# Survey Paper on Data Mining Techniques: Outlier Detection and Text summarization

Divya Goyal, Research Scholar, Hardeep Singh, A.P. Dept. CSE at LPU, Jalandhar.

***Abstract***—In this paper, we will discuss all the researches we have find till. And what we have concluded from that survey. We try to compare and combine two subjects that are Natural language Processing and Data Mining. We will work on Outlier Detection and Text summarization. Data mining is the process of extraction of data that would be of any kind and Outlier is detection of irrelevant data. Natural Language Processing is any kind of process that would be applied on natural language to make it understandable for other than any person and Text summarization is the task which can complete with the help of NLP procedures and is useful to compress the text**.**

***Index Terms***— Data mining, Natural Language Processing, Outlier Detection, Text summarization**.**

## I. INTRODUCTION

THISpaper is a survey for text mining with text summarization. In this we discuss our whole survey which we have done on Outlier Detection and Text Summarization till the time in the favor of partial fulfillment of thesis work for Masters of technology. I have studied various definitions of Outlier Detection and Text Summarization given by the different expert researchers from there research results.

Fromthis introduction part, reviewer can get how data is arranged in the paper.Literature Review is in second section in which we will discuss whole the researches done till which we have studied. Data mining is third section of our paper, which tells us about the extraction process on text and the various applications that come under the data mining, their respective uses and various other tasks which can be done with the help of data mining. Natural language Processing would be fourth part, which helps us to focus on Text summarization and describe briefly other tasks that come under Natural Language Processing. In fifth section, Result will show how we relate or combine these two subjects with each other. From which we conclude, that we have to concentrate on the research of Novel summarization by the Hybrid approach of Text summarization and outlier detection.

## II. LITERATURE REVIEW

A literature review goes beyond the search for information and includes the identification and expression of relationships between the literature and our field of research. While the layout of the literature review may fluctuate with different types of studies, the basic purposes remain constant: Base paper.

[2] **PrakashChandore -** Data mining is a field of research area where the work is based on the knowledge discovery. There is a problem with detecting the outliers over the dynamic data stream and the specific techniques are used for detecting the outliers over streaming the data in data mining. The data stream mining is an active research of data mining. A data stream is defined as a sequence of data elements that are continuously being generated at a faster rate. A large amount of data is being inserted and queried continuously in streaming. Outlier detection has many applications in data stream analysis. It has become an important aspect for finding and removing outliers over the data stream. There are two main groups that we can consider for large data set to detect outliers and to analyze the data stream. The first group can be referred as the data stream and data mining techniques and the other group can be referred as to mine the data stream using different efficient algorithm. Detecting the outliers and analyzing the data has led to some unexpected knowledge in fraud detection, web document, etc. Its efficiency is dependent on the type data and the distribution data. [3]**ManzoorElahi-** Anomaly detection is an active research problem in many fields and is involved in numerous applications. Much of the methods that exist are based on distance measure that can give better results as compared to other methods. But in data stream these methods are computationally not much efficient. To find some specified number of neighbors for each data element, these methods are improper for the data stream environment due to huge volume of data. To declare a point as an outlier as soon as it comes, can frequently lead us to a false decision. There is an algorithm that partitions the incoming stream into chunks. The preliminary clusters are then made in each chunk during the initial stage. In the next stage, for each cluster the K-nearest neighbor approach for outlier detection is applied. Several experiments on different datasets confirm that our technique can find suitable outliers with low computational cost than the other exiting distance based approaches for outlier detection over data stream.[4] **JiadongRen-** An outlier detection algorithm is based on a

I am DivyaGoyal. I am pursuing my Mtechin Computer Science from Lovely Professional University. My registration id is 11210803. My email id is goyal.d1010@gmail.com.. I have done my work under the guidance of Mr.HardeepSingh, assistant professor,DeptCSE,Lovely Professional University,Jalandhar.
Lovely Professional University, Jalandhar-Delhi G.T. Road(NH-1), Phagwara, Punjab(India)-144411.

