International Journal for Research in Engineering Application & Management (IJREAM)
ISSN : 2454-9150 Vol-03, Issue-08, Nov 2017

# Group Anomaly and Emerging Topics Discovery Using ATD Approach

[1]Shewale Yogita, [2]Prof. N. R. Wankhade

[1]PG Student, [2]Professor, [1,2]Department of Computer Engineering, Late G. N. Sapkal College of Engineering Nashik, Maharashtra, India.

***Abstract*** **- Anomaly detection is an important problem that has been researched within diverse research areas and application domains. Many anomaly detection techniques have been specifically developed for certain application. Domains Anomaly detection is an approach to detect anomalies from high dimensional discrete data. Several approaches for anomaly detection have been proposed which is only capable of detecting individual anomaly. It is very time consuming and infeasible task. With proposed approach group anomalies are detected. Some techniques used all features for anomaly detection which get fail. In our system, batch of text documents are given to discover anomalies therefore, topic based algorithmic approach is utilized. With group anomalies detection, emerging topic discovery by extracting links between social users is contributed in proposed system.**

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

## I. INTRODUCTION

Anomaly detection refers to the problem of finding patterns in data that do not conform to expected behavior. These non-conforming patterns are often referred to as anomalies, outliers, discordant observations, exceptions, aberrations, surprises, peculiarities or contaminants in different application domains..AD has several applications in credit card fraud detection, insurance fraud detection, network intrusions. Traditional approaches are only capable of detecting an individual anomaly from given input. Therefore, Anomalous Topic Discovery (ATD) approach is proposed. It contains two phases such as training and testing phase. In training phase, Parsimonious Topic Model (PTM) is used. Rather than LDA model, PTM is utilized in proposed work because, it can achieve better generalization accuracy and automatically estimates the number of normal topics from test batch. In PTM, normal data is extracted and used to construct null model whereas, in anomaly detection phase, null model is used as, reference model to identify group or clusters of anomalies from test batch of documents.

## II. RELATED WORK

Application based on HMM works on credit card fraud detection as well as works on an assumption. It creates training dataset by observing the user behavior. By analyzing training dataset it defines threshold. If upcoming test record has lower value than threshold then it generates an alert as fraud detected. In this fraud detection system user spending behavior is analyzed. Based on transaction history user profile is categorized in 3 categories as: low, medium and high. According to the profile and transaction history threshold is defined. Fraud is nothing but a anomalous entry is in transaction is identified in this system[1]. PTM is parsimonious topic model used for text corpora. In s Latent Dirichlet Allocation (LDA), all words are modeled topic-specifically, even though many words occur with similar frequencies across different topics. PTM model gives the sparse topic representation to determine subset of relevant topics for each document. BIC is derived for complexity and goodness of fit. BIC is minimized to determine entire model. Results are carried out on three text corpora and an image dataset to show that proposed model can achieve higher test set. The proposed form of BIC has differentiated cost terms, based on different effective sample sizes for the different parameter types in their model. The use of a shared feature representation essentially increases the sample size to feature dimension ratio[2]. A rule based anomaly detection algorithm identifies anomalous topic by some predefined rules. It mainly works on emergency dataset containing emergency cases of anmous pattern. It defines new algorithm WSARE based on the strategy: *"What's strange about recent events"*. Due to computational issues, the number of components for these rules is two or less[3]. Local anomaly detector identifies individual records with anomalous attribute values and then detect pattern where the number of anomalous record is higher than expected. It is able to accurately detect anomalous patterns in real world hospital container shipping and network intrusion data. The proposed APD is orthogonal to the local anomaly detection method[4]. The problems of new event detection and event tracking within a stream of broadcast news stories looking at subsequent stories, decision about one story is taken already. They represented a method based on miss and false alarm rates. In this bootstrap method is used to produce performance distribution for topic detection. The TDT tasks and evaluation were developed by a joint effort between DARPA. The group involved in the tasks found that the "State of the art" is capable of providing adequate performance for detection and tracking events, but there is high enough failure rate to warrant significant research into

