An Effective Use of Meta Information for Text Mining

Mr. Nitin J.Ghatge, Prof. Poonam D. Lambhate

Abstract— **Aim of this paper is immersed in effective clustering and mining approach with help of side information. Number of text mining applications, having side-information with them. This information may be of various forms, such as provenance information of the documents, the links in the document, web logs which contains user-access behavior, or other text document which are embedded into the non-textual attributes. These attributes may contain a lot of information for clustering purposes. However, the concerned importance of this side-information may be hard to count, especially when some of the information is noisy. In such cases, it can be hazardous to merge side-information into the mining process, because it can either enhance the quality of the representation or can add noise in the system. Therefore, literature study suggests way to design efficient algorithm which combines classical partitioning algorithm with probabilistic model for effective clustering approach, so as to maximize the benefits from using side information**

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

I. Introduction

The problem of text clustering arises in the context of many application domains such as the web, social networks, and other digital collections. The rapidly increasing amounts of text data in the context of these large online collections has led to an interest in creating scalable and effective mining algorithms. A tremendous amount of work has been done in recent years on the problem of clustering in text collections in the database and information retrieval communities. However, this work is primarily designed for the problem of pure text clustering, in the absence of other kinds of attributes. In many application domains, a tremendous amount of side information is also associated along with the documents. This is because text documents typically occur in the context of a variety of applications in which there may be a large amount of other kinds of database attributes or met 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: activity statistics, accessed files, path through the site, information about referring pages, search engines, browsers, operating systems 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 may be identified among various types of objects, include organizations, people 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.

A. *Contributions:*

For considering Meta information

- Using of side-information enhance the quality of text clustering and classification, while maintaining a high level of efficiency.
- Design a model to identify noisy information.
- Combines classical partitioning algorithms with probabilistic models in order to create an effective clustering approach.

II. Related Work

The study of concept based mining model for information retrieval system [1] [3] [4], A thought regarding 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. Paper Lyout

The rest of the paper is organized as follows: we go over formulate the problem, and define our privacy goal in Section IV. It highlights the key challenge in achieving our privacy goal, and presents the ON THE USE OF SIDE INFORMATION FOR MINING TEXT DATA 2014 that leads to our solution. In Section V, we formally present our solution, and proved that it achieves our privacy goal. In Section VI, we show the experimental values and analysis.

IV. Problem Formation

A. Problem Statement

Many application domains contain large amount of text data along with meta-information, while such meta-information can be useful in enhancing the quality of the clustering process. This paper discusses the importance of meta-information using clustering and classification techniques. But 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.

B. Proposed System

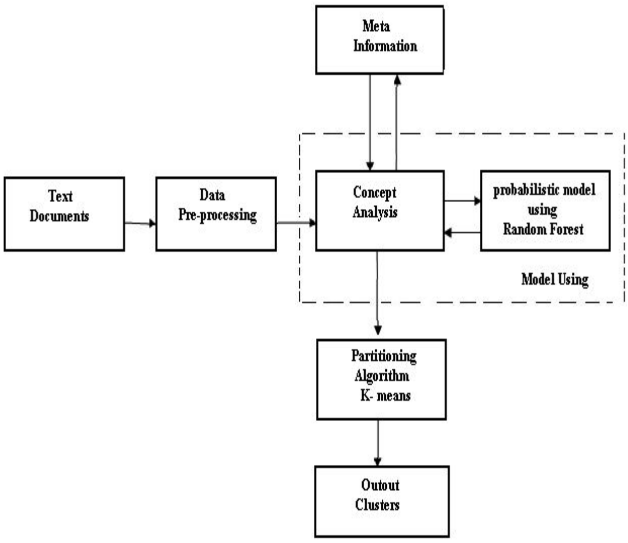


Fig. 1. Outline of proposed work

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 stop-words have been removed, and stemming has been performed in order to improve the discriminatory power of the attributes.

- Separate sentence
- Label terms
- Remove stop words
- 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 labeled. The lengths of the matched concepts and their verb argument structures are stored for the concept-based similarity calculations.

- Calculating conceptual term frequency
- Term frequency
- Document frequency

4) *Side 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 behavior from web logs.

5) *Concept Analysis:* The analyzed labeled 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.