heterogeneous data stream which divides the stream into chunks. Then that each chunk is clustered and the result of that clustering is then stored in the cluster references. The degree and the number of neighboring cluster references of each cluster reference are figured out to generate the final outlier references that include the possible outliers. The experimental results show that the approach has higher detection precision. In anomaly detection the data stream outlier mining is an important issue. Much of the existing outlier detection algorithms can only handle numeric attributes or categorical attributes. [5]**DragoljubPokrajac-** As we know that in many industrial and financial application the Outlier detection has become an important problem. The problem further becomes complicated by the fact when the outliers have to be figured from the data streams that arrive at a larger pace. An incremental LOF (Local Outlier Factor) algorithm that is appropriate for detecting the outliers in data streams, are suggested. It provides an equivalent detection performance as the iterated static LOF algorithm, thus requiring less computational time. This algorithm also dynamically updates the profiles of data points. As data profiles may change over time so it an important property. There is theoretical evidence that inserting a new data point or deleting an old data point effects only few or limited number of their closest neighbors. Thus, the total number of points $N$ in the data set are not equivalent to the number of updates per such insertion/deletion. The experiments that are performed on several simulated and real life data sets have shown that the proposed incremental LOF algorithm is computationally efficient, but it is successful in detecting outliers in various data stream applications. [6]**MaysoonAbulkhair**- There is a formative model which is been constructed for figuring Intelligent Discharge System (IDS) automatically without any human intervention. Manyorganizations and the healthcare providers are concerned on implementing the new tools and applications so as to improve the quality of the services that are provided and to speed up the processing time. The method is the DSR integration that is composed of text summarization techniques which provides the ability to fetch the data which is important from the huge size of text which is collected from the different parts of the patients HMR. Discharge summary report (DSR) is a restrict report that take in the physicians efforts and time to integrate it. There are only few researches that exists on automating the integration of DSR, which is based on combining various patient data from his/her Hospital Medical Reports (HMR). To reduce the time for generating the DSR we use Natural Language Processing (NLP) techniques such as Segmentation, Stemming and Stop word Filtering. As a result the DSR output which is produced and is been approved by the physician should contain all the required fields that in the report. [7]**Ha Nguyen Thi Thu "***The World Wide Web has brought us a vast amount of online information. When we search with a keyword, data feedback from many different websites and the user cannot read all the information. So that, text summarization has become a hot topic, it has attracted experts in data mining and natural language processing field. For Vietnamese, some methods of text summarization are based on that have been proposed for English which have brought some significant results. There is still some difficult problems that are remaining which are to treat with the Vietnamese language processing, typical in this is the Vietnamese text segmentation tool and text summarization corpus. There is a Vietnamese text summarization method based on sentence extraction approach using neural network for learning combine reducing dimensional features to overcome the cost when building term sets and reduce the computational complexity. The experimental results show that our method is really effective in reducing computational complexity, and is better than some methods that have been proposed previous".*[8]**Shu Wu-** Outlier detection can be considered as a pre-processing step that locate those objects which do not behave as a well-defined notation as per to the expected behavior in the data set. There are two practical 1-parameter outlier detection methods that require no user defined parameters for deciding whether an object is an outlier or not , we just need to give the number of outliers we want to detect and these methods are named as ITB_SS and ITB_SP which are very effective and efficient and can be used for large and high dimensional data sets. It is important in data mining for discovering the novel or exceptional phenomena,etc. A new concept of holoentropy which takes both entropy and total correlation into consideration is been proposed for a formal definition and optimization model of outlier detection. As it is a challenging problem as we face the difficulty of defining something meaningful which is similar to the categorical data.[9]**FabrizioAngiulli-** There is an introduction of a distributed method for detecting distance-based outliers in very large data sets. A small sub set of data that can be employed for predicting the novel outliers that can be done using the approach which is based on the concept of the outlier detection solving set. There is a vast time savings as it works on parallel computation and it shows excellent performance. For the increasing number of nodes the algorithm is efficient and its running time scales is quiet well. In order to decrease the overall runtime and to improve the communication cost, a basic strategy is been discussed so as to reduce the amount of data to be transferred. The temporal cost of this algorithm is expected to be at least the three orders of magnitude which is faster than the classical nested loop like approach to detect the outliers. Thus the solving set has the same quality as that is been produced by the centralized method which is been configured by this approach in a distributed environment. [10] **Tuomo W. Pirinen-** When errors are large and there is a huge amount of corrupted data the statistic is effective in measuring estimation reliability and identifying outliers. Estimating errors with less information that are encountered in sensor systems which requires outlier rejection and self-diagnosis is a difficult problem. In sensor arrays this work proposes a confidence statistic and an outlier detector for different estimates. It is applicable to a variety of estimates from a generalized difference quantity model. To detect the presence of outliers the confidence statistic is used. It is demonstrated that when the statistical assumptions are not met the analytical results are useful and the performance is examined.