 DOI : 10.18231/2454-9150.2017.0030 

how algorithms can be advanced. A TDT approach is presented as the vertical search engine in financial field. Results are grouped into multiple topics with stock as unit. The clustering method called as, hierarchical agglomerative. Online topic detection and topic tracking with the proposed approach splits, agglomerative hierarchical clustering method into two steps and considers the time factor. In final study the effect on the similarity between two stories or topics. The proposed method limited only for tracking and detection of financial news. A system for adaptive anomalous discovery based on adaptive AD algorithm. It works on high dimensional dataset. Ad algorithm uses score function and generates neighbor graph for n nominal values points in dataset. It identifies the points with smaller score value with respect to m topics generated by other points. The versions of SVM approximate for single-class classification in the context of information retrieval used as single-class SVM. It is robust with respect to small categories. But SVM is very sensitive to the parameter and selection on kernel. Multiple parameters are included in this proposed method involves the data representation and the decision involved in the modification on two-class method to individual class. However, it turns out to be surprisingly sensitive to specific choices of representation and kernel in ways which are not very transparent. To measure the performance of proposed method neural network method is required. The problem of individual anomaly detection is proposed to analyz that in most of the cases anomalies are occurred in the group form. Previous methods can able to identify group of anomaly which is already present in the dataset. In this paper author proposed a hierarchical Bayes model which is known as, GLAD i.e. "Group Latent Anomaly Detection". It can accept both pair-wise & point-wise data in the form of input. This proposed model can automatically infers the group & can simultaneously detect group anomalies. In processing step, MMSB and LDA model shared the group of membership distribution for given both input. The general approach for the iterative computations of maximum-likelihood estimates the observation views incomplete data. EM algorithm is remarkable process because of simplicity and the generality of associated theory. Concept similar to the EM algorithm is discussed in [11] by X. Li. Meng and D. V. Duk. Their main approach is to consider statistical construction of algorithms that are simple and fast. In this paper, they discussed about intrinsic connections between EM-type algorithm and the Gibb's sampler. It helps to detect individual wheat from the chaff from the thousands of incoming news stream. In this paper, DPM, Discriminative Probabilistic Model is proposed. It is simple and effective topic detection model. In this paper they focused on both online and offline topic discovery using DPM. DPM does not require any complicated generative models like vMF mixture and LDA. Clustering process is represented by variation of TFIDF under condition in that only discriminative words are used. A bursty phenomenon of words is utilized to discover discriminative features. Furthermore, they remark DPM soft-clustering performance on offline topic detection. As extend to this work author planned to explore non- Dirichlet process mixture models from topic evolution. A system for detecting group of anomalies is proposed to identify individual objects form large dataset. In some scenario group of anomalies may appear in sequential manner. It helps to identify sources or pattern of anomaly. It takes high dimensional discrete data as an input. The scope of proposed method is applied to multidimensional dataset containing only discrete features and not applied to regular text based document anomaly detection

M. Zhao et al[14], proposed non-parametric adaptive anomaly detection algorithm. The proposed algorithm derived from nearest neighbor graphs on nominal data point. Whenever, test samples falls down anomaly score get detected. The proposed algorithm is efficient and linear in dimension as well as quadratic into data size. K-nearest neighbor taken as an input to produced sample test score. With the computation of high dimensional quantities, it is reliably difficult in high dimensional feature spaces. Computing high dimensional quantities get avoided with the computation of score functions. Computational cost of proposed algorithm is grows linearly and quadratically in the dimension and in data size respectively.

S. Wilks, et al[15], applied the principle of maximum likelihood. A method is suggested for the functions of observation which is called as, "composite statistical hypothesis"/ "simple composite hypotheses". For test significance a number of statics are used to make significance test which expressed in terms of ʎ.

G. Schwarz[16], concentrated towards an appropriate modification of maximum likelihood. A leading terms of Bayes estimator turns into maximum likelihood estimator. Hence they have lower probability on lower-dimensional subspaces. A staticians problem of selecting an appropriate dimensionality model which fit to the given dataset.

## III. SYSTEM ARCHITECTURE

"To design and develop a system for group anomaly and emerging topic discovery using ATD technique."

