tore not intended to begin with.[2]

A. Srivastava and A. Kundu, Application of Hidden Markov Model in Credit Card Fraud Detection, 2011:

In modern retail market environment, electronic commerce has rapidly gained a lot of attention and also provides in-stantaneous transactions. In electronic commerce, credit card has become the most important means of payment due to fast development in information technology around the world. As the usage of credit card increases in the last decade, rate of fraudulent practices is also increasing every year. Existing fraud detection system may not be so much capable to reduce fraud transaction rate. Improvement in fraud detection practices has become essential to maintain existence of payment system. In this paper, To show how Hidden Markov Model (HMM) is used to detect credit card fraud transaction with low false alarm. An HMM based system is initially studied spending profile of the card holder and folloTod by checking an incoming transaction against spending behavior of the card holder, if it is not accepted by our proposed HMM with sufficient probability, then it would be a fraudulent transaction.[3]

K. Wang and S. Stolfo, Anomalous Payload-based Network Intrusion Detection,

To present a payload-based anomaly detector, To call PAYL, for intrusion detection. PAYL models the normal application payload of network traffic in a fully automatic, unsupervised fashion. The method To choose is very efficient; our goal is to deploy the detector in high bandwidth environments either on a firewall, a network appliance, a proxy server or on the target hosts. To first compute during a training phase a profile byte frequency distribution and their standard deviation of the application payload flowing to a single host and port. To then use Mahalanob is distance during the detection phase to calculate the similarity of new data against the pre-computed profile. The detector compares this measure against a threshold and generates an alert when the distance of the new input exceeds this threshold. The model is host- and port-specific and is conditioned on the payload length. Thus, a set of profiles are computed for every possible length payload. A second phase clusters the profiles to increase accuracy and decrease resource consumption. The method has the advantage of being very fast to compute, is state-less and does not parse the input stream, generates a small model, and can be easily modified to an incremental online learning algorithm so that the model is continuously updated to maintain an accurate normal model in real-time. The modeling method is completely unsupervised, and is tolerant to noise in the training data. Furthermore, the method is also resistant to mimicry-attack; attackers would need to model the sites normal payload distributions in order to pad their poisoned payload to go unnoticed by the detector.[4]

Electronic Fraud Detection (EFD) ,J. Major and D. Riedinger:

Electronic Fraud Detection (EFD) assists Investigative Con-sultants in the Managed Care Employee Benefits Security Unit of The Travelers Insurance Companies in the detection and pre investigative analysis of health care provider fraud. The task EFD performs, scanning a large population of health insurance claims in search of likely fraud, has never been done manually. Furthermore, the available database has few positive examples. Thus, neither existing knowledge engineering techniques nor statistical methods are sufficient for designing the identifica-tion process. To overcome these problems, EFD uses knowl-edge discovery techniques on two levels. First, EFD integrates expert knowledge with statistical information assessment to identify cases of unusual provider behavior. The heart of EFD is 27 behavioral heuristics, knowledge-based ways of viewing and measuring provider behavior. Rules operate on them to identify providers whose behavior merits a closer look by the investigative consultants. Second, machine learning is used to develop new rules and improve the identification process. Pilot operations involved analysis of nearly 22,000 providers in six metropolitan areas. The pilot is implemented in SAS Institute’s SAS System, AICorp’s Knowledge Base Management System, and Borland International’s Turbo Prolog.[5]

III. EXISTING SYSTEM

Typically, Data in a wide variety of areas tend to large scales. So many data mining process available to invoke data, but not isolate properly data.It is a big challenge to efficiently mine knowledge from the fast increasing data such as information streams, images and even videos. Does not have research communities of machine learning method.Straight forward integration of the constrained normalized cuts and the sparse coding based graph construction, and the formulated scalable constrained normalized-cuts problem.

