IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

*International Journal of Advanced Research in Computer and Communication Engineering*
*Vol. 4, Issue 11, November 2015*

# A Review on Categorization of Text Data Using Side Information

**Sandeep Jadhav[1], Dr. K. V. Metre[2]**

PG Student, Computer Engineering Department, MET'S IOE, Nashik, India[1]

Asst. Professor, Computer Engineering Department, MET'S IOE, Nashik, India[2]

**Abstract**: In today's digital environment, text databases are rapidly increases due to use of internet and communication mediums. Different text mining techniques are used for knowledge discovery and Information retrieval. Text data contains the side information along with the text data. Side information may be the metadata associated with text data like author, co-author or citation network, document provenance information, web links or other kind of data which provide more insights about the text documents. Such side information contains tremendous amount of information for the clustering purpose. Using such side information in the categorization process provides more refine clustered data. But sometimes side information may be noisy and results in wrong categorization which decreases the quality of clustering process. Therefore, a new approach for mining of text data using side information is suggested, which combines partitioning approach with probabilistic estimation model for the mining of text data along with the side information.

**Keywords**: Text data mining, categorization, side information, clustering.

## I. Introduction

Text mining is the important field in knowledge discovery and information retrieval due to rapidly growing hardware and software technologies which led into availability of large amount of text database. To handle such huge amount of text databases in various electronic forms, efficient mining techniques are required. Text mining is nothing but the pattern discover or knowledge discovery from text documents. Different text mining techniques are proposed by many authors [1]. In text document, side information may be present. Side information is nothing but the attributes or term or features which provide more insights about the text data. Such side attribute may contain metadata, web links, user access behavior, citation network or other attributes which gives more detail view of the text documents. For example:

1) User access behavior of documents stored in web log provides additional information about text data which provides correlation in contents.
2) Links present in text documents can also be treated as the side attribute which provide additional information by the correlations among the text documents. Such correlation cannot be handled by only pure data.
3) Meta-data associated with text data contains author-coauthor relationship, document provenance information or the origin of documents, citation network, and timestamp associated with origin of documents.
4) Other side attributes like location, user tags, data ownership or some temporal information provides more information about the underline text data.

Such auxiliary attributes are used in the text mining process which gives more refine results. While using such side information in text mining, some problems occurs which reduce the quality of mining result. Side attribute may be contains noise in text data. Such noisy attribute gives wrong results which affect the purity of text categorization process. Hence such approach of using side information in text mining either increase quality of categorization or decrease the quality of results.

The basic idea behind the system is to find the clusters in which side information and content only data provide similar hints about the behaviour of underlying data and ignore that side attribute which conflict. For this purpose, the method which combines the iterative partitioning technique with the probabilistic model which gives coherence of side information with the pure text documents in categorization process. The noise in the side attribute is eliminated by the cluster membership probability by the Bayes Independence approximation method [4]. Here, partitioning approach means iteratively find the clusters associated with closest seed centroid and update that seed by new centroid i.e. this is for pure data only. After this, probabilistic model is used for finding the importance of side attribute with pure text cluster.

## II. Literature Review

Different text data mining techniques has been studied by various authors. The categorical data stream clustering problem also has a number of applications to the problems of customer segmentation and real time trend analysis. A Framework for Clustering Massive Text and Categorical Data Streams gives an online approach with use of a statistical summarization methodology. Statistical summary data is stored in regular interval and from this summary information characteristics of different clusters are diagnosed. It provides a framework in which carefully chosen statistical summary data is stored at regular intervals. This results in a system in which it is possible to

DOI 10.17148/IJARCCE.2015.41196

diagnose different characteristics of the clusters in an effective way. In the context of a data stream, such a methodology seems quite convenient, since a fast data stream cannot be repeatedly processed in order to answer different kinds of queries. The methodology used is same as to online analytically processing algorithm in which summary information is created for the purpose of repeated query [2].

J. Chang and D. Blei uses Relational Topic Model (RTM) for document networks which provides a model for documents and the links between them. For each pair of documents, the RTM models their link as a binary random variable that is conditioned on their contents. The model can be used to summarize a network of documents, predict links between them, and predict words within them. It also provides efficient inference and learning algorithms based on variational methods and evaluate the predictive performance of the RTM for large networks of scientific abstracts and web documents [3].

