ISSN:0975-9646

An Effective Use of Meta Information Using Clustering and Classification Techniques for Text Mining: A Survey

Mr. Nitin J.Ghatge[1], Prof. Poonam D. Lambhate[2]

[1, 2] *Computer Engg. Department, JSMP's, JSCOE, Hadapsar*
Pune University, India

Abstract— **Aim and the review of this paper is immersed on effective clustering and mining approach with the help of Meta information. Meta information is nothing but information about information. Such information is presented as text documents in many text mining applications which may be different forms, such as document origin information, links in the documents, user access behaviour from web logs, other non textual attributes present into the text documents. Such Meta information can be useful in enhancing the quality of clustering process, but it is difficult to use the relative importance when some information is noisy. In such a case, it can aggravate the quality of the mining process. Therefore, we use an approach which combines classical partitioning algorithms with probabilistic models so that we can create an effective clustering method, so this proposed approach act as solution to maximize the benefits from using meta information.**

Keywords— **Data mining, Data clustering, Meta information, Text mining..**

I. INTRODUCTION

This in recent years there has been an increasing emphasis on big data in every field such as organizations, medicines, business, academics and government are exploring how large volume data can usefully be deployed to create and capture value for individuals, business and communities. A wide majority of data is stored in documents that are virtually unstructured. Text mining is the analysis and organizing of data contained in natural language text. Text mining works with phrases transposing words and transposing words in unstructured data into numerical values which can then be linked with structured data in a database and analysed with traditional data mining techniques. Clustering is a technique for automatically organizing a large collection of text. Text documents broadly occur in the context of a number of applications in which there may be a large amount of meta-information which may be useful to the clustering process. Some examples of such meta-information are as follows:

- We captured wed logs which contain meta-information about your site visitors: accessed files, information, path through, activity statistics the site about operating systems, referring pages, search engines, browsers and more. Such logs can be used to enhance the quality of the mining process.
- Various text documents having links in them, such as links contain a lot of useful information for mining purposes. This link used to evaluate relationships between nodes. The relationship can be identified among various types of objects, include people, organizations and transactions.
- Meta-Information in many web documents contains information about the origin of the documents, ownership and location of the documents which is also useful for mining purpose.

The main idea from the study of paper presented by Charu C. Aggarwal [8] is that, an effective use of meta-information, which may be used to improve the clustering techniques in order to design clustering methods. To achieve this, we will combine probabilistic evaluation method and partitioning technique which computes the importance of different kinds of meta-information. For that we will have to study both clustering and classification algorithms.

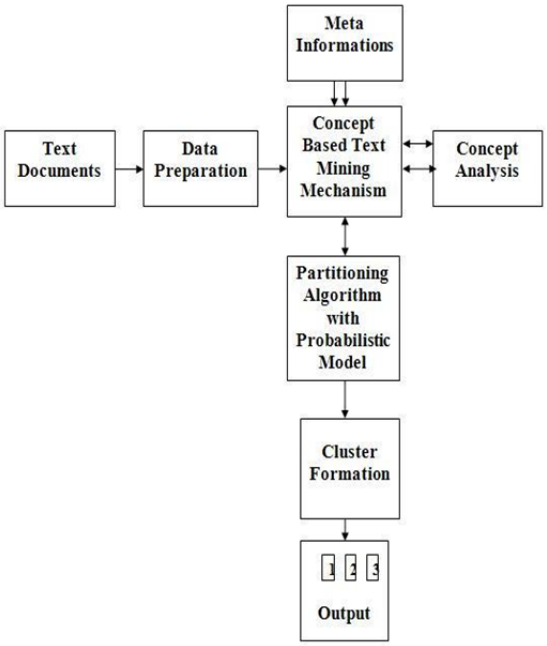
II. RELATED WORK

The study of concept based mining model for information retrieval system [1] [3] [4]. An idea about big data for text mining [2].. A general survey of clustering algorithm may be found in [5]. A comparative study of different clustering methods may be found in [6] [7]. A survey of text clustering methods may be found in [8] [9]. The Problem of text clustering has also been studied in the context of scalability [10] [11] [12] [13]. Co-clustering methods for textual data are proposed in [14] [15]. The studies of classification algorithm are proposed in [13] [14] [15]. Method of text clustering in the context of keyword extraction is discussed in [3] [17]. All of these methods are designed for cases in which text data are combined with other forms of data. Also, some limited work has been done on clustering text in the context of network- based linkage information [4] [24]. For coherence of text clustering and meta-information we will have to refer COATES algorithm which has given in [8] [9] [27]. Finally, the study of clustering and classification techniques and their comparison, we determined by using papers [21] [22] [23]. When the auxiliary information is important and provide effective guidance in creating more coherence clusters for that we refer paper [4] [8].