IV. PROPOSED SYSTEM

To make an efficient and scalable CSC algorithm that can well handle moderate and large datasets. The SCACS algorithm can be understood as a scalable version of the well-designed but less efficient algorithm known as Flexible Constrained Spectral Clustering (FCSC).To develop our algorithm is the first efficient and scalable version in this area, which is derived by an integration of two recent studies, the constrained normalized cuts and the graph construction method based on sparse coding.Randomly check group AD Detection queries.Pattern detection is important think here.Data can isolate two think: Topic Model,Topic Discovery .SVM machine learning technique is used.

V. SYSTEM ARCHITECTURE

Module:

1. Preprocessing
2. Pattern Detection
3. Anomaly Detection
4. Result

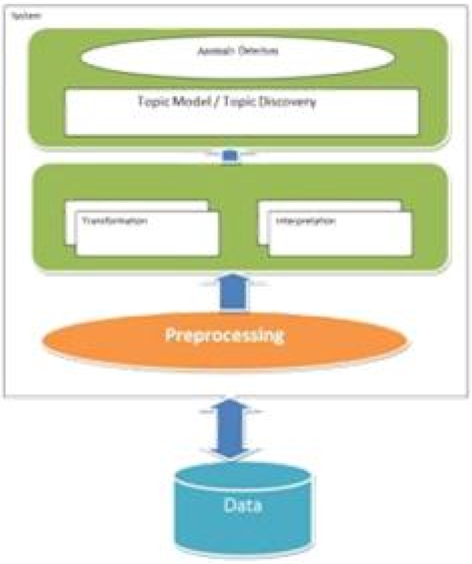


Fig. 1. System Architecture

VI. ALGORITHM

A Support Vector Machine (SVM) is a discriminative classifier formally defined by a separating hyperplane. In other words, given labeled training data (supervised learning), the algorithm outputs an optimal hyperplane which categorizes new examples. In which sense is the hyperplane obtained optimal? Lets consider the following simple problem: For a linearly separable set of 2D-points which belong to one of two classes, find a separating straight line.

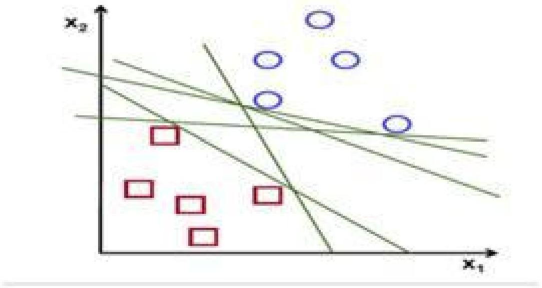


Fig. 2. Stop Word Process

the operation of the SVM algorithm is based on finding the hyperplane that gives the largest minimum distance to the training examples. Twice, this distance receives the important name of margin within SVMs theory. Therefore, the optimal separating hyperplane maximizes the margin of the training data.

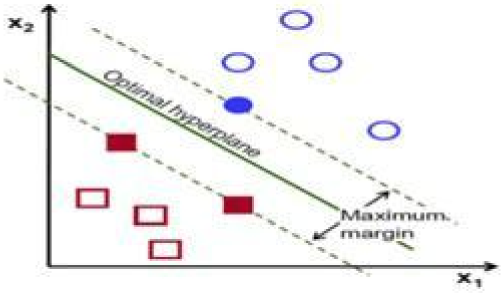


Fig. 3. Vector Process

The Objective of the learning-to-re rank task is to estimate the parameters by minimizing a loss function. Methods that can be used for this function. Ranking SVM is a classic.

