Survey on Mining the Text Data with Meta Information

Mrs. P. Nagajothi,
Assistant Professor,
Department of CSE,
K.S.Rangasamy college of Technology,
Thiruchengode, Namakkal (DT)

S. Vinothini,
M.E-I year,
Department of CSE,
K.S.Rangasamy college of Technology,
Thiruchengode, Namakkal (DT)

Abstract **- Text mining is used to extract hidden information from not-structured or semi-structured data. Aim of this paper is to provide a effective clustering and mining approach with the help of side information.In text mining application every text document has side information.the side information has anormous amout of information, which may be different forms, such as document origin information, links inthe documents, user access behaviour from web logs, other non textual attributes present into the text documents. Such side information can be useful in enhancing the quality of clustering process, or it will add noise to the information.To overcome the problem we use an approach which combines classical partitioning algorithms with probabilistic models so that we can create an effective clustering method. Thus, there should be a right way to do this mining process so that it will make use of side information to maximize their advantages.**

***Keywords – Text mining, information retrieval, information extraction, Text clustering, side information*.**

I. INTRODUCTION

The Text clustering appears in the context of many applications domains such as the web, social networks, and other digital applications. The rapidly enlarging the amount of text data in the context in the context. These huge an online collection has led to an interest in integrate scalable and effective algorithms.

Side information contain ownership of the document and it also contain link of the document .such link provide lot of information.Side data which are present with many web documents may correspond to different kinds of attributes such as provenance or other information about the source of the document.

Side information can be additional feature for raising the quality of the clustering process but it can be dangerous when Meta information is noisy. At that time it can actually degrade the quality of the mining process. Hence an approach is used which carefully ascertains the coherence of the clustering characteristics of the side information with that of the text content. It helps in managing the clustering effects and noisy data.

Text mining is a variation called data mining that tries to find interesting patterns from huge databases. Text databases are rapidly growing due to the increasing amount of information available in electronic form, such as electronic document, electronic publication, mail, World WideWeb, governments, various institutions, industry, business and other institutions information are stored electronically, in the form of text databases.

Data stored in most text databases are semi structured data in that they are neither completely unstructured nor completely structured. Is difficult to formulate effective queries for analyzing and extracting useful information from the data. Users need tools to compare various documents, rank the important of the document and relevance of the documents, or find patterns and trends across multiple documents. Thus, text mining has become popular and essential data mining.

The concept of this paper is to determine a clustering in which the text attributes and side-information provide same indications about the nature of the underlying clusters and at the same time ignore aspects in which conflicting indications are provided

II. TEXT MINING

Text mining process the unstructured information, extracts meaningful numeric indices from the text, and makes the information contained in the text accessible to the various data mining (statistical and machine learning) algorithms. Information can be extracted from the words of the documents, so the words can be determined and also the similarities between words and documents can be determined or how they are related to other variables in the data-mining project. Basically, text mining converts text into numbers that can be included in other analyses such as data mining projects, text clustering etc. Text mining is a text data mining, which refers extracting high-quality information from text.

There are various applications of Text mining like automatic processing of messages and emails. For example, it is possible to "filter" out automatically "junk email" based on certain terms; such messages can automatically be discarded. Such automatic systems for classifying electronic messages can also be useful in applications where messages need to be routed automatically to the most appropriate department. Another application is analyzing warranty or insurance claims, diagnostic interviews**.** In some business domains, the majority of information is collected in textual form.

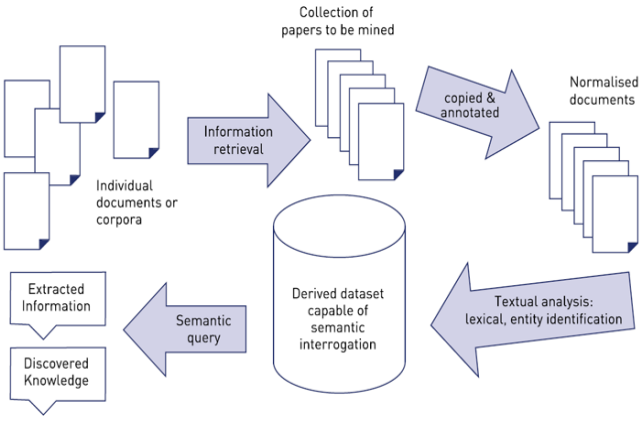

Fig.1. Block diagram for extracting the information

A. *Information extraction*

The general purpose of Knowledge Discovery is to "extract implicit, previously unknown, and potentially useful information from data". Information Extraction IE mainly deals with identifying words or feature terms from within a textual file. Feature terms can be defined as those which are directly related to the domain.