**[11]QiangShen-** To handle the data measurements for event detection there is a presentation of an adaptive online outlier detection. There is a need to design outlier detection algorithm so as to consider those limitations on the power, memory, etc due to the cheap and low quality sensor devices which are used for event detection in internet of things (IOT). This algorithm will provide the capability to tolerate that data which will be considered as outliers by traditional algorithms. TAOOD named as tolerance-based adaptive online outlier detection algorithm is proposed for an accurate ratio of outliers and a tolerance parameter. TAOOD by discarding the duplicate data and outliers, it decreases the amount of transmitted data andeliminates the limitation of original window-base outlier detection algorithm by adapting an accurate ratio of outliers and a tolerance parameter.[12]**Bo Jiang**- Spatial outlier detection method of studying multiple non-spatial attributes is based on the special objects. It uses the SOFMF algorithm. This algorithm summarize and analyze how to overcome the insufficiency of many clustering algorithms and find clusters in different shapes. It checks the non-sensitive input data sequence and processes the noise data and multi-dimensional data well and have multi-resolution. A novel idea for spatial clustering data is been proposed that proves this idea can be applied to spatial clustering. [13]**NaumanShahid**- The introduction of Support-Vector Machines (SVM) have received a great interest in the machine learning community especially in Outlier Detection in Wireless Sensor Networks (WSN). The Quarter-Sphere formulation of One-Class SVM (QS-SVM), extends the main SVM ideas from supervised to unsupervised learning algorithms. It has a non-ideal performance,as the QS-SVM formulation is based only on Spatio-Temporal correlations between the sensor nodes. Based on the novel concept of Attribute Correlations between the sensor nodes, this work presents a new One-Class Quarter-Sphere SVM formulation hence the name, Spatio-Temporal-Attribute Quartersphere SVM (STA-QS-SVM) formulation. Using this formulation Online and partially online approaches have been presented. Over the previous formulation (ST-QSSVM) the results indicate a remarkable reduction in the False Positive rates and a significant increase in the Outlier Detection rates. There by conserving significant computational and communication complexity, the results of this novel technique also suggest that the partially online approach is as efficient as the online approach.**[14] Michael T. Mills-** For natural language processing and natural language understanding there is a survey and analysis which presents the performance, functional components and maturity of graph-based methods for their potential for mature products. The capabilities that are resulted from the methods being surveyed include the redundancy reduction, text summarization, disambiguation, text entailment, similarity measures and novelty detection. Each method derives the estimated scores for coverage, scalability, performance and accuracy. From a collection of graph based methodologies with tables and graphs there is a unique abstraction of functional components and levels of maturity.**[15]YogitaThakran-** Streaming data is very challenging in outlier detection as it cannot be scanned multiple times and over time the new concepts keep evolving in coming data. Also the noisy attributes mislead the working with data streams. Here we propose a scheme which is based on clustering which does not require a labeled data but is an unsupervised outlier detection scheme for streaming data. It also takes the advantage of density based and partitioning clustering method. To reduce the effect of noisy attributes, weighted attributes are used and is adaptive to concept evolution. Results of this approach perform other existing approaches for outlier detection rate, increasing percentages and false alarm rate. **[16] ElioMasciari-** Trajectory data streams are continuously generated from different sources exploiting a wide variety of technologies. Thus, mining such a huge amount of data which belongs to time and position of moving objects is challenging as the possibility to fetch some useful information from this data is crucial in many applications. There are some interesting challenges which are possessed by spatial data for their proper definition and acquisition and hence for the classical point data the mining process becomes harder. Tn this there is a problem of trajectory data outlier detection that brings out some challenges when we deal with the data for whom the order of elements is not relevant. A complete framework is proposed which allows us to make an effective mining step. We propose a complete framework starting from data preparation task that allows us to make the mining step quite effective.

