ISSN No:-2456-2165

A Survey of Text Classification and Clustering Issues in Data Mining

M.A.MARIA PARIMALA[1], K. VINOTHA[2]
[1]M.C.A, M.Phil Assistant Professor, [2]Research Scholar
P.G. & Research Dept. of Computer Science, St. Joseph's College of Arts and Science (Autonomous) Cuddalore

Abstract:- Data mining is defined because the strategies to extract the records from huge quantity of information. It is the technique to outline to essential knowledge from the massive quantity of statistics stored in database or in the information warehouse. Clustering is an unmanaged technique of Data Mining. It manner grouping similar items collectively and setting apart the assorted ones. Clustering is to divide the facts in to comparable organization of facts the statistics institution can be together is referred to as clustering. The goal of this survey is to provide problems and challenges on clustering and class the usage of information mining techniques.

Keywords:- *Data Mining, Clustering, Classification, Clustering Algorithm.*

I. INTRODUCTION

Data mining is likewise referred to as information discovery studying or information. Is the process to analyses the records data. Data mining is process of looking and study of big information set to search out the guidelines and pattern. This step is known as expertise discovery learning in database. Data mining(DM) is way to construct the difference among the information and statistics. In system mastering technique has been used they may be supervised gaining knowledge of and un-supervised approach, Generally there are three principal records mining techniques are regression, class and clustering. Data mining is likewise referred to as expertise discovery in databases (KDD).

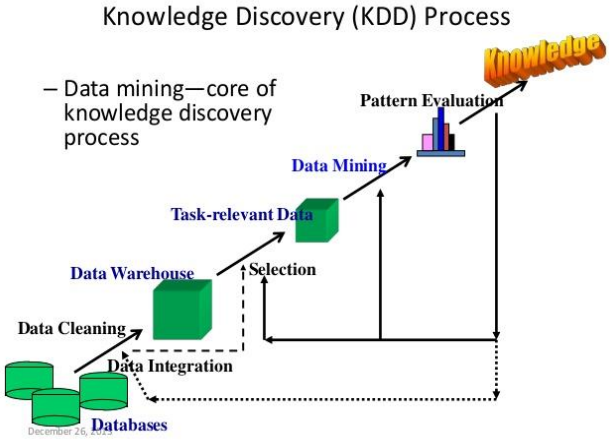

Fig 1

- Data Cleaning: It is the manner of dispose of noise and inconsistent data.
- Data Integrating: Is the method of integrate facts from multiple resources.
- Data choice: It is process of retrieving relevant facts from the database.
- Data transformation: In this technique records are converted or consolidated into forms appropriate for mining by using acting summary or aggregation operations.
- Data mining: It is critical system where shrewd strategies are implemented if you want to extract statistics styles.
- Pattern evolution: the patterns acquired in the facts mining degree are converted into information based totally on some interestingness measures.

II. LITERATURE SURVEY

Vikas K Vijayan,Bindu K.R,Latha Parameswaran[1] Text mining is the process of removing information and classifying the text documents, it includes spam Filtering, email routing, sentiment analysis , language identification, hidden semantics and high dimensional features are the limiting factor and it has its own pros and cons. So, outcome of classification is improve by using right choice classifier and appropriated reduction technique.

Victoria Bobicev[2] Text mining is an analysis to developed direction. Machine learning classification of sentiment labels and conditions in multiple labels. The average of F-measure is equal to 0.805 .For confusion the F-measure is 0.816 and gratitude the F-measure is 0.899 with the help of word based feature. The average result is sentiment is 0.805.

Xia Huosong and Liu Jian[3] Feature selection is a major step in text classification which damages accuracy and validity. It use four feature selection algorithm DF,MI,IG and CHI. Feature selection is used for integrated learning algorithm. The ILA is used for text categorization in aggressiveness of stop words. So the integrated learning algorithms improve the recall and precision.

A, Basu, C.Watters, and M.Shephered [4] Text categorization is to sort text document into predefined categorize of similar documents. Automatic classification is to improve the filtering and vocabulary to minimize the feature set. The text categorization of has been used two algorithm ANN and SVM. The SVM has better result for

precision and recall. And reduce the vector size as an improved performance.

Jiyuan An,Yi-ping Phoebe Chen[5] Text categorization is an application for machine learning .Some of the methods in machine learning like Rocchio method, naïve bayes method, SVM based method for text classification. It uses a robust algorithm based on enumeration for keyword. Its reduce keyword combination give the better result in robust algorithm.

M.Krendzelak and F.Jakab [6] Text categorization is the use of machine learning to a set of categories, with the use of hierarchical structure is classified to build. The performance of hierarchical trials are showed a difference .So the work in McCallum et al on improved method for probability classification in text hierarchies.