- Stemming

Stemming refers to identifying the root of a certain word. There are basically two types of stemming techniques, one is inflectional and other is derivational. Derivational stemming can create a new word from an existing word, sometimes by simply changing grammatical category (for example, changing a noun to a verb) [Wikipedia]. The type of stemming we were able to implement is called Inflectional Stemming. A commonly used algorithms is the 'Porter's Algorithm' for stemming. When the normalization is confined to regularizing grammatical variants such as singular/plural or past/present, it is referred to inflectional stemming [10].To minimize the effects of inflection and morphological variations of Words (stemming), our approach has pre-processed each word using a provided version of the Porter stemming algorithm with a few changes towards the end in which we have omitted some cases.e.g. apply – applied – applies and print – printing – prints – printed .In both the cases, all words of the first example will be treated as 'apply' and all words of the second example will be treated as 'print'.

B. *Information retrieval*

Information retrieval is considered as an extension to document retrieval where the documents that are returned are processed to condense or extract the particular information sought by the user. Thus document retrieval is followed by a text summarization stage that focuses on the query posed by the user, or an information extraction stage. The modularity of documents may be adjusted so that each individual subsection or paragraph comprises a unit in its own right, in an attempt to focus results on individual nuggets of information rather than lengthy documents.

III.PROPOSED SYSTEM

In many text mining application number of text documents, contains meta-information that is side information. Such side-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. So, we need a method to perform the mining process, so as to maximize the benefits from using this side information.

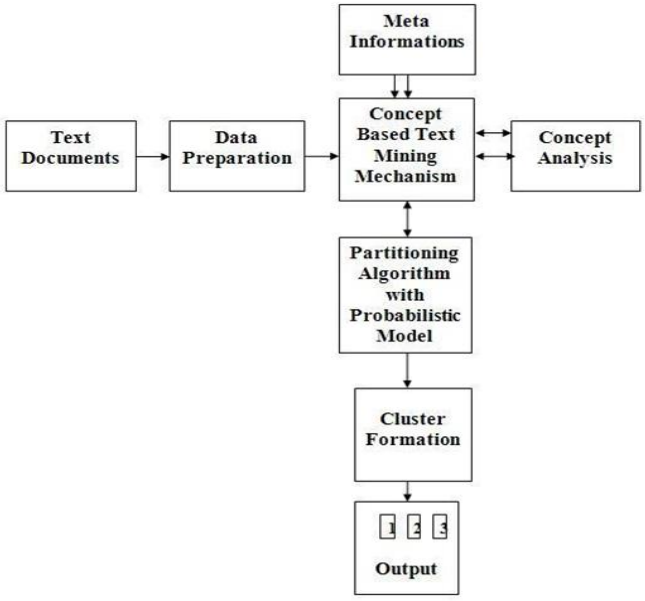

Fig.2. Flow diagram for TextMining

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.

6. *Probabilistic Model*

It combines classical partitioning algorithms with probabilistic models in order to create an effective clustering approach.

IV. 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 reassigned.This process is iterated until the k clusters stay unchanged.

2) *Coates Algorithm*

COATES throughout the paper, which corresponds to the fact that it is a COntent and Auxiliary, attribute based text clustering algorithm. Input to the algorithm is the number of clusters k. As in the case of all text-clustering algorithms, it is assumed that stop-words have been removed, and stemming has been performed in order to improve the discriminatory power of the attributes.

The algorithm requires two phases:

i) Inialization

The standard text clustering approach is used without any side-information. The reason that this algorithm is used, because it is a simple algorithm which can quickly and efficiently provide a reasonable initial starting point. The centroids and the partitioning created by the clusters formed in the first phase provide an initial starting point for the second phase. We note that the first phase is based on text only, and does not use the auxiliary information.

ii) Main Phase

The main phase of the algorithm is executed after the first phase. This phase starts off with these initial groups, and iteratively reconstructs these clusters with the use of both the text content and the auxiliary information. This phase performs alternating iterations which use the text content and auxiliary attribute information in order to improve the quality of the clustering. We call these iterations as content iterations and auxiliary iterations respectively. The combination of the two iterations is referred to as a major iteration. Each major iteration thus contains two minor iterations, corresponding to the auxiliary and text-based methods.

