International Journal of Modern Trends in Engineering and Research
www.ijmter.com
e-ISSN No.:2349-9745, Date: 2-4 July, 2015

Meta Information Based Text Clustering and Classification with the Use of COATES and COLT Algorithms.

Mrunal V. Upasani[1], Rucha C. Samant[2]
[1]*Computer Engg. Department, GES's R. H. Sapat College of Engineering, Management Studies and Research Nashik, Affiliated to Savitribai Phule Pune University, vrupasani2013@gmail.com*
[2]*Computer Engg. Department, GES's R. H. Sapat College of Engineering, Management Studies and Research Nashik, Affiliated to Savitribai Phule Pune University, ruchasamant25@gmail.com*

Abstract- Large amount of meta-information is available along with the text documents in many text mining applications. Such meta-information may be of different kinds, such as links in the document, user-access behaviour from web logs which can be useful for data mining. Tremendous amount of information can be found in unstructured attributes for clustering purposes. Therefore, the paper will use an approach which carefully ascertains the coherence of the clustering characteristics of the meta information with that of the text content. For improving the quality of the clustering both the text data and meta information will be helpful. In this paper, the design of an algorithm which combines classical partitioning algorithms with probabilistic models in order to create an effective clustering approach using meta information present in document will be perform. Then it shows how to extend the clustering approach to the classification problem. COATES and COLT algorithm for clustering and classification of text data along with the meta information are presented in this paper and it shows the advantages of using such an approach.
Keywords- Classification; clustering; data mining; meta information; text mining

I. INTRODUCTION

The problem of text clustering arises in the many application domains such as the social networks, web and other digital collections. The increase in amount of text data in the context of these large online collections has led to an interest in creating scalable and effective mining algorithms.
The set of disjoint classes called clusters are created in the process of clustering. Objects which are in the same cluster have similarity among themselves and dissimilarity to the objects belonging to other clusters. Clustering is having very important role in the text domain, where the objects which is to be clusters are of different sizes like documents, paragraphs, sentences or terms.
Many application domains contains large amount of meta information, which is also associated along with the documents. Text documents typically occur in the variety of applications in which there may be a large amount of other kinds of meta-information which may be useful to the clustering process.
The access behaviour of user may be captured in the form of web logs. For each document, the meta-information means the browsing behaviour of the different users. For enhancing the quality of the mining process which is more meaningful to the user these logs can be used. Many text documents contain links in between them, which can also be treated as meta-information attributes. Lot of useful information is available in these links which can be used for mining purposes. Web documents have meta information associated with them which correspond to different kinds of other information like ownership, location, or even temporal information about the origin of the document. This all are the examples of meta information related to the documents.
This type of meta-information can be useful in improving the quality of the clustering process [1]. The primary goal of this paper is to study the clustering of data in which auxiliary information is available with text. Such scenarios are very common in a wide variety of data domains. Therefore,

the paper extends the clustering approach to the problem of classification, which provides superior results because of the incorporation of meta information
Goal of this paper is to show the advantages of using meta-information extend beyond a pure clustering task, which can provide competitive advantages for a wider variety of problem scenarios.

II. LITERATURE SURVEY

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. The Survey of Text Clustering Algorithms is studied in [2] [3].
Hierarchical clustering creates the cluster hierarchy for which the leaf nodes correspond to individual documents, and the internal nodes correspond to the merged groups of clusters. A hierarchical clustering algorithm called CURE that is more robust to outliers, and identifies clusters having non-spherical shapes and wide variances in size is given in [4]. Another hierarchical clustering algorithm, Robust Clustering Algorithm for Categorical Attributes for data with Boolean and categorical attributes is studied in [5]. [6] Presents the hierarchical data clustering method BIRCH (Balanced Iterative Reducing and Clustering using Hierarchies) and it demonstrates that it is especially suitable for very large databases.
In particular, K-means uses the mean or median point of a group of points [2]. The simplest form of the k-means approach is to start with a set of k seeds from the original corpus, and assign documents to these seeds on the basis of closest similarity. In the next iteration, the centroid of the assigned points to each seed is used to replace the seed in the last iteration. In other words, the new seed is defined, so that it is a better central point for this cluster. Simultaneous Clustering and Dynamic Keyword Weighting for Text Documents takes place in [7]. It uses the approach to extend K-means algorithm, that in addition to partitioning the dataset into a given number of clusters, also finds the optimal set of feature weights for each clusters. [8] Combines an efficient online spherical k-means (OSKM) algorithm with an existing scalable clustering strategy to achieve fast and adaptive clustering of text streams.
The third type of document clustering is the hybrid clustering technique [2] [9]. Scatter-Gather clusters the whole collection to get groups of documents that the user can select or gather.
However, all of these methods are designed for the case of pure text data, and do not work for cases in which the text-data is combined with other forms of data. Some limited work has been done on clustering text in the context of network-based linkage information like graph mining and algorithms of graph mining in [10] [11] [12].
A wide variety of techniques have been designed for text classification in [13]. Probabilistic classifiers are designed to use an implicit mixture model for generation of the underlying documents. Decision Tree Classifiers performs the division of the data recursively. SVM is to determine separators in the search space which can best separate the different classes
All this work is not applicable to the case of general meta information attributes. The first approach of using other kinds of attributes in conjunction with text clustering was studied in [14]. This approach is especially useful, when the auxiliary information is highly informative, and provides effective guidance in creating more coherent clusters. The proposed work extends the clustering method to the problem of text classification.