Anomaly detection is an approach to detect anomalies from high dimensional discrete data. Several approaches for anomaly detection have been proposed which is only capable of detecting individual anomaly. It is very time consuming and infeasible task. Therefore, proposed system (ATD) aims to detect group of anomalies. Batch of text documents are given to discover anomalies and due to topic based approach proposed algorithm can efficiently discovers topics in text documents. System works on training corpus to detect normal topics based on PTM model technique. Pattern matching and group anomaly in cluster is then carried out into testing phase. In testing, similar documents based on similar patterns are club into clusters hence unusual or anomalous topic remains into side. In this process, topic relevance score is calculated. In each step of proposed algorithm candidate anomalous cluster (S) is detected which exhibits maximum

 DOI : 10.18231/2454-9150.2017.0030 

"deviance" from normal topic. Cluster significance is calculated to get d*. d* is candidate document belongs to S. Bootstrapping algorithm is utilized. New upcoming documents are matched with existing cluster and added to cluster having highest similarity match. If cluster size reached to limit then re-clustering is performed with specified threshold value.

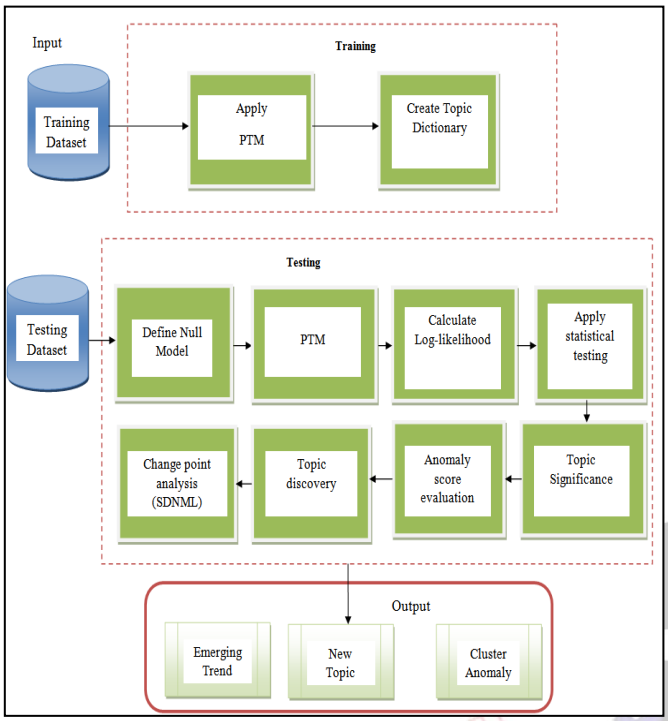


**Figure 1: System Architecture**

Figure 1 represents the architecture of proposed system. Our proposed ATD algorithm is to detect cluster of anomalies.

In experimental setup phase, performance of algorithm compared with baseline methods on synthetic data set and two text corpora dataset. To define anomalous classes a ground-truth class labels are used. In each data set, they have chosen some classes as anomalous and take all documents from those classes out of the training and validation sets. Then they normally select some documents from anomalous classes to build a test set. In this paper author do not considered any anomalous cluster actually exist in the test dataset. Document and upload time are mapped. Uploaded documents, those are anomalous but may relate to each other. System generates cluster for such topic as new trend appear with respect to time.

### A. Training Phase

*1. Upload training dataset*
*2. Apply stopword algorithm*
*3. Apply Stemming*
*4. Apply PTM*
*5. Save PTM parameters*
*6. Define Null model M0*

### B. Testing Phase

*1. Define testing dataset*
*2. Apply stopword algorithm*
*3. Apply Stemming*
*4. Define candidate anomalous cluster*
*5. Define M0*
*6. Define M1*
*7. Define word dictionary*
*8. Calculate BIC*
*9. Calculate degree of deviation*
*10. Calculate anomaly score*
*11. Add topic in cluster*
*12. Calculate word probabilities*
*13. Apply SDNML*
*14. Burst Detection*
*15. Calculate Anomaly score*
*16. View analysis report i.e. cluster anomaly and emerging trend*