Anmol Kumar,Amit kumar Tyagi[7] Mining of data is well known technique for automatic and intelligent to extract information or knowledge from a huge amount of data.Data mining is a multidisciplinary field,drawing work from areas including database technology,machine learning,pattern recognition, it discuss about challenges and issues in field of data mining.

Sukhvir Kaur [8] Data mining as the method to extraction of data to huge amount of information clustering is an important technique in data analysis and data mining application, the aim of the paper is clustering is the intrinsic grouping in a set.

- *Taxonomy of Clustering and Classification*

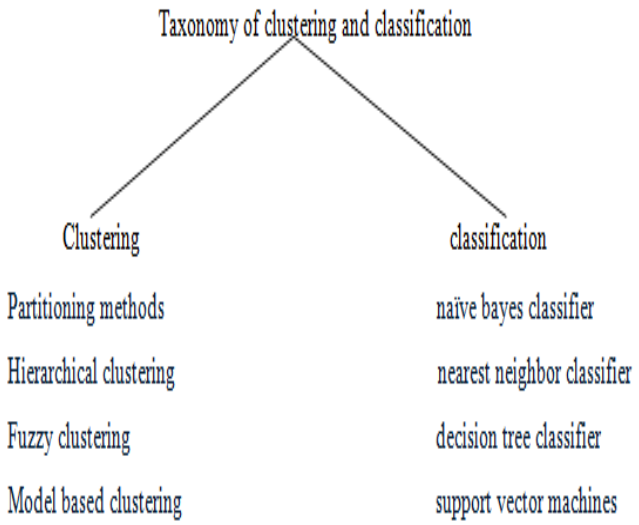

Fig 2

Clustering is method of organization the set of objects based totally on characteristics, aggregating them in keeping with their similarities. In information mining their method walls the facts implementing a algorithms. Clustering strategies is to pick out businesses of equal gadgets in statistics gathered. They are specific kinds of clustering techniques,[6]

A. Partitioning Strategies
B. Hierarchical Clustering
C. Fuzzy Clustering
D. Model Based Clustering

A. Partitioning Clustering

Partitioning algorithm is a clustering strategies which might be separate the information into a set of K businesses, wherein K is some of agencies pre-exact with the aid of the records .There are special styles of partitioning clustering techniques maximum famous is K - suggest clustering.

B. Hierarchical Clustering

Hierarchical clustering is special method examine to partitioning clustering is used to discover the group within the dataset .It might not require the pre-specify the wide variety of clusters to be generated. It is tree primarily based representation of the items.

C. Fuzzy Clustering

Fuzzy clustering is also called smooth method. Standard clustering tactics produce walls (K mean)it belong to most effective one cluster. This is known as tough clustering

III. CHALLENGES IN CLUSTERING

A. Large Quantity of Samples

The number of pattern to be processed could be very high. Algorithm, ought to be very aware of scaling issues. Clustering in preferred is NP-hard.[6]

B. High Dimensionality

The wide variety of functions could be very high and may even exceed the quantity of pattern.

C. Sparsity

Most capabilities are zero for most samples, i.e. The item-characteristic matrix is sparse. It affect the size of similarity and the computational complexity.

D. Strong No-Gaussian Distribution of Feature Values

The statistics is so skewed that it can be correctly modeled with the aid of normal distributions.

E. Significant Outliers

Outliers can also have big significance.Finding these outliers is relatively non-trivial, and putting off them isn't always applicable.

F. Legacy Clustering

Previous cluster analysis results are regularly available. This know-how should be reused rather than beginning every evaluation from scratch.[7]

