On The Use of Side Information

T. Saleem[1] A. Shiva Kumar[2]
[1,2]Tudi Ram Reddy Institute of Technology & Sciences, JNTU Hyderabad

Abstract— Text mining has been around for many years in order to extract latent information from textual documents. However, there is meta-data associated with the textual documents. Such data is nothing but the provenance information, links related to documents, user access related data. The meta-data plays a vital role in understanding the documents and their usage dynamics. Based on this information it is possible to achieve clustering of such documents. Textual and non-textual information can be used to help improve clustering process. However, knowing the importance of meta-data and how it is useful in clustering is non trivial. Therefore it is important to make use of meta-data that is important and reliable in order to use it for clustering process. There is noise that can be understood and removed in order to achieve quality in clustering. In this paper we propose and implement a mechanism that helps in making effective clustering. We built a prototype application that can be used to demonstrate the proof of concept. The empirical results revealed that the proposed mechanism works fine for clustering textual data based on associated meta-data.

Key words: Data mining, text mining, text clustering, classification, meta-data

I. INTRODUCTION

Text clustering is the process of grouping similar documents that have related information. Since the WWW is rich in textual data, it is widely used to perform text clustering. The text clustering process is meant for analysing textual data in order to group documents so as to help in different applications in the real world. When text documents are grouped together search in particular group becomes easy. This way many applications are possible on the clustered documents.

Traditionally clustering takes place based on the term frequency in the textual documents. However, with increased usage of web and datasets, there is associated data with every text document. The associated data is also known as side information. This data is very useful to make clustering decisions. This data is known as meta data. Meta data is of many types. The data which is related to a document in terms of links, number of users accessing the data and other related information. This data can be used in data mining or text mining in order to perform clustering, classification and other text mining activities.

In this paper our focus is clustering the documents based on the meta data available. We proposed and implemented a new mechanism that takes care of clustering documents based on the side information. We built a prototype application that can be used to demonstrate the proof of concept. The empirical results revealed that the proposed mechanism works fine for clustering textual data based on associated meta-data. The remainder of the data is structured as follows. Section II provides review of literature. Section III presents the proposed system in detail. Section IV presents experimental results while section V concludes the paper.

II. RELATED WORKS

Mining textual data has been around for many years. Processing data which is present in text files or unstructured data is very important as great majority of data in the WWW is in the form of text. Thus the text clustering came into existence. Text clustering is one of the approaches to process textual data. This is widely studied in the literature as explored in [1], [2] and [3]. The major part of the literature reveals that there has been increased research in the concept known as scalable clustering of data of different types with even multiple dimensions existed [4]. A good review of clustering algorithms that work on the textual data can be found in [5] and [6]. In the context of textual data only research is in abundance. Scatter-gether is the technique [7] well known for text clustering. It makes use of paritional and agglomerative clustering techniques. It does mean that it is the blend of both and gets the synergic effect in its results of clustering. There are many other methods available for clustering as explored in [8] and [9]. Co-clustering is another concept of clustering which is used in the literature. It is explored in [10] and [11]. Another method known as expectation maximization (EM) is used for text clustering [12]. Yet another method used for text – clustering is known as Matrix factorization as discussed in [13]. Based on the relevance this technique selects words from the documents in order to complete the clustering process. It is one of the alternative methods for expectation maximization method.

Topic-Modelling is another area which is similar to the text clustering. Other similar areas include text-categorization, and event tracking as explored in [14], [15], [16] and [17]. Topic driven clustering is used for clustering textual data based on the topics given [18]. With respect to key word extraction, text clustering methods were explored in [19]. Network based linkage information and its related work based on text clustering is found in [20], [21], and [22]. For side information attributes, these techniques are not suitable. In this paper we proposed as approach using certain attributes that work in tandem with text clustering.

Some limited work has been done on clustering text in the context of network-based linkage information [23], [24], [25], [26], [27], [28], [29], [30], though this work is not applicable to the case of general side information attributes. In this paper, we will provide a first approach to using other kinds of attributes in conjunction with text clustering. We will show the advantages of using such an approach over pure text-based clustering. Such an approach is especially useful, when the auxiliary information is highly informative, and provides effective guidance in creating more coherent clusters. We will also extend the method to the problem of text classification, which has been studied extensively in the literature. Detailed surveys on text classification may be found in [31], [32].

III. PROPOSED SYSTEM

This section provides the details of the proposed system. As shown in Figure 1, the proposed system has an algorithm meant for clustering documents. User provides many documents as input. The clustering algorithm extracts meta data and uses it for clustering decisions. The output is a group of documents that are related. These clusters are very useful for further processing in the real world applications.