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

V. Solution To Problem Statement

A. Mathematical Model

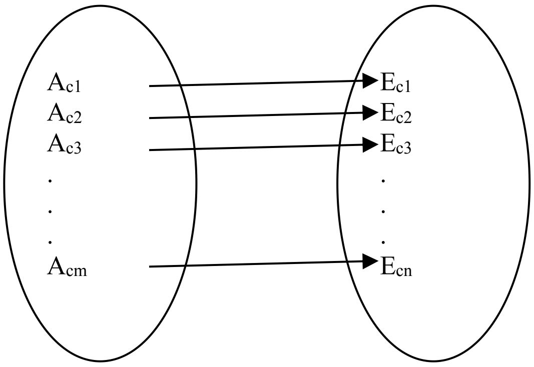
S = {I, E, X, Y, Z, DD, NDD, Fs}

I - Initial state
E- End state
X- Set of input
X= {T, A}
T- Set of text data
$T = \{T_1, T_2, T_3 \ldots T_n\}$
M- Set meta-data
$M = \{M_{11,} M_{12}, M_{13} \ldots\ldots M_{nn}\}$
C- Set of concepts
$C = \{C_1, C_2, C_3 \ldots. C_n\}$
Ac- Set of concepts with meta-data
$A_c = \{A_{c1}, A_{c2}, A_{c3} \ldots. A_{cn}\}$
Y- Set of intermediate output
$Y = \{C, A_c\}$
Z- Set of effective clusters
$Z= \{E_{c1}, E_{c2}, E_{c3} \ldots.. E_{cn}\}$

Sr. No.	Key notations.	
	Notation	*Definition*
1	T	Text Data
2	A	Auxiliary Data (Meta Data)
3	C	Concepts
4	Ac	Concepts with Auxiliary Data
5	Y	Intermediate Output
6	Z	Effective Clusters

Fs- Set of functions to be performed on input
Fs- $\{F_1, F_2, F_3, F_4\}$
F_1: Function to accept text input
F_2: $T \rightarrow C$
F_3: $A, C \rightarrow A_c$
F_4: $A_c \rightarrow E_c$

1) Mapping Function:



B. Algorithm Used

To achieve effective clustering we have divided our approach into two phases first phase will be initial phase for probabilistic model using random forest algorithm, where we will get known number of classes, by using this classes we will start second phase which is main phase, implemented by using k means algorithms

1) *K-means :* This outline of the algorithm assumes two clusters, and each individual's scores include two variables. Adding more clusters is as easy as adding another step like Step 4 or Step 5. Adding another variable for each individual is as easy as adding calculations within the type of step like Step 4 or Step 5.

a) Step 1:

- Choose the number of clusters.

b) Step 2:

- Set the initial partition, and the initial mean vectors for each cluster.

c) Step 3:

- For each remaining individual...

d) *Step 4:*

Get averages for comparison to the Cluster 1:

- Add individuals A value to the sum of A values of the individuals in Cluster 1, then divide by the total number of scores that were summed.
- Add individual's B value to the sum of B values of the individuals in Cluster 1, then divide by the total number of scores that were summed.

e) *Step 5:*

Get averages for comparison to the Cluster 2:

- Add individuals A value to the sum of A values of the individuals in Cluster 2, then divide by the total number of scores that were summed.
- Add individual's B value to the sum of B values of the individuals in Cluster 2, then divide by the total number of scores that were summed.

f) *Step 6:*

- If the averages found in Step 4 are closer to the mean values of Cluster 1, then this individual belongs to Cluster 1, and the averages found now become the new mean vectors for Cluster 1.
- If closer to Cluster 2, then it goes to Cluster 2, along with the averages as new mean vectors..

g) Step 7:

- If there are more individual's to process, continue again with Step 4. Otherwise go to Step 8.

h) Step 8:

- Now compare each individual’s distance to its own cluster's mean vector, and to that of the opposite cluster. The distance to its cluster's mean vector should be smaller than it distance to the other vector. If not, relocate the individual to the opposite cluster.

i) *Step 9:*

- If any relocation occurred in Step 8, the algorithm must continue again with Step 3, using all individuals and the new mean vectors.
- If no relocations occurred, stop. Clustering is complete.
- Again, in case the algorithm never settles on a final solution, it may be a good idea to implement a maximum number of iterations check.

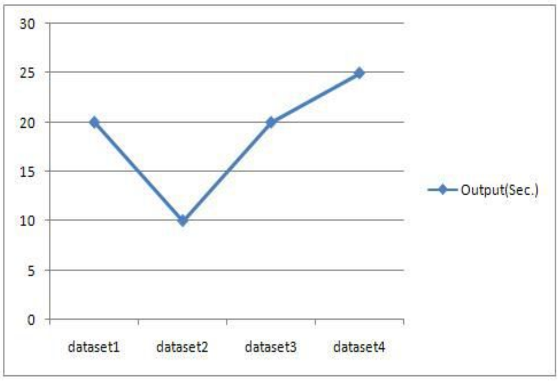
2) *Random Forest:* Random forest is a classification ensemble algorithm developed by Leo Breiman that uses multiple binary decision trees. Each of the classification trees is built using a sample of data and at each node a randomly chosen set of variables is considered for the best split. Random forest has become a major data analysis tool. It has been applied to large-scale tissue microarray data and genome-wide association studies for complex diseases node.

Each tree is constructed using the following algorithm:

- Let the number of training cases be N, and the number of variables in the classifier be M.
- The number of m input variables to be used to determine the decision at a node of the tree; m should be much less than M.
- Choose a training set for this tree by choosing N times with replacement from all N available training cases (i.e. take a bootstrap sample). Use the rest of the cases to estimate the error of the tree, by predicting their classes.
- For each node of the tree, randomly choose m variables on which to base the decision at that node. Calculate the best split based on these m variables in the training set.

