A Review on Clustering Techniques using Side-Information for Mining

Firdous Sadaf M.Ismail
Department of Computer Science & Engineering
J.D.College of Engineering & Management
Nagpur, India
sadaf_firdous2810@rediff.com

Prof. Amol G. Muley
Department of Computer Science & Engineering
J.D.College of Engineering & Management
Nagpur, India
amolmuley300@rediffmail.com

Abstract—In various applications of text mining, text documents are associated with side-information. Such side-information may be categories of different kinds; this may include document provenance information, links in documents, web logs of user access behavior or other non-textual attributes which are embedded into the text document. This attribute may consist of a tremendous amount of information for the purpose of clustering. However, this may be difficult to estimate the relative importance of this side information, especially when some of the information is noisy. So in such cases, the use of side-information for mining process can be risky, because it can either improve quality of mining process representation, or can add noise to the process. Therefore, the proposed approach gives the principled way to perform the mining procedure to maximize the advantages from using this type of side information. In this paper, proposed approach design such an algorithm which combine classical partitioning algorithm with probabilistic models in order to create an effective and efficient clustering approach. This algorithm then shows how to extend the approach in the classification problems. These present experimental results on a number of distinct data sets in order to illustrate the advantages of using such an approach.

Keywords*-Clustering, Data Sets, Data Mining*

*****

I. INTRODUCTION

In various application domains, huge amounts of side information are also linked along with the documents. This is because text document typically do occur in the context of a variety of applications in which there may be a large amount of other kinds of database attributes or Meta information which may be useful to the clustering process. This type of side information is also referred as auxiliary attributes. A huge amount of work has been done in recent years on the problem of clustering in text collections [3], [4], [5] in the database and communities of information retrieval [2][8].

However, the work is primarily developed for the problem of pure text clustering when the other kinds of attributes are not present. The user processes includes the rapidly increasing of large 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 problem of text clustering arises in the context of many application domains such as the web, social networks, and other digital collections.

In an application in which we track user access behavior of web document, the behavior of user-access may be captured in web logs form. For each and every document, meta-information may correspond to the browsing behavior of the different users. Such logs can be used to enhance the quality of the mining process in a way which is more meaningful for the user and it is also application-sensitive. Many text documents contain the various links among them. This links can be treated as attributes also. These types of links contain a lot of useful information for mining purposes. Many web documents have meta-data associated with them which correspond to different kinds of attributes such as the provenance or other information about the origin of the documents. In a number of various network and user-sharing applications, documents may be related with user-tags, which may be quite informative also. All these are the examples of Side Information which plays an important role for mining text data.

The core approach is to govern a clustering in which the text attributes and side-information will provide similar hints about the nature of underlying clusters. At the same time, it ignores those aspects where conflicting hints are provided. While such side-information can sometimes be useful in improving the quality of the clustering processes, it can be risky approach when the side-information is noisy. In such type of cases, it can actually aggravate the quality of the mining processes. Therefore, we will use an approach which carefully ascertains the coherence of the clustering characteristics of the side information with that of the text content. This approach helps in enlarging process of the clustering effects of both kinds of data.

In order to achieve this goal, we will combine a partitioning approach with probabilistic estimation processes, which governs the coherence of the side-attribute in the clustering process. The probabilistic model on side information does use the partitioning information (from text attributes) for the purpose of estimating the coherence of different clusters with side attributes. This approach helps in abstracting out noise in the membership behavior of different attribute. The partitioning approaches are specifically designed to be very well organized for big data sets. This approach can be important in scenarios where the data sets are very huge. We will show the experimental results on a number of various real data sets, and do illustrate the effectiveness and efficiency of approach.

II. DATASETS FOR CLUSTERING

To test our proposed approach we will use the real data set. Such datasets include KDD cup, mining paper set and Google search set. This data set is a collection of real data on which we apply clustering using side information for mining. Most commonly this corresponds to the database table or a single statistical data matrix. Datasets usually come from actual observations obtained by sampling a statistical population. In our proposed method we can use various data sets such as CORA (Coriolis Ocean database Reanalysis) is a global oceanographic temperature and salinity dataset produced and maintained by the French institute IFREMER. The real-time data are mostly the data that coming from different types of platforms. Platform include research vessels, underwater gliders, profilers, moored buoys, drifting buoys, sea mammals, and opportunity ships. An online collection of information related to movie is provided by Internet Movie Data Base (IMDB). This data set is capable of providing ten years of movie data. The data in it is distinguished by several genres. This genre includes Comedy, Drama, Short and Documentary. The data set provides four research areas related to data mining such is machine learning, information retrieval, database and Data mining.