III. DETAILS OF PROPOSED WORK

3.1 Problem Definition

Given a corpus S of documents denoted by $T_1..T_n$ and a set of auxiliary variables X_i associated with document T_i, determine a clustering of the documents into k clusters which are denoted by $C_1..C_k$ based on both the text content and the auxiliary variables. Generated clusters will be further

classified into number of classes with labels.

3.2 Proposed System Architecture

The Architecture of the proposed system is shown in the Figure 1 .The first phase of the proposed work is the preprocessing of the documents with both the text content and the meta content too. The Preprocessing phase includes the removal of stop words, removal of special characters etc.

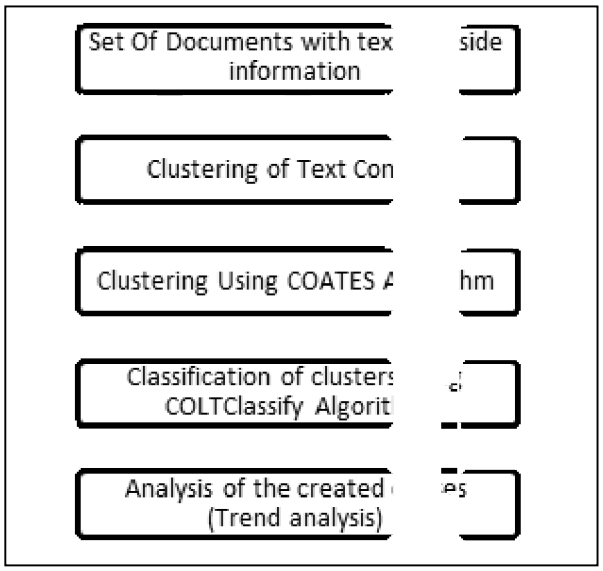

Figure 1. Block Diagram of Proposed System

For the creation of the primary clusters which will be used by the COATES and COLT algorithm, the system will use the clustering algorithms like KMeans clustering. The result will be given as input to the clustering with the meta information.

3.2.1 Clustering of Text Content

For the clustering of the text content the KMeans clustering algorithm is used. The result of this will be given as input to the next phases of the system.

3.2.2 Clustering Using COATES Algorithm

The algorithm which will be used for clustering of meta information is COATES Algorithm. This corresponds to the fact that it is a COntent and Auxiliary attribute based TExt cluStering algorithm [1].

The algorithm works in 2 steps.

3.2.2.1 Initialization

It is a lightweight initialization phase in which a standard text clustering approach like KMeans or hierarchical algorithm is used without any meta-information. The partitioning and the centroids created by the clusters formed in the first phase provide an initial starting point for the second phase. The first phase is based on text information only, not the meta information.

3.2.2.2 Main Phase

This phase iteratively reconstructs the clusters with the use of both the text content and the auxiliary information means the meta information. Alternating iterations which use the text content and auxiliary attribute information in order to improve the quality of the clustering are performed in this step. The iterations are content iterations and auxiliary iterations respectively. The combination of these two is referred as a major iteration.

3.2.3 Classification of clusters using COLTClassify Algorithm

For the purpose of classification, proposed work uses COLT algorithm, which refers to the fact that it is a COntent and auxiLiary attribute-based Text classification algorithm. This algorithm uses a supervised clustering approach in order to partition the data into different clusters. This partitioning

is then used for the purposes of classification [1].
The algorithm works in 3 steps.

3.2.3.1 Feature Selection

In the first step, system uses feature selection to remove the attributes, which are not related to the class label. It is performed both for the text attributes and the auxiliary attributes.

3.2.3.2 Initialization

In this step, system uses a supervised clustering approach in order to perform the initialization, with the use of purely text content. The class memberships of the records in each cluster are pure for the case of supervised initialization.

3.2.3.3 Cluster-Training Model Construction

In this phase, a combination of the text and meta-information will be used for the purposes of creating a cluster-based model.
The supervised clusters provide an effective summary of the data which can be used for classification purposes.

3.2.4 Analysis of the created classes (Trend analysis)

After the creation of the classes from the Coltclassify algorithm. The analysis of these classes will be done on the basis of the temporal information of the documents. And further the trend analysis of the created classes will be done accordingly, the statistical representations of the results will be made.

