Clustering and Classification of Documents based on Meta Information using COATES and COLT Algorithms

Mrunal V. Upasani
P.G. Student, Computer Engg. Department, GES's R.H. Sapat College of Engineering, Management Studies and Research, Nashik Affiliated to Savitribai Phule Pune University

Rucha C. Samant
Asst. Prof., Computer Engg. Department, GES's R.H. Sapat College of Engineering, Management Studies and Research, Nashik Affiliated to Savitribai Phule Pune University

ABSTRACT
The side information means the meta information of the documents can be used for the purpose of data mining applications like clustering, classification etc. Huge amount of meta-information is available along with the text documents in many text mining applications. Such meta-information is of different kinds, likes links in the document, user-access behavior from web logs etc. which can be useful for data mining. Tremendous amount of information can be found in this unstructured attributes for clustering purposes. Therefore, this system used 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 is helpful. In this system, 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 was performed. 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 used and it shows the advantages of using such an approach.

General Terms
Data mining, information retrieval, text mining

Keywords- Classification, clustering, data mining, meta information, text mining

1. INTRODUCTION
Text clustering is used 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 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 behavior of user can be captured in the form of web logs. For each document, the meta-information is the browsing behavior 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 clustering process [1]. The primary goal of this paper is to study the 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 of data in which auxiliary information is available with text. Such scenarios are very common in a wide variety of data domains. For the creation of the clusters using meat information the system uses COATES algorithm.

After clustering the system extends the approach to the problem of classification, which provides superior results because of the incorporation of meta information. COLT algorithm for clustering with the incorporation of meta information is used by the proposed system.

Goal of this system 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. After clustering the classification is done into number of classes with labels. The trend analysis of the text documents is done according to the temporal information available in the documents.

2. RELATED WORK
A tremendous amount of work has been done over the years on the clustering in text collections in the database and information retrieval communities. The detail survey of Text Clustering Algorithms was studied in [2] [3].

Hierarchical clustering which is one of the types of the 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 or meta information is highly informative, and provides effective guidance in creating more coherent clusters. The proposed system extends the clustering method to the classification of the text documents using the algorithm which uses the meta-information for the classification purpose.

3. DETAILS OF PROPOSED WORK

The focus of the proposed work is to show the advantages of using side-information for mining text data extend beyond a pure clustering task which provides competitive advantages for a wider variety of problem scenarios.

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 are classified into number of classes with labels. After classification the trend analysis of the text documents is done according to the temporal information of the documents.

3.2 Proposed System Architecture

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

For the creation of the primary clusters which is used by the COATES and COLT algorithm, the system uses the Kmeans clustering. The results of Kmeans are 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 clustering is given as input to the next phases of the system.

3.2.2 Clustering Using COATES Algorithm

The algorithm which is 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 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 performed in this step. The iterations are content iterations (used text only) and auxiliary iterations (used text and meta information) respectively. The combination of these two is referred as a major iteration.

This algorithm maintains a set of seed centroids, which gets refined in different iterations. In each content-based phase, the system assigns a document to its closest seed centroid based on a text similarity function. In each auxiliary phase, system creates a probabilistic model, which relates the attribute probabilities to the cluster-membership probabilities, based on the clusters which have already been created in the most recent text-based phase. The goal of this modeling is to examine the coherence of the text clustering with the meta information attributes

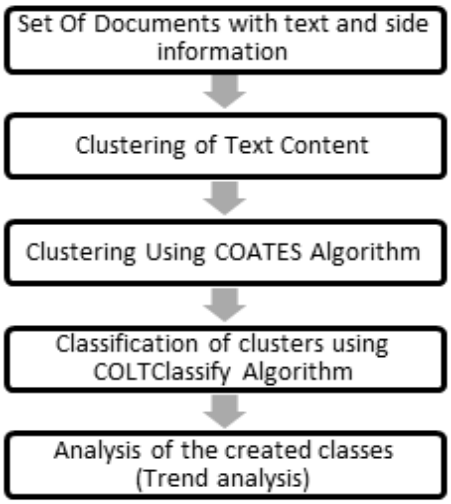

Fig. 1: Block Diagram of Proposed System

.

3.2.3 Classification of clusters using COLT Classify Algorithm

For the purpose of classification, proposed system 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 i.e. meta information.

The system computes the gini index for each attribute in the data with respect to the class label. If the gini index is below the average gini index of all attributes, then these attributes are pruned globally, and are never used further in the clustering process.

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.

