A review on mining text data with auxiliary attributes

[1]Greeshma RG, [2]Smitha ES
[1]P G Scholar, [2]Associate Professor
[1]Department of CSE, [2]Department of IT
[1]LBSITW, Poojappura, Thiruvananthapuram

***Abstract*-Clustering is a widely studied data mining problem in the text domains. The problem finds numerous applications. The clustering problem is defined to be that of finding groups of similar objects in the data. The similarity between the objects is measured with the use of a similarity function. The problem of clustering can be very useful in the text domain, where the objects to be clusters can be of different granularities such as documents, paragraphs, sentences or terms. Clustering is especially useful for organizing documents to improve retrieval and support browsing. Text mining is the analysis of data contained in natural language text. Text databases are rapidly growing due to the increasing amount of information available in various electronic forms. User need to access relevant information across multiple documents. In many text mining applications side information is available along with the text documents. Here side information is referred to as auxiliary attribute. This information corresponds to different kinds of attributes such as the document provenance information, information related to origin of documents etc. Such side information may contain a huge amount of information. This huge amount of information may be used for performing clustering. This paper represents review on most clustering techniques containing different kinds of data.**

***Index terms*- Classification, Clustering, Data mining, Side information, Text Mining**.

I. INTRODUCTION

Data mining is the exercise of automatically searching large stores of data to discover patterns and trends with simple analysis. Many researchers have used techniques such as classification, outlier detection, clustering, regression analysis etc. The clustering is used some special application. Clustering is mechanisms of combining set of physical or abstract objects into classes of similar objects. There are different orders or groups which is called cluster, subsist of objects that are correlated within themselves and unrelated to objects of other order or groups. Text mining is the discovery of new, previously unknown information by automatically extracting information from different written resources. Text mining is a variation on data mining that practically find out interesting patterns from large databases. The use of digital information is increasing day by day. Since increasing the amount of information it needs to extract relevant information from this huge amount of data for text mining. This proves to a reason in creating scalable and efficient mining algorithms. The clustering of data in the pure form is done till now. But to manage such large quantity of data we require indexing the data according to the users need. Large number of web documents contains side information. This side information can be sometimes called meta-data. These meta-data are exactly matching to the various different kinds of attributes such as the origin or other information related to the origin of the document. Data such as location, possession or even temporal information may prove to be informative for mining purposes in other cases.[9] Documents may be linked with user-tags in many network and user-sharing applications. This may also be quite informative for doing effective text mining. The process of deriving high quality information from text is known as text data mining. The side-information can sometimes provide useful information for improving the quality of clustering process, but when the side-information is noisy it can be a risky approach. A method is needed to discover the coherence of clustering characteristics of side information with the text content and at the same time reject those aspects in which incompatible clues are provided.

II. TEXT PREPROCESSING

Mining from a pre-processed text is easy as compare to natural languages documents. So, it is important to pre-process the text before clustering. To reduce the dimensionally of the documents words, special methods such as filtering and stemming are applied. Filtering methods remove those words from the set of all words which are irrelevant. Stop word filtering or stop words removal is a standard filtering method. Words like conjunctions, prepositions, articles, etc. are removed. Stemming is a technique used to find out the root/stem of a word. Stems are thought to be useful for improving retrieval performance because they reduce variants of the same root word to a common concept. For example, consider the words user, users, used, and using. The stem of these words is use. Similarly the stem of words engineering, engineered, and engineer is engineer. Furthermore, stemming has the secondary effect of reducing the size of the indexing structure because the number of distinct index terms is reduced. This can be done as follows.

- if a word ends with a consonant other than *s*, followed by an *s*, then delete *s*.
- if a word ends in *es*, drop the *s*.
- if a word ends in *ing*, delete the *ing* unless the remaining word consists only of one letter or of *th*.

- If a word ends with *ed*, preceded by a consonant, delete the *ed* unless this leaves only a single letter.

Many of the most frequently used words in English are worthless in text mining. These words are called stop words. For example the, of, and, to, etc. in stop word removal the stop words such as the , to etc. are removed.

III. LITERATURE SURVEY

Text clustering becomes a problem in many application domains due to the increasing amount of unstructured data. A general survey of text clustering algorithm can be found in [1]. In [2] major fundamental clustering methods are discussed. These methods can be classified as partitioning methods, hierarchical methods, density based methods and grid based methods [3]. K-means and K-medoids methods come under partitioning methods. They are distance based methods. In Distance-based Clustering Algorithms there is a use of similarity function which measures the closeness between the text objects takes place. The most well-known similarity function which is used commonly is the cosine similarity function. Hierarchical methods can be classified as agglomerative and divisive methods. In these methods clustering is a hierarchical decomposition. The general concept of agglomerative clustering is to successively merge documents into clusters based on their similarity with one another DBSCAN, DENCLUE are some of density based methods. With density based methods we can find arbitrarily shaped clusters. CLIQUE is one of the grid based methods. These methods use a multi-resolution grid data structure.

In [4] the cosine similarity to find the similarity of documents is explained. The documents can be represented as vectors. To compute the similarity between two vectors the following steps are used.