## IV. ALGORITHMS

**Input:**

- Test dataset D and PTM model with 'M'

**Where, D**: set of documents indexed by d ε {1, 2 …D} and **'M':** Normal topics

- Link between social user's

**Output:**

- Discovered cluster S with significance measure p-value(S).
- Current trend

**Processing steps:**

**1. READ** Test dataset $D_t$ and $M_0 = \{\theta_0, H_0\}$ on Dt Where, $M_0$ is null model,
**Where**,
'$\theta_0$': Topic-specific word probabilities
'H0': Topic proportions

**2. COMPUTE** $l_0$ (d) ∀d ε Dt
**Where,** $l_0$ (d) : length of document
**3. REPEAT**
**SET** S=Ø **Where,**
**S :** Set of normalized topics
**4. SELECT** d*= $\text{argmin}_{d \in Dt}$ $1/L_d$ $l_0(d)$
**5. REPEAT**
**6. SET** S←S U {d*}
**Where,** d* : selected documents from 'S'
**7. READ** $M1=\{\theta_1,H_1\}$ on S
**Where, M1** : alternative model
**8. COMPUTE** $l_1(d)$ ∀dε Dt-S

**9. SELECT** d*=

**10. CALL** algorithm (3), to test significance of topic M+1 in document d*
**11. UNTIL** topic M+1 is insignificant in d*
**12. EVALUATE** score(S)
**13. CALL** algorithm (4), to test significance of 'S'
14. $D_t$←$D_t$ - S
**15. TILL** 'S' is significant
**16. APPLY Kleinberg's burst detection method & Sequentially Discounting Normalized Maximum Likelihood (SDNML)coding**

$$P^b sw(1\text{-}Psw)^{n-b} \Pi^n_{t=1} f_{exp}(x_t ; \alpha i_t)$$

 DOI : 10.18231/2454-9150.2017.0030 

**Where**,

- *p*sw is a given state transition probability,
- *b* is the number of state transitions in the sequence,
- *it* (*t* =1, . . . , *n*),
- $f_{exp}(x ; \alpha)$ is the probability density function of the exponential distribution with rate parameter α,
- *xt* is the $t^{th}$ inter-event interval.

**18. GET emerging trend**

***1. Algorithm to Generate Bootstrap Document***

**Input**:

d*: candidate document

$D_v$ : No. of document in validation set

**Processing**:

Step 1: Compute similarity between document d and d* using Cosine similarity measure

Step 2: Find document sparcity d'.

d' =argmaxd ρd∗ (d) ∀d = 1,... Dv

Step 3: Randomly choose one of the documents from D', d' ~ uniform (D').

Step 4: Then, from the Ld' words in document d',

randomly choose Ld∗ words with replacement.

**Where,** Ld: length of document

**Output:** Document $d_b$={$w_{1b,.....}w_{Ld*b}$}

***2. Algorithm for testing significance of topic M+1 in document d****

**Input:**

d* : candidate document

Dv: No. of document in validation set

$M_0$: Null Model

$M_1$: Alternative Model

**Processing:**

Step 1: **Evaluate** actual scope of new topic θˆ d∗

For b=1 to B1 do

Step 2: **Generate** Bootstrap document b from algorithm 2

Step 3: **Identify** the scope of new topic under $M_1$

Step 4: **Compute** θˆ $_b$

**End for**

**Output**: t(θˆ d∗) i.e. Significance of the new topic in candidate document d ∗

**Where,** θˆ : Significance of new topic

***Algorithm 3: Testing Significance of 'S'***

**Input:**

Candidate cluster 'S'

Score ($S_b$)

**Processing:**

Step 1: For b=1 to $B_2$

Step 2: Set $S_b$= ∅

Step 3: for d=1 to |S| do

Step 4: Generate bootstrap documents for d using algorithm2

Step 5: $S_b$← $S_b$ U {$d_b$}

Step 6: Compute score ($S_b$)

Step 7: End for

Step 8: Identify $M_0$ & $M_0$ on $S_b$

Step 9: Compute Score ($S_b$)