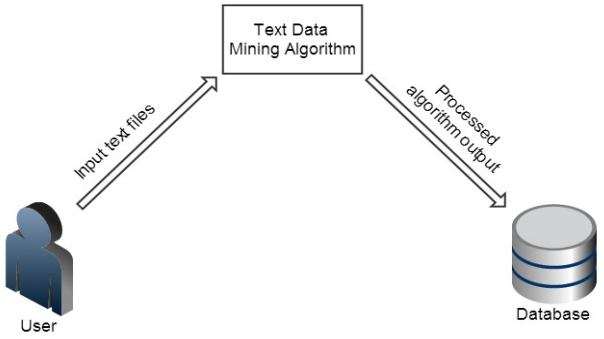

Fig. 1: Proposed mechanism

As can be seen in Figure 1, it is evident that the proposed mechanism throws light into text mining especially clsustering of given documents. We proposed an algorithm to achieve this. The algorithm has two phases namely extracting meta data of documents and the clustering process. In the first phase, the algorithm extract meta data while in the second phase actual clustering takes place based on the side information available.

A. *Text Clustering Algorithm*

Algorithm: Text Clustering Algorithm
Inputs: Text Documents ***D***
Output: Clusters ***C***
01 Initialize Metadata ***MD***
02 For each ***d*** in ***D***
03 Extract metadata ***md*** related to the ***d***
04 Add ***md*** to ***MD***
05 End For
06 Initialize ***C*** with different documents
07 For each ***d*** in ***D***
08 Use **md** for comparison
09 Add ***d*** to ***C*** to update clusters
10 End For
11 Return ***C***

Fig. 2:

As can be seen in the proposed algorithm, there are two things important. First one is set of documents and the second one is the associated meta data for each document. The meta dat is used to make clustering decisions. Initially the clusters are with a single document and as the process goes on the clusters are updated with relevant documents. The aim of this clustering is to group documents with high intra-cluster similarity and low inter-cluster similarity.

B. *Datasets Used*

We used to datasets for experiments in this paper. They are known as Cora dataset and DBLP-Four-Area dataset. The former has plenty of scientific publications while the latter also has data of different publications. These datasets provide ample opportunity to explore different aspects of text clustering and also clustering based on the additional information available with the publications.

IV. EXPERIMENTS AND RESULTS

Experiments are made with the prototype application. The datasets are used as described in the previous section. The results reveal that there is clustering of documents based on the side information or meta data. The results are as follows.

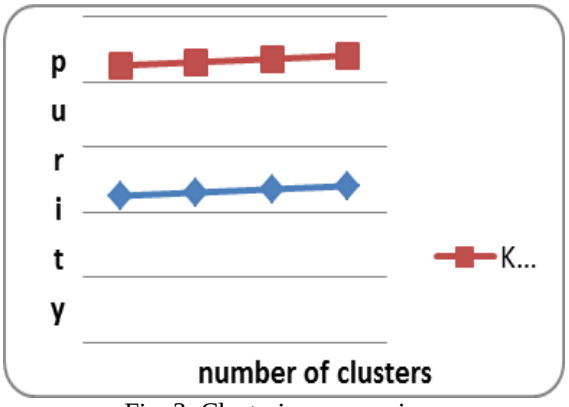

Fig. 3: Clustering comparison

As can be seen in Figure 2, it is evident that there is clustering with two approaches and the results are compared. The results revealed that K-means achieved more purity than other approach. The results are on purity versus number of clusters.

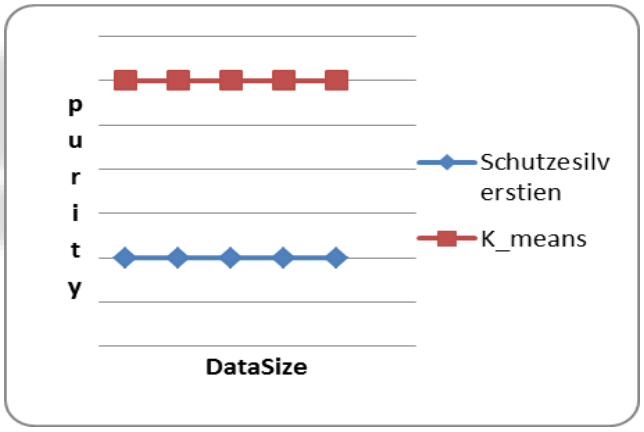

Fig. 3: Clustering Comparison with Data Size

As can be seen in Figure 3, it is evident that there is clustering with two approaches and the results are compared. The results revealed that K-means achieved more purity than other approach. The results are on purity versus size of dataset.