III. SYSTEM ANALYSIS

A. Proposed System

Number of text documents, contains meta-information which is used in a number of text mining applications. Such meta-information may be of various forms, such as document origin information, the links in the document, web logs which contains user-access behavior, or other text document which are present in the non-textual attributes. Therefore, we need a decent way to perform the mining process, so as to maximize the benefits from using this side information.



.Fig. 1 Outline of proposed work

Modules are as follows:

1) Text Documents:

The document is given as input to the proposed model

2) *Data Preparation:* As in the case of text clustering algorithms, it is assumed that the stemming has been performed and stop-words have been removed in order to improve the discriminatory power of the attributes.

a. Separate sentence
b. Label terms
c. Remove stop words
d. Stem words

3) *Concept Based Text Mining:* The concept-based analysis algorithm describes the process of calculating the ctf, tf and df of the matched concepts in the documents. This strategy begins with processing a new document which has well defined sentence boundaries. Each sentence is semantically labelled. The lengths of the matched concepts and their verb argument structures are stored for the concept-based similarity calculations.

a. Calculating conceptual term frequency
b. Term frequency
c. Document frequency

4) Meta Information: Here the side-information is input, side-information is available along with the text documents may be of different kinds, such as the links in the document, document origin information, non-textual attributes which are enclosed into the text document or user-access behaviour from web logs.

5) *Concept Analysis*: The analysed labelled terms are the concepts that capture the semantic structure of each sentence. Second, to measure the contribution of the concept to the meaning of the sentence we have to use term frequency. Last, the document frequency is used to the number of documents that contains the

6) *Probabilistic Model*: It combines classical partitioning algorithms with probabilistic models in order to create an effective clustering approach.

B. Clustering Algorithms

1) K-means Algorithm: K-means is one of the most popular methods which produce a single clustering. It requires the number of clusters k, to be specified in advance. Initially, k clusters are specified. Then each document in the document set is re-assigned based on the similarity between the document and the k-clusters. Then the k clusters are updated. Then all the documents in the document set are re-assigned. This process is iterated until the k clusters stay unchanged. [22].

2) K-medoid clustering algorithm: When we use K-means algorithm, it is sensitive for objects which having extremely value, which can create problem for data so we have to modify K-means algorithm. Instead of taking mean value, k-medoid used most centrally located object in a clustering as a reference point. By doing this partitioning method can be performed based on principle of minimizing the sum of irregularity between each object and its corresponding reference point. The basic idea behind K-mediod clustering is to search k cluster in n object by first arbitrarily finding a representative object for each cluster, then reaming object is clustered with mediod to which it is most similar [22].

3) *Hierarchical clustering:* Hierarchical clustering constructs a cluster hierarchy, also known as a dendrogram. Every cluster node contains child clusters; sibling clusters Partition the points covered by their common parent. Using such technique, it allows us to explore data on different levels of granularity. Hierarchical clustering generally fall into two types.

a. Agglomerative: This is a bottom up approach. Every observation starts in its own cluster, and merges two or more most appropriate clusters.
b. Divisive: This is a top down Approach. Each observation starts their working in one cluster and splits are performed one by one, as one moves down the hierarchy

In the general case, the complexity of agglomerative clustering is O(n2), which makes them too slow for large volume datasets. The complexity of divisive clustering with an exhaustive search is O(2n)

Advantages of hierarchical clustering:

- It is very flexible for the level of granularity
- It is easier for handling of any forms of distance or similarity
- Applicable to any attribute types

Disadvantages of hierarchical clustering algorithm are as follows:

- Once we design hierarchical algorithms, it becomes hard to reconstruct clusters with the purpose of their improvement

C. Classification Algorithms

Classification is a process that is used for dividing data into different classes according to some constraints. In other word we can say that classification is the process of organizing the data according to different instance. There are so many kinds of classification algorithms include SVM, Decision Tree, Bayesion, and Neural Network Classifier [13].

1) Decision Trees: Decision tree is a classification model in the form of a tree structure that includes root node which is the top most node in the tree, branch node that denote the outcome of the text and leaf node which hold the class label. In a decision tree, each internal node splits the instance space into two or more sub-spaces according to a certain discrete function of the input attributes values. Most frequent cases, every test taken as a single attribute. So that instance space is partitioned according to the attribute's value. But in the case of numeric attributes, it refers to a range. Leaf node is assigned to one class representing the most appropriate target value. Iteratively, the leaf may hold a probability vector indicating the probability of the target attribute having a certain value.. Internal nodes are represented as circles and leaves are denoted as triangles [13].