Step 10: end for

**Output:**

p-value to measure significance of the candidate cluster

Our algorithm consists of two steps. First is the training step in which we learn PTM as our null model M 0 to generate all document in test set. Second is the detection phase in which we utilized document bootstrapping algorithm for clustering of candidate documents (S) in the test set. Furthermore, as a part of contribution we focus on emergence of topics Use APPLY Kleinberg's burst detection method & Sequentially Discounting Normalized Maximum Likelihood (SDNML) coding for Change Point Analysis. Signaled by social aspects by discovering links between social users. Aggregating anomaly scores from hundreds of users, In Proposed show that we can detect emerging topics only based on the reply/mention relationships in documents. proposed approach can efficiently detects a cluster of anomalies and emerging topic in test set.

## V. MATHEMATICAL MODEL

S is the system of anamoly detection such that,

S = {I, F, O} I is the input to the system F is system functions O is Systems output

***Table 1: Math model***

| | |
|---|---|
| I: {I1, I2, I3},Set of input data | I1= Training Dataset |
| | I2= Testing Dataset |
| | I3= Node |
| F: {F1, F2, F3, F4, F5, F6, F7, F8, F9, F10, F11, F12, F13,F14,F15,F16,F17} | F1= Upload training dataset |
| | F2=Apply preprocessing i.e. stemmer and stopword algorithm |
| | F3=Word Extraction |
| | F4= Apply PTM |
| | F5= Save PTM Parameters |
| | F6= Define null model |
| | F7= Upload testing datset |
| | F8= Apply preprocessing |
| | F9= Define candidates anomalous cluster |
| | F10= Compute M0 |
| | F11= Define M1 |
| | F12= Define word dictionary |
| | F13= Calculate BIC |
| | F14= Add topic in cluster |
| | F15= Calculate word probabilities |
| | F16= Calculate anomaly score |
| | F17= Display report |
| O :{ O1, O2, O3,O4} | O1 = Preprocessed word |
| | O2= Word probabilities |
| | O3= Anomaly score |
| | O4 =Cluster of anomalous topic |

In Above Mathematical model the input, output, and the execution process is given

There are two phases involved in proposed approach:

- 1. Training phase.
- 2. Testing phase.

Typically, AD techniques used to detect individual anomalies from the test set. To extend this, discovery of cluster of anomalies have been proposed in this paper. The proposed

 DOI : 10.18231/2454-9150.2017.0030 

ATD algorithm contains two phases, in first phase training is provided to the test set to identify PTM i.e. Parsimonious Topic Model which contains normal patterns as null model. Other is detection of anomaly cluster. For ATD algorithm PTM model is selected rather than LDA model. PTM model typically achieves better generalization accuracy than the LDA. It can automatically estimate the number of normal topics. In anomalous topic discovery, model order selection is crucial task. Therefore to compute significance of any anomaly topic will be measure with the null model, either over fitting the null can lead to false discovery of anomalous clusters due, respectively, to limited modeling power or to poor generalization.

To design and develop anomaly detection system, in which group anomalies are detected. Also to discover latest emerging trend by extracting communication links between users.

## VI. EXPERIMENTAL SETUP

We have developed a desktop application using java- Jdk1.7. Mysql 5.3 is used to store database. Core i3 machine with 4GB ram is used for development and testing. Netbeans-8.0.1 IDE is used to build and test the system using J unit.

***Dataset:***

1. Newsgroup dataset [17]: It contains 20 different news topics with news article. It contains approximately 20,000 newsgroup documents
   Some categories are very closely related to each other. Following is a list of the 20 newsgroups, partitioned according to subject matter