A Neighbourhood Based Approach for Clustering of Linked Document Collections proposed the problem of automatically structuring linked document collections by using clustering. In contrast to traditional clustering, clustering problem in the light of available link structure information for the data set (e.g., hyperlinks among web documents or co-authorship among bibliographic data entries). It uses an iterative relaxation of cluster assignments, and can be built on top of any clustering algorithm. Content based clustering along with the link structure used to form the final cluster. Document d represent the vertex in graph G. Hyperlinks are represented as edge between two vertexes along with its weight [4].

C. Aggarwal and S. Gates used partial supervision methods for text categorization which gives merits of building text categorization systems by using supervised clustering techniques. Traditional approaches for document classification on a predefined set of classes are often unable to provide sufficient accuracy because of the difficulty of fitting a manually categorized collection of documents in a given classification model. This is especially the case for heterogeneous collections of Web documents which have varying styles, vocabulary, and authorship. Hence, use of clustering in order to create the set of categories and its use for classification of documents. Completely unsupervised clustering has the disadvantage that it has difficulty in isolating sufficiently fine-grained classes of documents relating to a coherent subject matter. Hence, information from a pre existing taxonomy in order to supervise the creation of a set of related clusters, though with some freedom in defining and creating the classes. The advantage of using partially supervised clustering is that it is possible to have some control over the range of subjects that one would like the categorization system to address, but with a precise mathematical definition of how each category is defined. An extremely effective way then to categorize documents is to use this a priori knowledge of the definition of each category. The preexisting taxonomy or data is used to supervise the cluster creation by some freedom in defining and creation of the classes [5].

The Simple Bayesian Classifier under Zero-One Loss provides simple Bayesian classifier which is known to be optimal when attributes are independent given the class, but the question of whether other sufficient conditions for its optimality exist has so far not been explored. Although the Bayesian classifier’s probability estimates are only optimal under quadratic loss if the independence assumption holds, the classifier itself can be optimal under zero-one loss (misclassification rate) even when this assumption is violated by a wide margin. The region of quadratic-loss optimality of the Bayesian classifier is in fact a second-order infinitesimal fraction of the region of zero-one optimality. This implies that the Bayesian classifier has a much greater range of applicability than previously thought [7].

A survey of text classification algorithms provide detailed survey on classification [6]. The techniques studied earlier are based on only pure text data for text mining and they have not used any kind of side information with pure data for clustering. So, the system which works on side information along with pure data to find the coherence of text data for the categorization purpose is suggested.

## III. SYSTEM ARCHITECTURE

The system uses side information with the pure text data for the mining process. The input to the system is set of text documents. Input data is pre-processed for the removal of noisy data. Pre-processing techniques like stop word removal, stemming, etc. are used. Cosine similarity measure is used for to determine the closest cluster to centroid. For clustering purpose, COATES algorithm is used which is “COntent and Auxiliary based Text CluStering Algorithm”. This algorithm starts with some initial clusters as input. It has two iterations, namely content based iteration and auxiliary based iteration. Content based iteration takes pure text data while auxiliary based iteration takes side information along with content based data. In first content based iteration cosine similarity measure is used for finding closet clusters to seed centroid of initial clusters. By adding the new data points in clusters, centroids are updated.

In Auxiliary based iteration, it takes input from the previous content based iteration. For noise removal, Gini index is calculated for each side feature and feature which have value lower than mean standard deviation are taken as non-discriminatory. Only for discriminatory attributes (Ri) in the document Ti the posterior probability is calculated using Bayes Independence Approximation. The posterior probability $P^n$ (Ti ∈ Cj|Ri)i.e. probability of finding document Ti in the cluster Cj given that the value of auxiliary attribute Ri is 1.

In this way, iteratively content and auxiliary based iterations are carried out by calculating prior and posterior probability to find the more précised clusters.