3) *COLT algorithm*

COntent and auxiLiary attribute-based Text classification algorithm.The algorithm uses a supervised clustering approach in order to partition the data into *k* different clusters. This partitioning is then used for the purposes of classification.

The steps used in the training algorithm are as follows:

• *Feature Selection*

The use feature selection to remove those attributes, which are not related to the class label. This is performed both for the text attributes and the auxiliary attributes.

• *Initialization*

Use a supervised kmeans approach in order to perform the initialization, with the use of purely text content. The main difference between a supervised *k*-means initialization and an unsupervised initialization is that the class memberships of the records in each cluster are pure for the case of supervised initialization. Thus, the *k*-means clustering algorithm is modified, so that each cluster only contains records of a particular class.

• *Cluster-Training Model Construction*

The method is to a combination of the text and side-information is used for the purposes of creating a cluster-based model. As in the case of initialization, the purity of the clusters in maintained.

4) Classification Algorithms

Classification is a process that is used for dividing data into different classes according to some constraints. There are so many kinds of classification algorithms include SVM, Decision Tree, Bayesion, and Neural Network Classifier [2].

a) 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. Internal nodes are represented as circles and leaves are denoted as triangles [2].

b) 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.

c) 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.

V. CONCLUSION

The papers are discussed for mining text data with making use of side information. Side information may be presented in many forms of text database which are used to enhance the clustering process. It can be a risky approach to merge meta-information in the mining process because it can add noise in the process.Therefore the way to design clustering and classification algorithms, is that it combines classical partitioning algorithm with probabilistic model for effective clustering. So we will get more benefits of metainformation for mining text data. This general method is used to design both clustering and classification algorithms.

REFERENCE

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

[3] Hassan, A., Jones, R., and Klinkner, K., "Beyond dcg: user behavior as a predictor of a successful search," ser. WSDM'10, 2010, pp. 221–230.
Teevan, J., Adar, E. Jones, R. and Potts M., "Information reretrieval: repeat queries in yahoo's logs," ser. SIGIR '07, New York, NY, USA, 2007, pp. 151–158.

[4] White, R., Bailey, P. and Chen, L., "Predicting user interests from contextual information," ser. SIGIR '09, 2009, pp. 363–370.

[5] Radlinski, F. and Craswell, N., "Comparing the sensitivity of information retrieval metrics," ser. SIGIR '10, 2010, pp. 667–674.

[6] Shen, X., Tan, B. and Zhai, ChengXiang, "Context-sensitive information retrieval using implicit feedback," ser. SIGIR '05, 2005, pp. 43–50.

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

[8] M. Franz, T. Ward, J. S. McCarley, and W. J. Zhu, "Unsupervised and supervised clustering for topic tracking," in Proc. *ACM SIGIR* Conf., New York, NY, USA, 2001, pp. 310–317.

[9] G. P. C. Fung, J. X. Yu, and H. Lu, "Classifying text streams in the presence of concept drifts," in Proc. *PAKDD* Conf., Sydney, NSW, Australia, 2004, pp. 373–383.

[10] H. Frigui and O. Nasraoui, "Simultaneous clustering and dynamic keyword weighting for text documents," in Survey of Text Mining, M. Berry, Ed. New York, NY, USA: Springer, 2004,pp. 45–70.

[11] S. Guha, R. Rastogi, and K. Shim, "CURE: An efficient clustering algorithm for large databases," in Proc. *ACM SIGMOD* Conf., New York, NY, USA, 1998, pp. 73–84.

[12] Q. He, K. Chang, E.-P. Lim, and J. Zhang, "Bursty feature representation for clustering text streams," in Proc. *SDM* Conf., 2007, pp. 491–496.

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

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

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

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

[17] C. C. Aggarwal and P. S. Yu, "A framework for clustering massive text and categorical data streams," in *Proc. SIAM* Conf. Data Mining, 2006, pp. 477–481.

[18] C. C. Aggarwal and P. S. Yu, "On text clustering with side information," in *Proc. IEEE ICDE* Conf., Washington, DC, USA, 2012.

[19] R. Angelova and S. Siersdorfer, "A neighborhood-based approach for clustering of linked document collections," in Proc. *CIKM* Conf., New York, NY, USA, 2006, pp. 778–779.