In this system, News Group dataset is given as input to system. It get preprocessed to extract news categories form it. List of news categories is generated and then partitioned by user into training and testing phase. Parsimonious Topic Model (PTM) outputs normal pattern from given dataset. In both phases, clusters are generated as per similar topics from list. Significance of normal topics i.e. `S' is calculated and matched with the normal clusters. From the set of all clusters, data point which exhibits the pattern with maximum \deviance" from normal topics. Then, a statistical test is performed to measure the significance of S and the topic exhibited by it, compared to the normal topics hypothesis

| | | |
|---|---|---|
| comp.graphics<br>comp.os.ms-windows.misc<br>comp.sys.ibm.pc.hardware<br>comp.sys.mac.hardware<br>comp.windows.x | rec.autos<br>rec.motorcycles<br>rec.sport.baseball<br>rec.sport.hockey | sci.crypt<br>sci.electronics<br>sci.med<br>sci.space |
| misc.forsale | talk.politics.misc<br>talk.politics.guns<br>talk.politics.mideast | talk.religion.misc<br>alt.atheism<br>soc.religion.christian |

**Figure 2: Image of dataset**

**Result Tables:**

*Table 2: Comparative analysis of anomaly detection*

| Total documents | Anomalous Documents(Existing System) | Anomalous Documents(Proposed System) |
|---|---|---|
| 50 | 14 | 10 |
| 100 | 25 | 15 |
| 150 | 33 | 16 |
| 200 | 45 | 36 |

Table 2, represents the performance analysis between existing and proposed system. For system testing we have used set of documents such as, 50,100,150,200. As per observation as compared to existing system, proposed system consists of fewer anomalies for each input document set due to utilization of clustering and new topic discovery techniques.

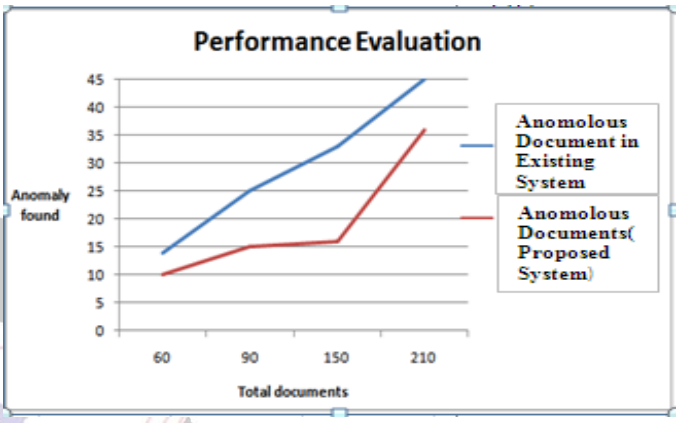


**Figure 2: Graph of system comparison**

In figure 3, graphical view of comparative analysis is shown. On X-axis, numbers of documents are given whereas; on Y-axis extracted anomalies are given.

*Table 3: Topic Discovery*

| Total documents | Topic discovery(Actual results) | Topic discovery(Expected results) |
|---|---|---|
| 60 | 4 | 6 |
| 90 | 10 | 13 |
| 150 | 7 | 11 |
| 210 | 9 | 5 |

Table 3 represents topic discovery analysis. Readings are manually evaluated by analyzing results of anomaly detection from table 9.2 given table 9.3 depicts the variance or comparison between actual and expected grouped anomaly results.

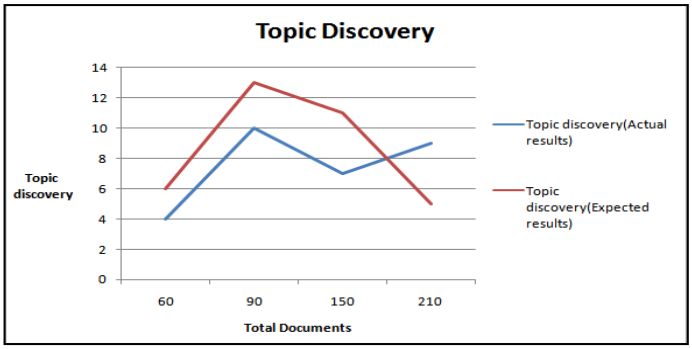


**Figure 4: Analysis of topic discovery**

 DOI : 10.18231/2454-9150.2017.0030 

In figure 4, depicts graph of topic discovery in which X-axis represents test documents and Y-axis represents topic discovery results.