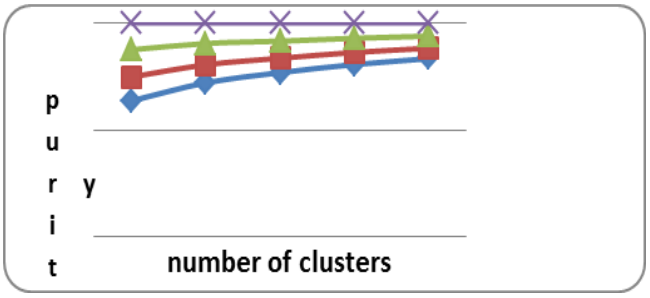

Fig. 4: Clustering Comparison

As can be seen in Figure 4, it is evident that there is clustering with three approaches and the results are compared. The results revealed that the proposed approach achieved more purity than other approach. The results are on purity versus number of clusters.

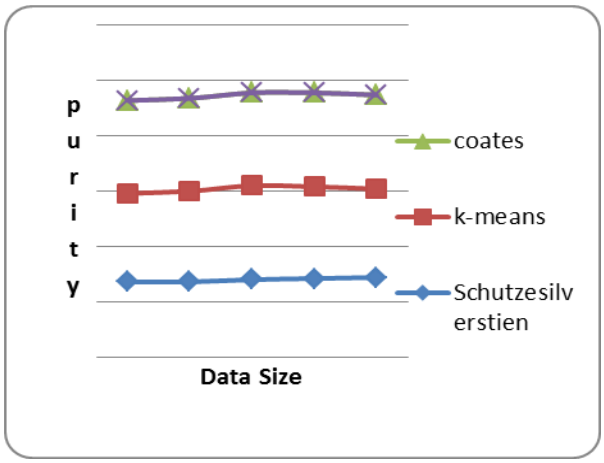

Fig. 5: Clustering Comparison Based on Data Size

As can be seen in Figure 5, it is evident that there is clustering with three approaches and the results are compared. The results revealed that the proposed approach achieved more purity than other approach. The results are on purity versus data size.

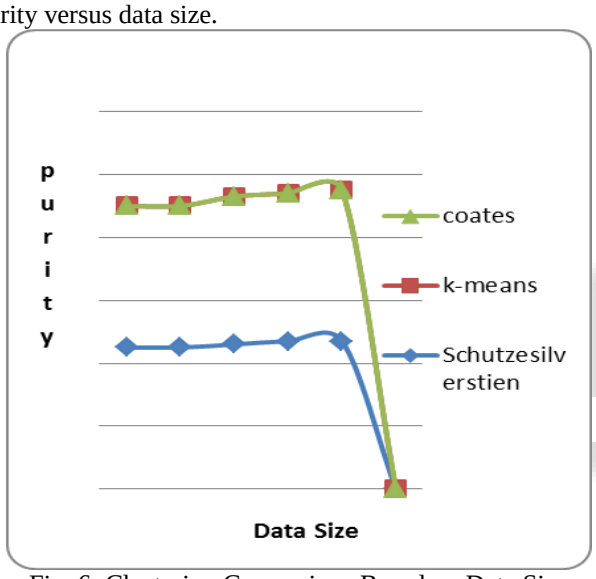

Fig. 6: Clustering Comparison Based on Data Size

As can be seen in Figure 6, it is evident that there is clustering with three approaches and the results are compared. The results revealed that the proposed approach achieved more purity than other approach. The results are on purity versus data size.

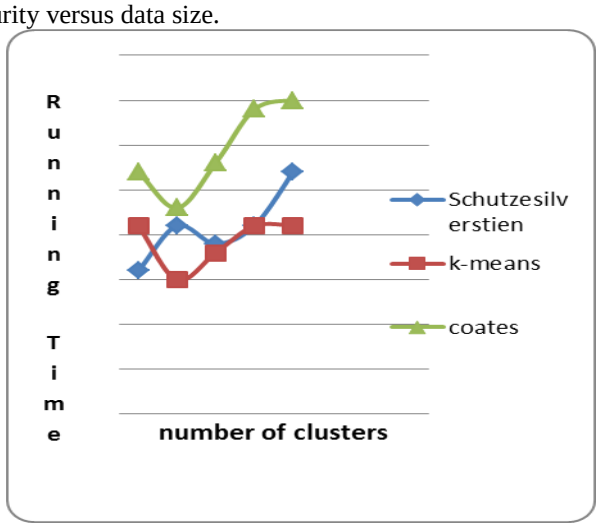

Fig. 7: Clustering Comparison Based Number of Clusters Vs. Time

As can be seen in Figure 7, it is evident that there is clustering with three approaches and the results are compared. The results revealed that the proposed approach needs more running time.