- Consider two vectors (say A and B).
- Take the union of those vectors.
- Find the dot product of vectors A and B.
- Calculate the magnitude of vector A and B.
- Multiple the magnitudes of A and B.
- Divide the dot product of A and B by the product of the magnitudes of A and B.

$$\text{Similarity} = \cos(\theta) = \frac{A \cdot B}{\|A\|\|B\|} = \frac{\sum_{i=1}^{n} A_i \times B_i}{\sqrt{\sum_{i=1}^{n} (A_i)^2} \times \sqrt{\sum_{i=1}^{n} (B_i)^2}}$$

In [5] C.C.Aggarwal explained about the different clustering algorithms. Data classification is a two-step process consisting of learning step and classification step. In learning step a classification algorithm builds the classifier by learning from a training set. In classification step a model is used to predict class labels for given data. Some methods which are commonly used for text classification are as follows. Decision trees, Rule based classifier, SVM classifiers, Bayesian classifier etc. Decision tree is a hierarchical decomposition of training set in which a condition on attribute value is used in order to divide the data space hierarchically. In rule based classifier a set of rules are used to model the data space. SVM classifiers attempt to partition the data space with the use of linear or non-linear delineations between the different classes. Neural network classifiers are related to SVM classifiers, both are in the category of discriminative classifiers. Bayesian classifiers build a probabilistic classifier based on modelling the underlying word features in different classes.

In [6] a scatter-gather technique is discussed. It is a cluster based approach to browse large document collections. It uses document clustering as its primitive operation. Here initially system scatters the collection into clusters and present short summaries to the user. Based on the summaries one or more groups are selected. The selected groups are gathered to form sub collection. This is an iterative process. Here partitioning methods are defined to partition the collection into clusters. Buckshot and fractionation algorithms are used to find initial clusters. Buckshot is a fast clustering algorithm needed for reclustering. Fractionation is a clustering algorithm with great accuracy. In [7] a co-clustering approach for documents and words is explained. Here documents and words are clustered simultaneously. The document collection can be represented as a word by document matrix. This word by document can then be represented as a bipartite graph. The dual clustering problem is done in terms of finding minimum cut in bipartite graph. A spectral algorithm is used to solve the partitioning problem. High dimensionality of feature space is a challenge for clustering algorithms [8].

Feature extraction and feature selection techniques are used to reduce feature space dimensionality. In feature extraction it extracts a set of new features from original features through some functional mapping. In feature selection it chooses a subset from the original feature set according to some criteria. Document frequency, information gain, term strength are some of the feature selection methods. Unsupervised feature selection methods are much worse than supervised feature selection. In order to utilize the efficient supervised method an iterative feature selection method that iteratively performs clustering and feature selection is proposed in this paper.

IV. CONCLUSION

This paper gives brief introduction about document clustering and classification. The increasing amount of text data in large collections led to the creation of scalable and efficient mining algorithms. Many works has been done by considering the text clustering problem. However all these works deals with pure text clustering. It does not consider other kinds of attributes. In many applications tremendous amount of side information can be seen. Sometimes this side-information may be noisy and it worsens the quality of clustering. So this work needs a way for performing the mining process, so that the advantages from using this side information can be maximized. This work needs to use an approach which carefully discovers the connection of the clustering characteristics of the side information with that of text content.

V. ACKNOWLEDGEMENT

I have taken efforts in this review of clustering for mining using side information. However, it would not have been possible without the kind support and help of many individuals. I am highly indebted to Associate Prof. Smitha ES for her guidance and constant supervision as well as for providing necessary information regarding this approach.

REFERENCES

[1] C. C. Aggarwal and C.-X. Zhai, "Mining Text Data", New York, NY, USA: Springer, 2012.

[2] J. and Kamber, M., Data Mining: Concepts and Techniques, 2nd ed., Elsevier, Morgan Kaufmann, 2006.

[3] A. Jain and R. Dubes, "Algorithms for Clustering Data", Englewood Cliffs, NJ, USA: Prentice-Hall, Inc., 1988.

[4] G. Salton, "An Introduction to Modern Information Retrieval. London", U.K.: McGraw Hill, 1983.

[5] C. C. Aggarwal and C.-X. Zhai, "A survey of text classification algorithms," in Mining Text Data. New York, NY, USA: Springer, 2012.

[6] D. Cutting, D. Karger, J. Pedersen, and J. Tukey, "Scatter/Gather: A cluster-based approach to browsing large document collections," in Proc. ACM SIGIR Conf., New York, NY, USA, 1992, pp. 318–329.

[7] I. Dhillon, "Co-clustering documents and words using bipartite spectral graph partitioning," in Proc. ACM KDD Conf., New York, NY, USA, 2001, pp. 269–274

[8] T. Liu, S. Liu, Z. Chen, and W.-Y. Ma, "An evaluation of feature selection for text clustering," in Proc. ICML Conf., Washington, DC, USA, 2003, pp. 488–495

[9] C. C. Aggarwal, S. C. Gates, and P. S. Yu, "On using partial supervision for text categorization," IEEE Trans. Knowledge. Data Eng., vol. 16, no. 2, pp. 245–255, Feb. 2004.