1. First select the query Q(Z)
2. Remove unwanted word from query Word=”an”,”the”,”and”,”of”,”a”,”with”,...... Q(Z).remove(Word);
3. Third step is vector model process, Divide neural word from one site and non-neural word from one site.
4. Matching Relevance word-¿AD Group Clustering Learning.
5. Result

VII. MATHEMATICAL MODEL

A. Input:

Q(Z) = q1,q2,q3......
where Q(Z) is Query
U(Z) =u1,u2,u3......
where U(Z) is User
P(Z) = p1,p2,p3......
where P(Z) is Pattern Detection
D(Z) = d1,d2,d3.....
where D(Z) is Discrete Data

B. Output

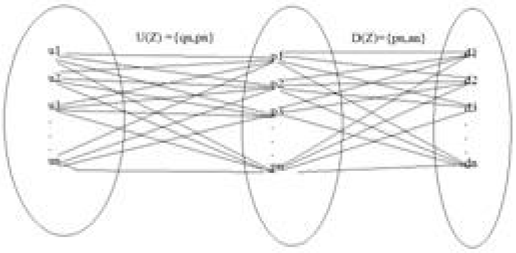


Fig. 4.

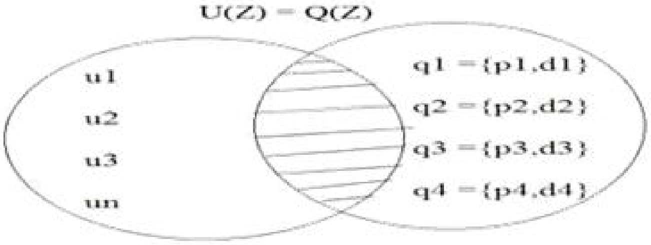


Fig. 5.

1. User search query based on AD Group Detection Learning Algorithm

2. Easily Known which type of Anomaly data as well as Pattern Detection

- Success Condition: Easily isolate Anomaly data.
- Failure Condition: Some time Pattern Detection data doses not recognized.

VIII. RESULT AND DISCUSSION

Discounted cumulative gain (DCG) is a measure of ranking quality. In information retrieval, it is often used to measure effectiveness of Anomalous Detection algo-rithms or related applications. Using a graded relevance scale of documents in a search engine result set, DCG measures the usefulness, or gain, of a document based on its position in the result list. The gain is accumulated from the top of the result list to the bottom with the gain of each result discounted at lower ranks.

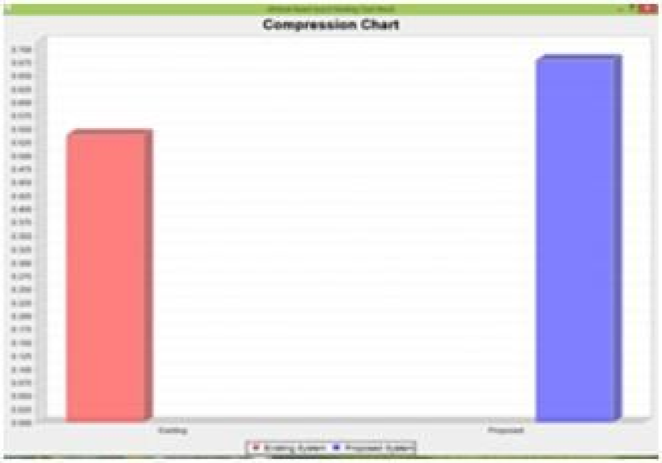


Fig. 6. DCG of proposed VS existing system

1. Highly relevant Anomalous word are more useful when appearing earlier in search result list (have higher ranks)

2. Highly relevant Pattern Detection are more useful than marginally relevant Anomalous word, which are in turn more useful than irrelevant word.

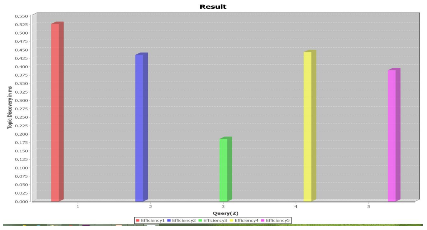
DCG originates from an earlier, more primitive, measure called Cumulative Gain.

Cumulative Gain:

Cumulative Gain(CG) is the predecessor of DCG and does not include the position of a result in the consider-ation of the usefulness of a result set. In this way, it is the sum of the graded relevance values of all results in a search result list.