IV. RESULTS

The CORA Dataset will be used for the implementation of this system. The Cora data set contains scientific publications in the computer science domain. The dataset further contains two types of meta information from the data set: citation and authorship. These will be used as separate attributes in order to assist in the clustering process. The result of text clustering algorithm i.e. KMeans algorithm is shown in Table No. 1.The partial result of trend analysis of the paper published per year is shown in Figure 2.
The expected results of the system will be the creation of the coherent clusters and the classes with the use of this meta information which will be differ from the results which will be generated by the use of text content only. After creation of the classes the expected result is the trend analysis of the classes based on its temporal information

Table No. 1. Text Clustering Results

Cluster No	Paper Id
1	4, 8, 9, 10, 11, 15, 21, 26, 27, 31, 34, 35, 36, 38, 40, 43, 53, 59, 60, 61, 62, 68, 69, 70, 73, 75, 79, 82, 83, 85, 89, 92, 93, 94, 95
2	0, 2, 5, 6, 7, 16, 17, 19, 22, 24, 28, 29, 30, 32, 37, 42, 46, 47, 48, 50, 51, 52, 54, 55, 56, 58, 63, 64, 65, 67, 71, 72, 74, 77, 78, 80, 81, 86, 87, 88, 90, 99, 100
3	1, 4, 7, 8, 10, 11, 14, 17, 21, 28, 34, 35, 41, 45, 46, 59, 61, 66, 75, 76, 81, 82, 89, 92, 95, 96

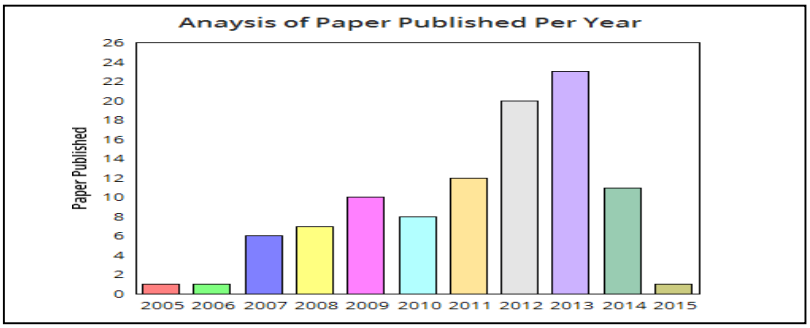

Fig 2. Trend Analysis

CONCLUSION AND FUTURE WORK

This paper provides a first approach to using other kinds of attributes in conjunction with text clustering. This approach is useful, when the meta information is provides effective guidance in creating more coherent clusters. In order to design the clustering for meta information, the proposed work will use the combination of an iterative partitioning technique with a probability estimation process, which computes the importance of different kinds of meta-information. For the clustering purpose it will use COATES algorithm and COLT algorithm will be used for the classification. These two approaches will enhance the quality of text clustering and classification, while maintaining a high level of efficiency.

For the future work the system can also consider the other kinds of meta information like access time, access frequency, publication details of the document etc. using this information the clustering and classification will be done. The system can be used as a recommendation system for the user who wants to access the documents of the particular domain, using this coherent clusters and classes.

REFERENCES

[1] Charu C. Aggarwal, Fellow, IEEE, Yuchen Zhao, and Philip S. Yu, Fellow, IEEE,"On the Use of side Information for Mining Text Data", *IEEE Transactions on Knowledge and Data Engineering*, Vol. 26, No. 6, June 2014.

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

[3] M. Steinbach, G. Karypis, and V. Kumar, "A comparison of document clustering techniques," *in Proc. Text Mining Workshop KDD*, pp. 109-110, 2000.

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

[5] S. Guha, R. Rastogi, and K. Shim, "ROCK: A robust clustering algorithm for categorical attributes," *Inf. Syst.*, vol. 25, no. 5, pp. 345-366, 2000

[6] T. Zhang, R. Ramakrishnan, and M. Livny, "BIRCH: An efficient data clustering method for very large databases," *in Proc. ACM SIGMOD Conf.*, New York, NY, USA, pp. 103-114, 1996.

[7] 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*, pp. 45-70, 2004.

[8] S. Zhong, "Efficient streaming text clustering," *Neural netw.*, vol. 18, no. 56, pp. 790-798, 2005

[9] 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, pp. 318-329, 1992.

[10] Y. Sun, J. Han, J. Gao, and Y. Yu," iTopicModel: Information network integrated topic modelling," *in Proc. ICDM Conf.*, Miami, FL, USA, pp. 493-502 2009.

[11] C. C. Aggarwal and H. Wang, "Managing and Mining Graph Data," New York, NY, *USA: Springer*, 2010

[12] C. C. Aggarwal, "Social Network Data Analytics," New York, NY, *USA: Springer*, 2011

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

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