2) SVM Classifiers: SVM Classifiers try to partition the data space with the use of linear or non-linear delineations between the different classes. The main idea behind such classifiers is to determine the optimal boundaries between the different lasses and use them for the purposes of classification.

3) *Neural Network Classifier:* In data-rich environments neural networks are suitable and mostly used for extracting embedded knowledge in the form of clustering, self-organization, feature evaluation and dimensionality reduction, classification and regression. There are many nice features of neural networks, which make them attractive for data mining. These features include fault tolerance, learning and generalization ability, content addressability, robustness, self-organization and simplicity of basic computations.

IV. CONCLUSIONS

Many application domains contain large amount of text data along with meta-information, such meta-information can be useful in enhancing the quality of the clustering process. The paper discusses the importance of meta-information using clustering and classification techniques. It can be a risky approach to merge meta- information in the mining process because it can add noise in the process, so for improving quality of clustering, we have to remove such noisy data.

Therefore, after studying these techniques we come to the conclusion that a way to design clustering and classification algorithms, which combines classical partitioning algorithm with probabilistic model for effective clustering. So we will get more benefits of meta-information for mining text data.

ACKNOWLEDGMENT

I would like to express my deep sense of gratitude towards Prof.P. D. Lambhate for his motivation and useful suggestions which truly helped me in improving the quality of this paper.

REFERENCES

[1] Shady Shehata, "*An Efficient Concept-Based Mining Model for Enhancing Text Clustering*," IEEE Trans. Knowl. Data Eng., vol. 22, no. 10, Oct. 2010

[2] Xindong Wu and Xingquan Zhu, "Data Mining with Big Data" IEEE Trans. Knowl. Data Eng., vol. 20, no. 1, January 2014.

[3] Magnus Rosell, "Introduction to Information Retrieval and Text Clustering".KTH CSC,Aug 2006.

[4] Shady Shehata and Fakhri Karray, "Enhancing Text Clustering using Concept-based Mining Model". ICDM'06,IEEE, 2006.

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

[6] Michael Steinbach and George Karypis, "A Comparison of Document Clustering Techniques". Technical Report #00-034,2013

[7] G.manimekalai and k.sathiyakumari,"comparative study of fuzzy models in document clustering," in Proc. Text Mining Workshop KDD, 2000, pp. 109–110.

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

[9] C. C. Aggarwal, "social network data analytics". IBM T. J. Watson Research Center, Yorktown Heights, NY 10598, USA : Springer, 2011.

[10] Shi Zhong, "Efficient Streaming Text Clustering" An abbreviated version of some portions of this article appeared in Zhong (2005), published under the IEEE copyright.

[11] Vishal Gupta and Gurpreet S. Lehal, "A Survey of Text Mining Techniques and Applications" in journal of emerging technologies in web intelligence, vol. 1, no. 1, august 2009.

[12] R.Jensi, Dr.G.Wiselin Jiji, "a survey on optimization approaches to text document clustering," International Journal on Computational Sciences & Applications (IJCSA) Vol.3, No.6, December 2013.

[13] Michael W. Berry, "Survey of Text Mining Clustering, Classification, and Retrieval Scanned by Velocity" Springer, 2004.

[14] 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.

[15] *I. Dhillon, S. Mallela, and D. Modha, "Information-theoretic coclustering," in Proc. ACM KDD Conf., New York, NY, USA, 2003,pp. 89–98.*

[16] Raj Kumar, "Classification Algorithms for Data Mining:A Survey," in International Journal of Innovations in Engineering and Technology (IJIET).

[17] Q. Mei, D. Cai, D. Zhang, and C.-X. Zhai, "Topic modeling with network regularization," in Proc. WWW Conf., New York, NY, USA, 2008, pp. 101–110.

[18] G. Salton, an introduction to modern information retrieval. London, U.K.: McGraw Hill, 1983.

[19] H. Schutze and C. Silverstein, "Projections for efficient document clustering," in Proc. ACM SIGIR Conf., New York, NY, USA, 1997, pp. 74–81.

[20] *C. Silverstein and J. Pedersen, "Almost-constant time clustering of arbitrary corpus sets," in Proc. ACM SIGIR Conf., New York, NY, USA, 1997, pp. 60–66.*

[21] Guoqiang Peter Zhang, Neural Networks for Classification: A Survey," in IEEE Trans. Knowl. Data Eng VOL. 30, NO. 4, NOVEMBER 2000.

[22] T. Velmurugan and T. Santhanam, "Computational Complexity between K-Means and K-Medoids Clustering Algorithms for Normal and Uniform Distributions of Data Points," in Journal of Computer Science 6 (3): 363-368, 2010 ISSN 1549-3636 © 2010 Science Publications.