Each cluster is associated with a particular class and all the records in the cluster belong to that class. This goal is achieved by first creating unsupervised cluster centroids, and then adding supervision to the process.

3.2.3.3 Cluster-Training Model Construction

In this phase, a combination of the text and meta-information is 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.

In order to perform the classification of the data, the system computed the *r* closest clusters to the test instance with the use of both content and auxiliary attributes. Specifically, for content attributes, the use of similarity based on the content attributes is done and for the case of the auxiliary attributes, the system determines the probability for the test instance.

*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 was done on the basis of the temporal information of the documents. And further the trend analysis of the created classes done accordingly, the statistical representations of the results was made.

4. EXPERIMENTAL SETUP AND RESULTS

4.1 Experimental Setup

The CORA Dataset was 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 side information from the data set: citation and authorship. These were be used as separate attributes in order to assist in the clustering process. The citation of the documents is used as a meta-information for the implementation of this system. The citations were converted into 0-1 variable, which indicates whether or not the ith document has is having the citation of rth document or not..

This information can be used in order to cluster the documents in a site in a more informative way than a techniques which is based purely on the content of the documents.

The implementation environment for the proposed system uses the windows operating system, Java SE Development Kit 8 and Eclipse Juno version.

4.2 Results

The result of text clustering algorithm i.e. Kmeans algorithm is shown in Table 1. It shows the clustering results of Kmeans algorithm when the 50 documents are given to the system which does not contain any meta information. The documents were clustered in 4 clusters. The clustered documents are the document ids of the input documents. Cluster 1 contains the documents which are having the largest similarity among themselves. This clustering is done for the text only documents. So for the similarity purpose the cosine similarity concept is used.

Table 1. Result of Kmeans without meta information

Cluster Name	Cluster Node
1	19,20,21,23,38,40,44
2	2,3,4,5,6,8,11,12,13,14,18,22,24,25,26,27,29,30,31,32,33,35,36,37,39, 41,42,43,45,47,48,49
3	16,17
4	0,1,7,9,10,28,34,46

The result of COATES algorithm is shown in Table 2. For this algorithm the input is the documents with the meta information which is in the form of binary values i.e. 0 and 1. The meta information used here is the citations given in the documents. The results of COATES algorithm are better than the Kmeans which was implemented without meta information. The numbers of the clusters for the result of the COATES are also 4 but the result differs in the accuracy and the purity of the clusters. The difference between result of Kmeans and COATES is evaluated in next section.

Table 2. Result of COATES with meta information

Cluster Name	Cluster Node
1	19,20,21,22,23,24
2	0,1,2,3,4,5,17,25,28,31,34,36,38,42,47,6,7,12,13,14,15,18,26,27,29,30,32,33,35, 37,49
3	16,39,40,41,43,44,45,46,48
4	2,8,9,10,11

After COATES, the COLT algorithm is applied on the documents with the meta information, the output of this algorithm is the prediction of the most probable class for the test document according to its label. The result of COLT

classification is shown in Table 3. The result of the COLT algorithm using Meta information for the classification was better than the traditional classification algorithm like Naïve Bayes and ID3 Classifier. The classifier classifies the documents into the dominant class. The evaluation of the result of COLT Classifier as compare to other 2 is given in next section.

Table 3. Result of COLT

Classifier	COLT	Naïve Bayes	ID3
Test Document number	21	21	21
Predicted Label	1	2	2

4.3 Evaluation Metrics

The evaluation metrics of the clustering is the cluster purity which is defined as the fraction of documents in the clusters which correspond to its dominant class. The cluster purity always lies between 0 and 1.

Table 4 shows the various class labels used for the implementation of the system.

Table 4. Label number with name

Number	**Label Name**
1	Information Retrieval
2	Operating System
3	Human Computer Interface
4	Encryption
5	Databases
6	Artificial Intelligence

Accuracy of Kmeans without meta information which is evaluated by the means of purity**.** Results of label assignments for the clusters created by Kmeans are shown in Table 5. The accuracy of the Kmeans can be calculated by the predicated cluster label of the documents. The labels are the dominant classes in which the documents are belonging. For cluster 1 the label assigned is 6. Means the cluster 1 should contain the number of documents which are from label 6. So the cluster 1 contains the 4 documents which are belonging to label 6. For the cluster 2 the label is 1 and cluster 2 contains 14 documents which are belonging to label 1. Cluster 3 contains the 2 documents of label 4. And finally cluster 4 contains the 5 documents of label 1. So from this the purity of Kmeans is 0.5.