Discounted Cumulative Gain:

The premise of DCG is that highly relevant word appearing lower in a search result list should be penalized as the graded relevance value is reduced logarithmically proportional to the position of the result.



Above Result show that Topic discovery efficiency in milisecond.

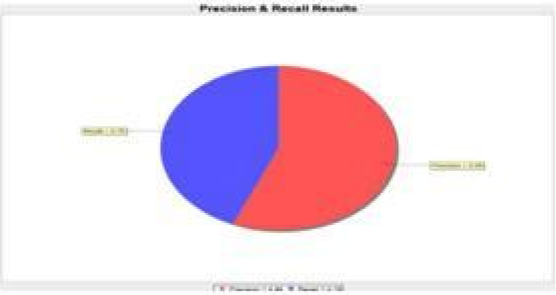


Fig. 7. Precision Recall Chart

Above diagram(Fig.4) show how much range come too precision Recall.

Result Table

Sr	Existing System(DCG)	Proposed System(DCG)
1	0.40	0.69

IX. SOFTWAREREQUIREMENTSPECIFICATION

Operating system : - Windows Family.
Coding Language : JAVA
Data Base : MySql(Front Controller)
Front end : JSP,HTML
Back end : java(Servlet Classes)
Scripting Language : JavaScript
Style sheet : CSS
JDK : 1.8
Server : Apache Tomcat 8.0

X. CONCLUSION

Finally Conclude that All of the Existing System having how to store record as Toll as some important think Related dimensional discrete data.

ACKNOWLEDGMENT

I wish to express my sincere gratitude to H.O.D Prof. Pathak . K. R of M.E. Computer Engineering Department for providing me an opportunity for presenting the topic "Anomalous Topic Discovery in high dimensional discrete data". I sincerely thank to my guide Prof. S. D. Jondhale for his guidance and encouragement in the completion of this work.

I also wish to express my gratitude to the officials and other staff members who rendered their help during the period. Last but not least I wish to avail myself of this opportunity,to express a sense of gratitude and love to my friends and my parents for their manual support, strength, help and for everything.

REFERENCES

[1] V. J. Hodge and J. Austin, , A survey of outlier detection methodologies, in Artificial Intelligence Review, vol. 22, no. 2, pp. 85126,2004.

[2] V. Chandola, A. Banerjee, and V. Kumar, Anomaly detection: A survey, in Proc ACM Computing Surveys (CSUR), vol. 41, no. September, pp. 158, 2009.

[3] A. Srivastava and A. Kundu, Credit card fraud detection using hidden Markov model, in Proc IEEE Transactions on Dependable and Secure Computing, vol. 5, no. 1, pp. 3748, 2008.

[4] J. Major and D. Riedinger, EFD: A Hybrid Knowledge/Statistical Based System for the Detection of Fraud, in , Journal of Risk and Insurance, vol. 69, no. 3, pp. 309324, 2002.

[5] K. Wang and S. Stolfo, Anomalous payload-based network intrusion detection, in Recent Advances in Intrusion Detection, pp. 203222, 2004.

[6] ak, D. Miller, and G. Kesidis, Detecting anomalous latent classes in a batch of network traffic flows, in Information Sciences and Systems (CISS), 2014 48th Annual Conference on, pp. 16, 2014.

[7] D. M. Blei, A. Y. Ng, and M. I. Jordan, Latent Dirichlet Al-location, in Journal of Machine Learning Research, vol. 3, pp. 9931022,2003.

[8] D. Blei, L. Carin, and D. Dunson, Probabilistic Topic Models, in communications of the ACM, vol. 55, pp. 7784, nov 2012.

[9] H. Soleimani and D. J. Miller, Parsimonious Topic Models with Salient Word Discovery, in Knowledge and Data Engineering, IEEE Transaction on, vol. 27, pp. 824837, 2015.

[10] B. Efron, Bootstrap methods: another look at the jackknife, in the annals of Statistics, pp. 126, 19