Volume : 4 | Issue : 7 | July 2015 • ISSN No 2277 - 8179

Research Paper

Clustering Using Side-Information with NLP Operations for Data Mining

Medical Science

KEYWORDS : Clustering, Side Information, NLP Libraries, Data Mining

Firdous Sadaf M.Ismail

Department of Computer Science & Engineering J.D.College of Engineering & Management Nagpur, India

ABSTRACT

In previous procedures of application of mining data, on basis of algorithms of pure text clustering number of cluster are developed. Although such type of pure text clustering algorithms are useful in the mining process for making the clusters but this can be a risky approach since text may have repeated or redundant information. This type of noisy and ambiguous information is noisy and provides conflicting hints which reflects the performance of mining process and degrades the efficiency of clustering outcome. So in proposed approach with the use of text clustering we are using the side information which is present with the document. This Auxiliary Attribute or annotations which is referred to as side information clustering are performed after the datasets are got filtered using the Natural Language Processing libraries. That is the reason proposed approach provides far better efficiency than the previous clustering algorithm.

INTRODUCTION

In several application domains, large amount of side information are also associated 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 big amount of other types of database attributes or Meta information which may be useful to the clustering process. This type of side information is also known as an auxiliary attributes. In recent years on the problem of clustering in text collections a huge amount of work has been done [3], [4], [5] communities of information retrieval and in the database [2][8]. However, when the other kinds of attributes are not present the previous work was primarily designed for the problem of pure text clustering. 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. Many application domains such as several digital collection and social networks and web has been suffering from the problem of text clustering.

The basic idea of proposed approach is to govern a clustering in which the text attributes and side-information (Annotations) will provide similar hints about the nature of underlying clusters. At the same time where conflicting hints are provided, it ignores those types of aspects. Although 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 ambiguous and noisy. In such type of cases, it can actually spoil the accuracy and efficiency of mining process. Hence, we have suggested an approach where it carefully ascertains the coherence of the clustering characteristics of the side information with that of the text content. The proposed approach helps in developing process of the clustering effects.

To achieve this objective, we combined a partitioning approach with probabilistic estimation processes, which guide the coherence of the side-attribute in the clustering process with NLP processed information for clustering. This approach helps in abstracting out unwanted information in the membership behavior of different property. The partitioning approaches are specifically developed to be organized for big data sets very well. This approach can be important in scenarios where the data sets are extremely large. We have shown the experimental results on a number of various real data sets, and illustrated the effectiveness and efficiency of proposed approach.

PROBLEM DEFINITION

In previously proposed approaches, it is observe that since Side-information can be useful for improving the quality of clustering process, but it can be sensitive approach when the side-information is redundant and ambiguous. In those cases, the accuracy of mining processes is actually of no use to the user for the needed text data. This work [6] isn't applicable to the case of general auxiliary/annotated attributes. A dataset or document is represented as a bag of words always in [11] data clustering. Performance of clustering algorithms' will decline dramatically due to certain problems of data sparseness and large dimensionality. It is very desirable to reduce the feature [4] space dimensionality. To create Scatter/Gather, fast document clustering is a necessity otherwise the required operational speed for the user will slow down. The problem [3] of categorical data streams and clustering text also affect the efficient clustering methods. Difficulty is equivalent seen in a number of web related applications such as target marketing, text crawling, and news group segmentation for electronic commerce.

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. S. Zhong extends the work to cluster text streams based on efficient Online Spherical k-Means. This combination improves the efficiency and effectiveness of OSKM as well as it provide the ability of exponentially reduce the contribution of history data. The problem of multi-dimensional 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 methods is called as Scatter/Gather methods. This method uses document clustering as its primitive operation. 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 needs filtering and quick summarizations. In those methods, It often needs an effective and efficient methods. Tremendous web crawlers continuously harvest hundreds of thousands of web pages on the web, which is subsequently summarized by human effort. In various e-commerce applications, large volumes of transactions are processed on the World Wide Web (WWW). Those 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. Whenever volume of crawls is significant, it is not possible to achieve such goal by human efforts. Applications where 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 topic modeling & event tracking. They show 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

Volume : 4 | Issue : 7 | July 2015 • ISSN No 2277 - 8179

hybrid scheme for topics modeling over documents streams. It provided a good trade-off between accuracy and speed while performing unsupervised learning over a huge volume of text. By comparing the performance of different off-line topic modeling algorithms and proposed an on-line vMF algorithm that outperforms on-line versions of LDA and DCM in efficiency and performance.

RESEARCH METHODOLOGY

The main objective of this research is to determine a clustering in which the NLP processed data and 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 during the implementation of clustering, and clustering accuracy level. It also measures 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 enhanced the quality of the mining process in a way which would be more meaningful for the users and would be application sensitive.

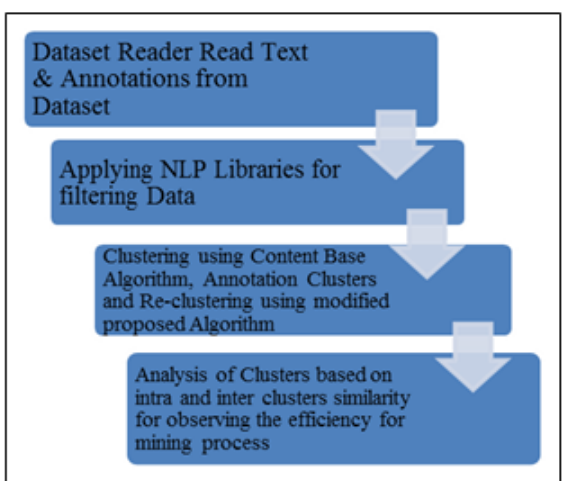

Figure 1 Flow of Operation Proposed Approach

We have used NLP processed COATES algorithm by using side information for text clustering, which indicate the fact that algorithm is Content and Auxiliary attribute based Text clustering algorithm. In this method, an input is number of clusters k to the COATES algorithm. In COATES algorithm text clustering idea is used. Therefore for this purpose, we are using the algorithm explained in [1] and for calculating the centroid we used k-means algorithm. 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 for the second phase.

Next phase starts with initial groups, and then it reconstructs the clusters iteratively by using of both NLP processed text content and Side information (auxiliary annotations). These iterations are referred 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. When the input datasets are provided than first the NLP libraries are applied on it which will eliminate the unwanted text from the datasets. After that the above said procedure of clustering is applied. The resultant clusters are comparing according to their intra and inter clusters differences which clearly show the efficiency of proposed work.

CONCLUSION

In this Paper, We have proposed an approach for clustering the text data based on NLP processed side information which upgraded the quality of the mining process in a meaningful way required by the user. 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 as compare to previously used methods.

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. Kemal U. Koche for his guidance and constant supervision as well as for providing necessary information regarding this approach.

REFERENCE

| [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 documentclustering 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 approachfor clustering of linked document collections," in Proc. CIKM | [7] A. Banerjee and S. Basu, "Topic models over text streams: A studyof batch and online unsupervised learning," in Proc. SDM Conf.,2007, pp. 437–442. | [8] H. Schutze and C. Silverstein, "Projections for efficient documentclustering," in Proc. ACM SIGIR Conf., New York, NY, USA, 1997, | [9] S. Guha, R. Rastogi, and K. Shim, "CURE: An efficient clusteringalgorithm for large databases," in Proc. ACM SIGMOD Conf., New | [10] R. Ng and J. Han, "Efficient and effective clustering methods forspatial 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.