III. Related Work

R. Angelova and S. Siersdorfer proposed iterative relaxation of cluster assignments [6] that can be built on top of any clustering algorithm such as k-means and DBSCAN. The proposed techniques by them results in robust self-organization, better overall accuracy and higher cluster purity. By their proposed approach, they observe that cluster distance metric improve the clustering based on graph-based and this approach outperforms all pure content based procedures. S. Zhong extends the work to cluster text streams based on efficient online spherical k-means. The extension is essentially a combination of scalable clustering techniques with the online spherical k-means (OSKM) algorithm. With this combination it improves the effectiveness and efficiency of OSKM as well as it provide the ability of exponentially reduce the contribution of history data. He designed experiments to investigate the forgetting mechanism’s effect. He concludes that at a series of stream points clustering quality is high when the decay parameter is small. The problem of multidimensional clustering have been observed [9][10]. The D. Cutting, D. Karger, J. Pedersen, and J. Tukey proposed an approach to document Clustering [4]. They asked how clustering can be effective as an access method in its own right rather than dismissing document clustering as a poor tool for enhancing near-neighbor search. They described a document browsing method, called Scatter/Gather. This method uses document clustering as its primitive operation. This technique is directed towards information access with non-specific goals and serves as a complement to more focused techniques they introduced two near linear time clustering algorithms in which experimentation has shown to be effective. The problem of text-clustering has been studied in [3]. The applications of categorical and text data stream clustering involves many portals on the World Wide Web which provide real time news and other articles. It requires filtering and quick summarization. These methods often require efficient and effective methods for text segmentation. Many web crawlers continuously harvest thousands of web pages on the web, which is subsequently summarized by human effort. In many electronic commerce applications, large volumes of transactions are processed on the World Wide Web. Such transactions can take the form of categorical or much market basket record. In such cases, it is often useful to perform real time clustering for target marketing. When the volume of such crawls is significant, it is not possible to achieve such goal by human effort. In such applications, data stream clustering algorithms can be helpful in organizing the crawled resources into coherent sets of clusters. The A. Banerjee and S. Basu proposed closely related area of event tracking, topic-modeling, and text-categorization. They demonstrated that while LDA is good, at finding word-level topic, vMF is more efficient and effective at finding document-level clusters. They presented a practical hybrid scheme for topics modeling over documents streams. It provided a good tradeoff between accuracy and speed while performing unsupervised learning over a huge volume of text. By comparing the performance of different offline topic modeling algorithms and proposed a online vMF algorithm that outperforms online versions of LDA and DCM in efficiency and performance.

IV. Problem Definition

According to previously proposed approaches, it is observe that although Side-information can be useful for improving the quality of clustering process, but it can be risky approach when the side-information is noisy. In such cases, the quality of mining processes is actually useless to the user for the required text data. This [6] work is not applicable to the case of general side-information attributes. A document or text is always represented [11] as a bag of words in text clustering. So, this representation raises one severe problem such as the high dimensionality of the feature space and the inherent data sparsity, since a single document has a sparse vector over the set of all terms. The performance of clustering algorithms will decline dramatically due to the problems of data sparseness and high dimensionality. Therefore it is highly desirable to reduce the feature space dimensionality [4].To implement Scatter/Gather, fast document clustering is a necessity otherwise it will slow down the needed operational speed for the user. The problem [3] of categorical data streams and clustering text also affect the efficient clustering methods. This problem is relevant in a number of web related applications such as text crawling, news group segmentation, target marketing and text crawling for electronic commerce.

V. Research Methodology

The core objective of this research is to determine a clustering in which the side information provides similar hints about the underlying cluster and to ignore those aspects of clustering methods in which conflicting hints are provided. The performance measures are expected from this approach can be computed in terms of time that means delay in clustering, accuracy level of clustering and similarity between inter and intra domain clustering. This paper primarily proposed for the problem of pure text clustering when the other kinds of attribute are not present. Our Objective is to show that the advantages of using side-information extend beyond a pure clustering task. It can provide competitive advantage for a wider variety of problem scenarios. This paper provides an approach which will enhance the quality of the mining process in a way which would be more meaningful for the users and would be application sensitive. We will use an approach which carefully ascertains coherence of clustering characteristics of side information with text content. This helps in enlarging the clustering effects of several kinds of data.