***Table 4: Average significance of ATD***

| No. Of Documents | Avg. Significance | Avg. Significance |
|---|---|---|
| 50 | 0.141286828 | 0.063787304 |
| 100 | 0.151248 | 0.08652 |
| 150 | 0.14982 | 0.09236 |
| 200 | 0.2472 | 0.06932 |

Table 4 represents the average document significance between proposed and existing system. From observation of table it is clear that existing system consisting more significance than proposed system. Becuse in proposed system relevance between anomalous topics get calculated and it get grouped hence there exist less anomaly which depicts the less average significance.

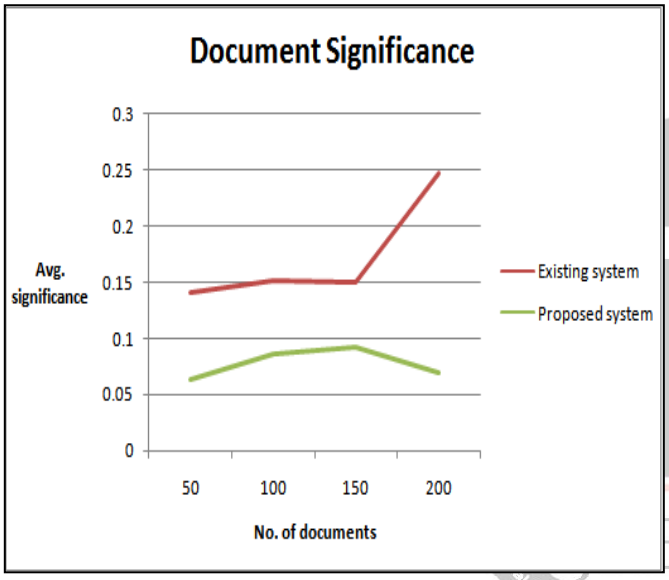


**Figure 5: Graph of average significance**

Figure 5 represents the graph of anomaly significance. In this X-axis represents the number of documents and Y-axis represents the average significance in anomalies.

*Table 5: 'p'-value measure*

| Total documents | Precision | recall | f-measure |
|---|---|---|---|
| 30 | 0.85 | 1 | 0.9 |
| 60 | 0.9 | 1 | 0.9 |
| 90 | 0.9 | 1 | 0.9 |
| 120 | 0.95 | 1 | 0.95 |
| 210 | 0.75 | 1 | 0.8 |

Table 4 represents the performance of `p'-measures. It consists of average precision, recall and f-measure values of three domains such as, Atheism, Hockey and Anomali(Graphics).

For testing purpose we have used set of documents such as, 30,60,90,120,210.

Figure 5 represents the graph of `p'-value measures. In this X-axis represents the number of documents and Y-axis represents the `p'-value measures.

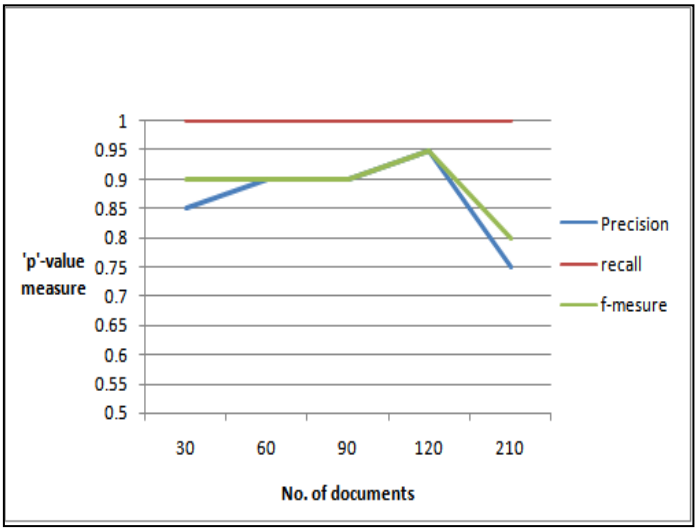


**Figure 5: Graph of 'p'-value measure**

## VII. CONCLUSION

We proposed ATD approach to detect cluster of anomalies from input dataset. Traditional approaches of anomaly detection such as, MGMM and FGM can efficiently works on high density dataset. But it can only detect individual anomaly from huge data input which is infeasible task. Hence, proposed approach mainly aims to discover group anomaly. PTM model is utilised for normal topic discovery in training phase whereas, in testing phase, it is used to construct $M_1$. Anomalies are nothing but abornal patterns, in cluster formation relevance score is used for construct anomaly cluster's. With proposed work system contributes emerging trend detection. With experimental set up, propopsed system proves it's efficiency in terms of accuracy.