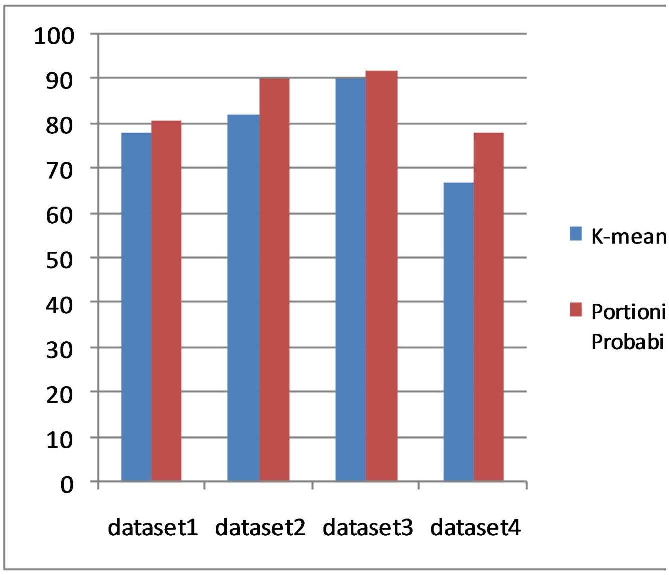
- Each tree is fully grown and not pruned (as may be done in constructing a normal tree classifier).

VI. Experimental Analysis



Time Complexity	
Input	OutPut(sec)
DataSet1	20
DataSet2	10
DataSet3	20
DataSet4	25

Time Complexity		
Input	K-means(sec)	Portioning + Probabilistic
DataSet1	78	81
DataSet2	82	90
DataSet3	90	92
DataSet4	67	78



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] Charu C. Aggarwal, “On the Use of Side Information for Mining Text “ Data,IEEE Trans. Knowl.Data Eng.vol.26,no.6,JUNE 2014,pp.14151420.

[2] Xindong Wu and Xingquan Zhul,” Data Mining with Big Data”,IEEE Trans. Knowl.Data Eng.vol.20,no.1, January 2014,,pp.97-107.

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

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

[5] Shady Shehata and Fakhri Karray, Enhancing Text Clustering using Concept-based Mining Mode ,ICDM06,IEEE,2006.

[6] Yung-Shen Lin, Jung-Yi Jiang, Similarity Measure for Text Classification and Clusterin ,IEEE Trans. Knowl. Data Eng.vol.26,no.7,JULY 2014,pp.1575-1590.

[7] C. C. Aggarwal, S. C. Gates, and P. S. Yu, On using partial supervision for text categorization ,IEEE Trans. Knowl. Data Eng.vol.16 ,no.2, Feb 2004.

[8] C. C. Aggarwal and P. S. Yu, On text clustering with side information ,Proc. IEEE ICDE Conf., Washington, DCUSA,2012.

[9] Michael Steinbach and George Karypis, A Comparison of Document Clustering Techniques ,Technical Report 00-034,2013 Conf.New York, NY, USA, 2006.

[10] G.manimekalai and k.sathiyakumari, comparative study of fuzzy models in document clustering ,Proc. Text Mining Workshop KDD 2000.

[11] Shi Zhong, Efficient Streaming Text Clustering ,IEEE Trans. Knowl. Data Eng.,published under the IEEE copyright,2005.

[12] Vishal Gupta and Gurpreet S. Lehal, A Survey of Text Mining Techniques and Applications ,ETWI, vol.1,no.1 AUGUST,2009

[13] R.Jensi, Dr.G.Wiselin Jiji, A survey on optimization approaches to text document clustering ,IJCSIT, vol.3, no.6,December 2013.

[14] Margaret H. Dunham, Data Mining Introductory and advanced Topics ,2006.

[15] I. Dhillon, S. Mallela, and D. Modha, Information-theoretic coclusterin ,Proc. ACM KDD Conf. New York, NY, USA, 2003.

[16] S. Guha, R. Rastogi, and K. Shim, ROCK: A robust clustering algorithm for categorical attributes ,f. Syst., vol. 25, no. 5 pp. 345366,

[17] XindongWu Vipin Kumar, Top 10 algorithms in data mining ,KnowlInf Syst (2008).14:137.

[18] Charu C. Aggarwa, On the Use of Side Information for Mining Text Data ,IEEE Trans. Knowl. Data Eng.vol.26,no.6,JUNE 2014. ,pp.14151420.

***Mr. Nitin J.Ghatge**, Computer Engineering Department, JSPM's, JSCOE, Hadapsar, Pune, India , 7588619822*

***Prof. Poonam D. Lambhate.** Computer Engineering Department, JSPM's,JSCOE,Hadapsar,Pune,India.*