## III. DATA MINING

The term 'data mining' either refers to detect the irrelevant data from huge amount of collection with respect to a particular topic or extraction of useful data form huge amount of irrelevant data. We explain these both definitions with the help of real life experiments: 1) when we detect irrelevant data from huge amount that should be similar with extract impurities from the water as show in fig 1. And 2) when we extract useful data from huge amount of irrelevant data that should be similar with extraction of iron from impurities as show in fig 2. In first point the relevant data is huge in amount and irrelevant is less that's why we prefer the detection because it takes less processing time. On the other hand, Second point state the data which is useful for us is less and the data which is useless is much more in size, so we prefer extraction of data that is useful.

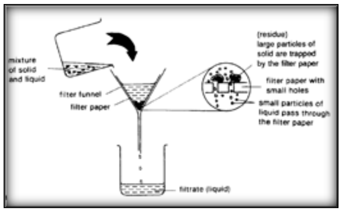

Fig1. Water purification

Fig2. Iron extraction

Data mining has 4 main types: Text, Audio, Image and Web. Text mining is used where we have collection of words which either makes sentence or paragraph.Audio mining is used by the forensic experts mainly to recognize the voice of any criminal or recognize any other voice which should be helpful to solve the case. Image mining generates a new subject. It is a complete study of digital image processing. Web mining is further divided into two types: Web Usage Mining and Web Log Mining.

Tasks come under the data mining are:

- **Anomaly Detection / Outlier Detection** : Separate the irrelevant data from the large data set of relevant data. Using seven methods we can detect the outliers or anomaly:

i. Sliding Window Based**:** Work with fix window size.

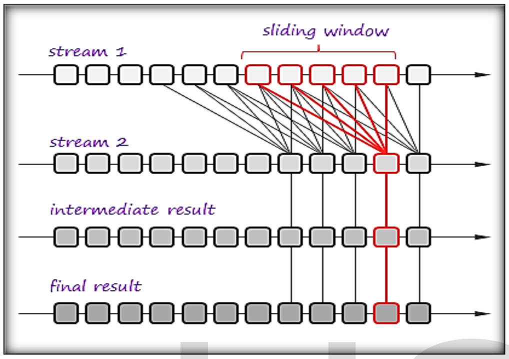

Fig.3 Sliding Window Concept

ii. Auto Regression Based: Metrics and model based.
iii. Clustering Based: Grouping of similar datasets.
iv. Density Based: Nearest neighbor density comparison
v. Statistical
vi. Distance based: distance from fixed dataset either greater or lower.
vii. Derived: Numbers of techniques are getting together.

- **Association Rules**: The association rules are established between those datasets who have some similar properties with each other but are different in nature. There are various kind of association rules which we are discussing below:

i. Contrast set learning
ii. Weighted Class learning
iii. High-order pattern Discovery
iv. K-Optimal pattern Discovery
v. Generalized Association Rules
vi. Quantitative Association Rules
vii. Interval Data Association Rules
viii. Maximal Association Rules

Most common types of algorithms are used under the heading of Association are:

1. Apriori Algorithm
2. Eclat Algorithm
3. FP-Growth Algorithm

- **Clustering**: objects with more similarities are grouped into a single group, remaining belongs to different. This task can be used in data mining.

- **Classification**: divide the large data into number of different classes.

- **Summarization**: to decrease the size of text in length but do not put much effect on meaning.

## IV. Natural Language Processing

The term Natural Language Processing refers to the process applied on natural language to make it understandable for any other person. Tasks come under NLP:

i. Machine translation
ii. Word sense disambiguation
iii. Named Entity Recoganisation
iv. Text Summarization
v. Morphologic
vi. Segmentation
vii. Sentence breaking
viii. Parsing
ix. Automatic summarization

After the completionof theprocess of data mining, we have to re arrange the sentences. So we need to process the data using NLP.