## REFERENCES

[1] A.Srivastava and A. Kundu, "*Credit card fraud detection using hidden Markov model*," IEEE Transactions on Dependable and Secure Computing, vol. 5, no. 1, pp. 37–48, 2008.

[2] H. Soleimani and D. J. Miller, "*Parsimonious Topic Models with Salient Word Discovery*," Knowledge and Data Engineering, IEEE Transaction on, vol. 27, pp. 824–837, 2015.

[3] W. Wong, A. Moore, G. Cooper, and M. Wagner, "*Rule-based anomaly pattern detection for detecting disease outbreaks*," 2002.

[4] K. Das, J. Schneider, and D. B. Neill, "*Anomaly pattern detection in categorical datasets*," 2008.

[5] X. Dai, Q. Chen, X. Wang, and J. Xu, "*Online topic detection and tracking of financial news based on hierarchical clustering*," in Machine Learning and

 DOI : 10.18231/2454-9150.2017.0030 

Cybernetics (ICMLC), 2010 International Conference on, pp. 3341–3346, 2010.

[6] X. Dai, Q. Chen, X. Wang, and J. Xu, "*Online topic detection and tracking of financial news based on hierarchical clustering*," in Machine Learning and Cybernetics (ICMLC), 2010 International Conference on, pp. 3341–3346, 2010.

[7] M. Zhao and V. Saligrama, "*Anomaly Detection with Score functions based on Nearest Neighbor Graphs*," in Advances in neural information processing systems, pp. 2250–2258, 2009.

[8] L. M. Manevitz and M. Yousef, "*One-Class SVMs for Document Classification*," Journal of Machine Learning Research, vol. 2, pp. 139– 154, 2001.

[9] R. Yu, X. He, and Y. Liu, "*GLAD : Group Anomaly Detection in Social Media Analysis*," in Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining, pp. 372–381, 2014.

[10] A.P. Dempster, N. M. Laird, and D. B. Rubin, "*Maximum likelihood from incomplete data via the EM algorithm*," Journal of the Royal Statistical Society., vol. 39, no. 1, pp. 1–38, 1977.

[11] X.-L. Meng and D. Van Dyk, "*The EM algorithm–an old folk-song sung to a fast new tune*," Journal of the Royal Statistical Society: Series B (Statistical Methodology), vol. 59, no. 3, pp. 511–567, 1997.

[12] Q. He, K. Chang, E.-P. Lim, and A. Banerjee, "*Keep it simple with time: A reexamination of probabilistic topic detection models*," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 32, no. 10, pp. 1795–1808, 2010.

[13] H. Soleimani, D.j. Miller, "*ATD: Anomalous Topic Discovery in High Dimensional Discrete Data*", IEEE transaction on knowledge and data mining.

[14] M. Zhao and V. Saligrama, "*Anomaly Detection with Score functions based on Nearest Neighbor Graphs*," in Advances in neural information processing systems, pp. 2250–2258, 2009.

[15] Shewale Yogita, Prof. Shinde Jayashri, "*A Discovery of Group Anomaly and Emerging Topics Using ATD Approach*," International Journal for Research in Engineering Application & Management (IJREAM), ISSN : 2494-9150 Vol-02, Issue 08, pp. 06–09, 2016

[16] S. Wilks, "*The large-sample distribution of the likelihood ratio for testing composite hypotheses*," The Annals of Mathematical Statistics, vol. 9, no. 1, pp. 60–62, 1938.

[17] G. Schwarz, "*Estimating the dimension of a model*," Annals of Statistics, vol. 6, no. 2, pp. 461–464, 1978.

[18] Dataset: http://www.daviddlewis.com/resources/test collections /reuters 21578/

 DOI : 10.18231/2454-9150.2017.0030 