V. Conclusions and Future Work

In this paper we studied clustering of textual documents. The existing techniqniues on text clustering are mostly based on the term frequency and other metrics that are widely used in the literature. Howver, every textual document has associated meta data that can provide valuable information in processing text documents. Especially, in text mining, it is possible to use such meta data for classification or clusteing tasks. In this paper we used meta data in order to perform clustering of documents. We proposed an algorithm that works in two phases. In the first phase, it extracts the meta data of each document and prepares a vector. In the second phase, this vector information is used in order to perform clustering techniques. The similarity measure is from 0.0 to 1.0 inndicating dissimilar and highly similar respectively. The value between these two indicates the relative similarity. The proposed algorithm makes clusters of given documents abased on the Meta data. Our prototype application demonstrated the proof of concept. The empirical results revealed that the algorithm works fine when Meta data is available and there is relevant metadata. This research can be improved further in order to improve the quality of clustering by fine-tuning the side information associated with documents.

References

[1] T. Zhang, R. Ramakrishna, and M. Livny, "BIRCH: An efficient data clustering method for very large databases," in Proc. ACM SIGMOD Conf., New York, NY, USA, 1996, pp. 103–114.
[2] R. Ng and J. Han, "Efficient and effective clustering methods for spatial data mining," in Proc. VLDB Conf., San Francisco, CA, USA, 1994, pp. 144–155.
[3] C. C. Aggarwal and C.-X. Zhai, Mining Text Data. New York, NY, USA: Springer, 2012.
[4] 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.
[5] A. Jain and R. Dubes, Algorithms for Clustering Data. Englewood Cliffs, NJ, USA: Prentice-Hall, Inc., 1988.
[6] 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.
[7] D. 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, 1992, pp. 318–329.
[8] H. Schutze and C. Silverstein, "Projections for efficient document clustering," in Proc. ACM SIGIR Conf., New York, NY, USA, 1997, pp. 74–81.
[9] C. Silverstein and J. Pedersen, "Almost-constant time clustering of arbitrary corpus sets," in Proc. ACM SIGIR Conf., New York, NY, USA, 1997, pp. 60–66.

[10] I. Dhillon, "Co-clustering documents and words using bipartite spectral graph partitioning," in Proc. ACM KDD Conf., New York, NY, USA, 2001, pp. 269–274.
[11] I. Dhillon, S. Mallela, and D. Modha, "Information-theoretic coclustering," in Proc. ACM KDD Conf., New York, NY, USA, 2003, pp. 89–98.
[12] C. C. Aggarwal, Social Network Data Analytics. New York, NY, USA: Springer, 2011.
[13] W. Xu, X. Liu, and Y. Gong, "Document clustering based on nonnegative matrix factorization," in Proc. ACM SIGIR Conf., New York, NY, USA, 2003, pp. 267–273.
[14] 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.
[15] 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.
[16] A. Banerjee and S. Basu, "Topic models over text streams: A study of batch and online unsupervised learning," in Proc. SDM Conf., 2007, pp. 437–442.
[17] 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, pp. 245–255, Feb. 2004.
[18] Q. Mei, D. Cai, D. Zhang, and C.-X. Zhai, "Topic modeling with network regularization," in Proc. WWW Conf., New York, NY, USA, 2008, pp. 101–110.
[19] 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.
[20] Y. Zhou, H. Cheng, and J. X. Yu, "Graph clustering based on structural/attribute similarities," PVLDB, vol. 2, no. 1, pp. 718–729, 2009.
[21] F. Sebastiani, "Machine learning for automated text categorization," ACM CSUR, vol. 34, no. 1, pp. 1–47, 2002.
[22] Y. Sun, J. Han, J. Gao, and Y. Yu, "iTopicModel: Information network integrated topic modeling," in Proc. ICDM Conf., Miami, FL, USA, 2009, pp. 493–502.
[23] C. C. Aggarwal and H. Wang, Managing and Mining Graph Data. New York, NY, USA: Springer, 2010.
[24] 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.
[25] J. Chang and D. Blei, "Relational topic models for document networks," in Proc. AISTASIS, Clearwater, FL, USA, 2009, pp. 81–88.
[26] C. C. Aggarwal and C.-X. Zhai, "A survey of text classification algorithms," in Mining Text Data. New York, NY, USA: Springer, 2012.
[27] T. Yang, R. Jin, Y. Chi, and S. Zhu, "Combining link and content for community detection: A discriminative approach," in Proc. ACM KDD Conf., New York, NY, USA, 2009, pp. 927–936.
[28] Y. Zhao and G. Karypis, "Topic-driven clustering for document datasets," in Proc. SIAM Conf. Data Mining, 2005, pp. 358–369.
[29] M. Steinbach, G. Karypis, and V. Kumar, "A comparison of document clustering techniques," in Proc. Text Mining Workshop KDD, 2000, pp. 109–110.
[30] Y. Sun, J. Han, J. Gao, and Y. Yu, "iTopicModel: Information network integrated topic modeling," in Proc. ICDM Conf., Miami, FL, USA, 2009, pp. 493–502.
[31] W. Xu, X. Liu, and Y. Gong, "Document clustering based on nonnegative matrix factorization," in Proc. ACM SIGIR Conf., New York, NY, USA, 2003, pp. 267–273.