## V. Conclusion and Future Work

In this paper we discussed the techniques and methods used by different researchers and basic introduction and algorithms are discussed in brief.At the end, we conclude that the outlier detection is must to increase the speed of processing of data and also helpful in reduction of processing cost of data. We have to compare all the techniques which are effectively used till the time and get the best one as the result of this comparison and reliable to implement. Try to achieve the objectives which are defined previously.

## References

### WEBSITES:

[1].AvailableAt:[http://www.anderson.ucla.edu/faculty/ jason.frand/teacher/technologies/palace\ datamining.htm ; Dated: oct,2013.

[17].Availableat:[http://en.wikipedia.org/wiki/Automatic_summarization]

[18].Avaiableat:[http://www.kdnuggets.com/gpspubs/aimag-kdd-overview-1996-Fayyad.pdf ]

### JOURNALS:

[2]. P.Chandore, P.Chatur,'*Outlier Detection Techniques over Streaming Data in Data Mining: A Research Perspective*', *International Journal of Recent Technology and Engineering (IJRTE)* ISSN: 2277-3878, Volume-2, Issue-1**,** *March 2013***.**

[3].ManzoorElahi, XinjieLv, "DB-Outlier Detection Algorithm using Divide andConquer approach over Dynamic,International Conference on Computer Science and Software Engineering DataStream", 2008.

[4].JiadongRen, Qunhui Wu, Jia Zhang, '*Efficient Outlier Detection Algorithm for Heterogeneous Data Streams*',Sixth International Conference on Fuzzy Systems and Knowledge Discovery, *2009.*

[5].DragoljubPokrajac, '*IEEE Symposium on Computational Intelligence and Data Mining (CIDM)*',April 2007.

[6]. MaysoonAbulkhair,"*Intelligent Integration of Discharge Summary: a Formative Model*" ,2013. 4th International Conference on Intelligent Systems, Modelling and Simulation

[7]. Ha Nguyen Thi Thu, "*A Supervised Learning Method Combine With Dimensionality Reduction In Vietnamese Text Summarization*", 978-1-4673-6044-9/13/$31.00 ©2013 IEEE

[8].Shuwu , '*Information-Theoretic Outlier Detection for Large-Scale Categorical Data* ', *ieee transactions on knowledge and data engineering*, vol. 25, no. 3, M*arch 2013.*

[9]. FabrizioAngiulli, '*Distributed Strategies for Mining Outliers in Large Data Sets*', *ieee transactions on knowledge and data engineering*, vol. 25, no. 7, July *2013.*

[10]. Tuomo W. Pirinen, '*A Confidence Statistic and an Outlier Detector for Difference Estimates in Sensor Arrays*', *ieee sensors journal, vol. 8, no. 12,* December 2008

[11].QiangShen,'*Tolerance-based Adaptive Online Outlier Detection for Internet of Things*', *2010 IEEE/ACM International Conference on Green Computing and Communications & 2010 IEEE/ACM International Conference on Cyber, Physical and Social Computing.*

[12].Bo Jiang ,'*Spatial Outlier Detection Algorithms Based on Knowledge Discovery*',©2009 IEEE.

[13].NaumanShahid,'*Quarter-Sphere SVM: Attribute and Spatio-Temporal Correlations based Outlier & Event Detection in Wireless Sensor Networks*', 2002 IEEE wireless communication and networking Conferece: Mobile and wirless Networks.

[14]Michael T. Mills,'*Graph-Based Methods for Natural Language Processing and Understanding—A Survey and Analysis*',IEEE Transactions On Systems, Man, And Cybernetics: Systems.

[15].YogitaThakran,'*Unsupervised Outlier Detection in Streaming Data UsingWeighted Clustering*',2012 IEEE.

[16.]ElioMasciari ,'*Trajectory Outlier Detection Using An Analytical Approach*', 2011 23rd IEEE International Conference on Tools with Artificial Intelligence.