For categorization purpose COLT algorithm is used. COLT stands for “COntent and auxiLiary attribute based Text Clustering algorithm”. Firstly, feature selection take place for removal of those attribute which are not related to class label. It uses supervised k mean algorithm for

DOI 10.17148/IJARCCE.2015.41196

IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

*International Journal of Advanced Research in Computer and Communication Engineering*
*Vol. 4, Issue 11, November 2015*

finding the supervised clusters so that each cluster contain records of particular class label only. It again uses cosine similarity for similarity measure and gini index calculation for noise attribute removal. In training model it uses both text content data and side attributes. It iteratively finds supervise clusters using content and auxiliary data by finding the prior and posterior probability. Fig. 1 shows COATES and COLT algorithms for categorization of text documents. Clusters are associated with the class label and label which has largest presence of such clusters is treated as class label for that test instance.

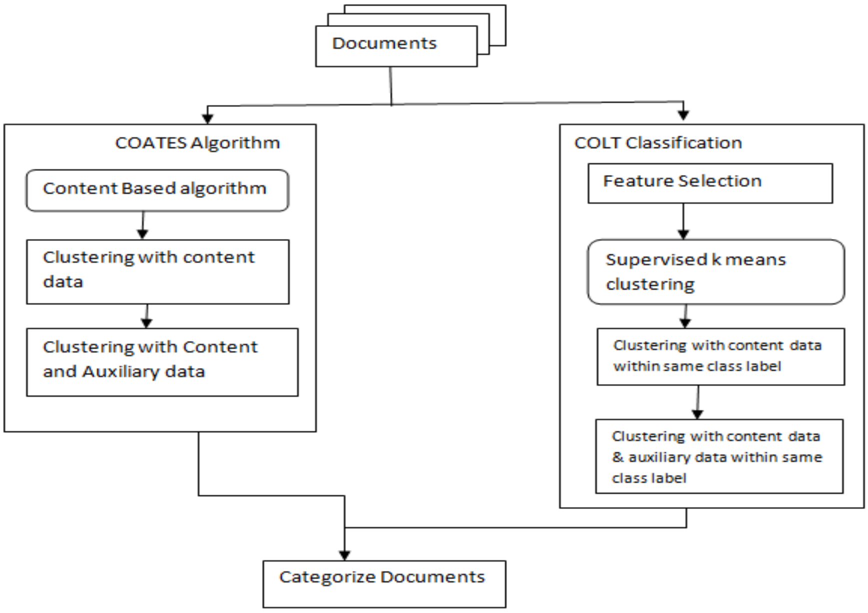
Documents
COATES Algorithm
Content Based algorithm
Clustering with content data
Clustering with Content and Auxiliary data
COLT Classification
Feature Selection
Supervised k means clustering
Clustering with content data within same class label
Clustering with content data & auxiliary data within same class label
Categorize Documents

Fig. 1 : Proposed System Architecture

## IV. CONCLUSION

Text data contain large amount of side information along with text content and it can be used in the process of text categorization which improve the efficiency of categorize data. In order to make use of such side information, clustering methods combine an iterative partitioning technique with a probabilistic estimation process which computes the importance different kind of side information with the content data. This general approach is used in order to design both clustering and classification algorithms.

## REFERENCES

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

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

[3] Chang, D. Blei, "Relational topic model for document network,"in Proc. ASIASIS, Clearwater, FL, USA 2009, pp. 81-88.

[4] R. Angelova, S. Siersdorfer,"A neighborhood- based approach for clustering of linked documents collection", in Proc. CIKM Conf. New York, NY, USA, 2006.

[5] C. C. Aggarwal, S. C. Gates,"On using partial supervision for text categorization," IEEE Trans. Know. Data Eng. Vol. 16, no.2, pp. 245-255, Feb. 2004.

[6] C.C Aggarwal, C. X. Zhai ,"A survey of text classification algorithm", in Mining Text Data, New York, USA: Springer, 2012

[7] P. Domingos, M. J. Pazzani,"On the optimality of simple baysian classifier under zero- one loss", Match. Learn, vol.29, no.2-3, pp. 103-130, 1997.

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

DOI 10.17148/IJARCCE.2015.41196