We will present a COATES algorithm by using side information for text clustering, which indicate the fact that algorithm is COntent and Auxiliary attribute based Text cluStering algorithm. Here our assumption is that an input to the COATES algorithm is number of clusters k. As we examined in case of all types of text-clustering algorithm, the assumption is that all stop-word has been removed, and linguistic morphology has been executed in order to improve the unfairly power of the attributes.

The COATES algorithm requires two phases. In the Initialization phase without using any side-information, a standard text clustering idea is used. Therefore for this purpose, we are using the algorithm explained in [1]. The reason behind using this algorithm is that, since it is very simple algorithm which has the capability to process quickly and very efficiently by providing the reasonable initial starting point. The centroids and partitioning made by clusters created in the first phase give an initial starting point to use in the second phase. We should take care in mind that the first phase is based on text only. First phase do not uses the Side (auxiliary) information.

In Second phase (Main Phase) of COATES algorithm are executed after the execution of first phase. Main phase do start off with these initial groups, and then iteratively reconstructs the clusters by using of both text content and Side (auxiliary) information. Main phase performs alternating iterations which uses text content and side attribute information in order to enhance the quality of the clustering process. We refer these iterations as content iterations and auxiliary iterations respectively. The combinations of these two iterations are referred to as a major iteration. Therefore, major iteration contains two minor iterations, corresponding to the side and text-based methods respectively.

The main focuses of first phase are simply to construct an initialization that provides good starting point for a clustering process based on text contents. Since a key technique for content and side information integration is in the second phase of algorithm, we will mainly focus on most of subsequent discussion on the second phase of the algorithm. The first phase is a direct application of text clustering algorithm. The overall approach of algorithm uses alternating minor iteration of content-based and auxiliary (side) attribute-based clustering. These two phases are referred to as content-based and auxiliary attribute-based iteration respectively. These algorithms maintain a collection of seed centroids that are subsequently refined in different iterations.

VI. Conclusion

In this Paper, We proposed an enhanced approach for clustering the text data based on side information which would upgrade the quality of the mining process in a meaningful way required by the user. It determines a clustering in which side-information provide similar required data about the behavior of the underlying clusters and ignoring the aspects where the conflicting data (hints) are provided. We have presented the methods for mining text data with the use of side-information. In order to design clustering, we combined an iterative partitioning technique which computes the importance of different kinds of side-information. The result showed that use of side-information can greatly enhance the quality of text clustering and classification, while maintaining a high level of efficiency.

VII. Acknowledgment

I have taken efforts in this approach of clustering for mining using side information. However, it would not have been possible without the kind support and help of many individuals. I am highly indebted to Prof. Amol G. Muley for his guidance and constant supervision as well as for providing necessary information regarding this approach.

References

[1] Charu C. Aggarwal, Yuchen Zhao, and Philip S. Yu " On the Use of Side Information for Mining Text Data" in IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, VOL. 26, NO. 6, JUNE 2014,pp. 1415-1429.

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

[3] C. C. Aggarwal and P. S. Yu, "A framework for clustering massive text and categorical data streams," in *Proc. SIAM Conf. Data*

[4] 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.

[5] S. Zhong, "Efficient streaming text clustering," *Neural Netw.*, vol. 18, no. 5–6, pp. 790–798.

[6] R. Angelova and S. Siersdorfer, "A neighborhood-based approach for clustering of linked document collections," in *Proc. CIKM*

[7] 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.

[8] H. Schutze and C. Silverstein, "Projections for efficient document clustering," in *Proc. ACM SIGIR Conf.*, New York, NY, USA, 1997,

[9] S. Guha, R. Rastogi, and K. Shim, "CURE: An efficient clustering algorithm for large databases," in *Proc. ACM SIGMOD Conf.*, New

[10] 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.

[11] T. Liu, S. Liu, Z. Chen, and W.-Y. Ma, "An evaluation of feature selection for text clustering," in *Proc. ICML Conf.*, Washington, DC, USA, 2003, pp. 488–495.