Purity=4+14+2+5/50=0.5

Table 5. Cluster number and label number for Kmeans

Cluster Number	Label Number
1	6
2	1
3	4
4	1

Accuracy of COATES algorithm which considers the meta information while clustering evaluated by the means of purity**.** Results of label assignments for the clusters created by COATES are shown in Table 6. For cluster 1 the label assigned is 6. So the cluster 1 contains the 6 documents which are belonging to label 6. For the cluster 2 the label is 1 and cluster 2 contains 14 documents which are belonging to label 1. Cluster 3 contains the 7 documents of label 4. And finally cluster 4 contains the 5 documents of label 1. So from this the purity of COATES is 0.62.

Purity=6+14+7+5/50=0.62

This shows that the proposed COATES algorithm for clustering creates the more pure cluster than Kmeans which uses text content only for clustering.

Table 6. Cluster number and label number for COATES

Cluster Number	Label Number
1	6
2	1
2	2
4	1

For the testing of COLT Classify algorithm the test document with document id 21 which is from Artificial Intelligence label is given as an input to the COLT classifier, it predicted its class label as 1, because from the output of COATES it was predicted that the cluster1 belongs to the class label 6 and it classifies test document correctly , whereas result of naïve bayes classifier and ID3 classifier is not correct for the classification of test document which was given as the input , they both predicts the wrong class for the test document.

The comparison of results of clustering of the documents by Kmeans and COATES is shown in Fig. 2

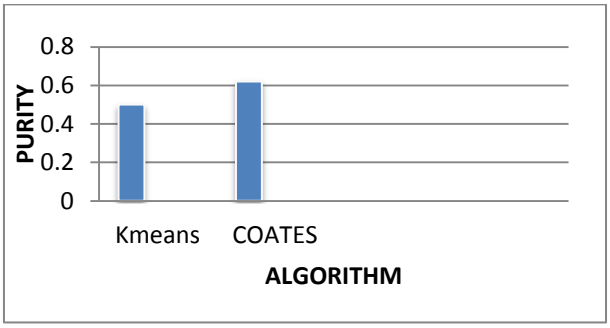

Fig. 2: Comparison of Kmeans and COATES

Finally the trend analysis was done on the documents for making the statistical analysis of the paper published per year. Fig. 3 shows the result of trend analysis.

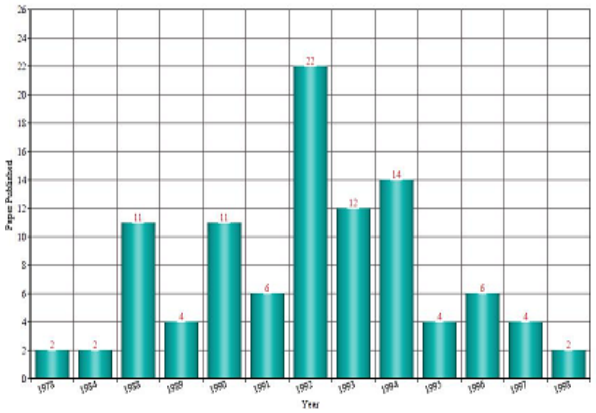

Fig 3: Result of trend analysis

5. CONCLUSION AND FUTURE WORK

This system provides a first approach to using other kinds of attributes in conjunction with text clustering. This approach is

useful, when the meta information provides effective guidance in creating more coherent clusters. In order to design the clustering for meta information, the proposed work used 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 used COATES algorithm and COLT algorithm was used for the classification. The result shows that these two approaches greatly 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 is 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.

6. ACKNOWLEDGMENTS

We are glad to express our sentiments of gratitude to all who rendered their valuable guidance to us. We would like to express our appreciation and thanks to Prof. Dr. P. C. Kulkarni, Principal, G. E. S. R. H. Sapat College of Engg., Nashik. We are also thankful to Prof. N. V. Alone, Head of Department, Computer Engg., G. E. S. R. H. Sapat College of Engg., Nashik. We thank the reviewers of IJCA for their comments.

7. REFERENCES

[1] Charu C. Aggarwal, Yuchen Zhao,Philip S. Yu, "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 ,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 WorkshopKDD*, pp. 109-110, 2000.

[4] S. Guha, R. Rastogi, 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, 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, M. Livny, "BIRCH: An efficient data clustering method for very large databases," *in Proc. ACMSIGMOD 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, 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, Y. Yu," iTopicModel: Information network integrated topic modelling," *in Proc. ICDM Conf.*, Miami, FL, USA, pp. 493-502 2009.

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

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