➢ *Steps and Techniques used in Clustering*

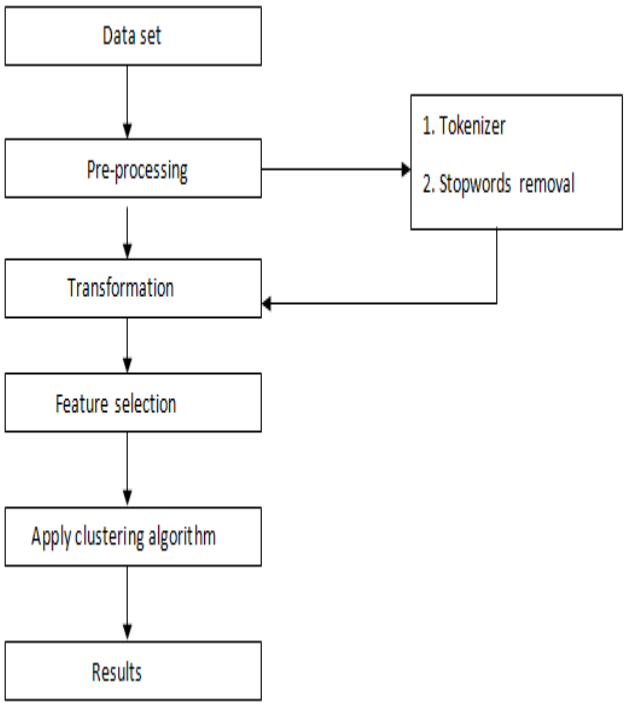

Fig 3

IV. CLASSIFICATION

Text type is the method of classifying text report into a predefined set of class it may now not straightly locate the herbal language we trade the format then only the computer can apprehend the text illustration model is vector space version(VSM).It as a supervised mastering approach in which a training set of files.[8]

A. Naïve Bayes

Naïve bayes classifier is the most popular studying technique grouped by using similarities that paintings on the famous bayes theorem of opportunity and document category.[3]

B. Support Vector System

Support vector machine (SMV) is supervised learning type set of rules for type or regression troubles where the dataset teaches SVM about the training in order that SVM can classify any new statistics.

C. Nearest Neighbor Classifier

Nearest Neighbor classifier is a proximity-based totally classifier which use distance-based measures to carry out the type.The documents which belong to the equal elegance are more “similar” based on similarity measure.

D. Decision Tree Classifiers

Decision tree is a graphical illustration that uses branching methodology to all feasible outcomes of a decision, based on positive situations.The tree represents the outcome of the check and the leaf node represents a particular class label.

V. TEXT CLASSIFICATION MEASURES

If there are a number of the measures are used to evaluate the classification methods include [8]

➢ *Accuracy*
➢ *Precision*
➢ *Recall*
➢ *F-degree*

➢ *Precision*

He field of data retrieval, precision is the fraction of retrieved files which might be applicable to the search.

➢ *Recall*

Recall in information retrieval is the fraction of the documents that are relevant to the query that are successfully retrieved.

➢ *F-Measure*

The weighted harmonic mean of precision and recall, the traditional F-measure or balanced F-score is:

VI. ISSUES OF CLASSIFICATION

A. Data Education

➢ Data cleaning: Preprocessing step to reduce noise and manage missing values
➢ Relevance analysis (feature selection) “Remove inappropriate or wrong attributes
➢ Data transformation and reduction:
➢ " Generalize data to (better concepts, discretization) “
➢ " Reduce the dimensionality of the facts”

B. Evaluating Classification Methods

➢ Accuracy:
➢ Classifier accuracy: the ability of a classifier to expect class labels
➢ Predictor accuracy: how near is the predicted price from one
➢ Speed:Time to assemble the version (training time)time to apply the model (type/prediction time)
➢ Robustness: managing noise and lacking values
➢ Scalability :efficiency in disk-resident databases
➢ Interpretability: Level of understanding and insight supplied by way of the version

VII. CONCLUSION

In this paper we have included the clustering and type algorithm. We have additionally defined the issues faced in clustering and the class. At final we've described the number of the textual content classification degree utilized in clustering.

REFERENCES

[1]. V. K. Vijayan, K. R. Bindu, and L. Parameswaran, "A Comprehensive Study of Text Classification Algorithms," pp. 1109–1113, 2017.
[2]. P. Agarwal, M. A. Alam, and R. Biswas, "Issues , Challenges and Tools of Clustering Algorithms," vol. 8, no. 3, pp. 523–528, 2011
[3]. X. Huosong, "The Research Of Feature Selection of Text Classification Based On Integrated Learning Algorithm," 2011.
[4]. A. Basu, C. Watters, and M. Shepherd, "Support Vector Machines for Text Categorization," pp. 1–7, 2002.
[5]. J. An and Y. P. Chen, "Keyword Extraction for Text Categorization," pp. 556–561.
[6]. M. Krendzelak and F. Jakab, "Text Categorization with Machine Learning and Hierarchical Structures."
[7]. Anmal Kumar, amit Kumar Tyagi,"Data Mining:Various Issues and Challenges for Future A Short discussion on Data Mining issues for future work"
[8]. Sukhvir Kaur "A survey of different data clustering algorithm"IJCSMC,Vol.5,Issues.5, May 2016,Pg.584-588
[9]. S. Kaur, "SURVEY OF DIFFERENT DATA," vol. 5, no. 5, pp. 584–588, 2016..
[10]. M. Krendzelak and F. Jakab, "Text Categorization with Machine Learning and Hierarchical Structures."