International Journal of Advance Research in Science and Engineering
Vol. No.6, Issue No. 01 , January 2017
www.ijarse.com
IJARSE
ISSN (O) 2319 - 8354
ISSN (P) 2319 - 8346

SURVEY ON ENHANCING QUALITY OF TEXT CLUSTERING BASED ON SIDE INFORMATION

Sushama Pawar, Pankaj Vanwari

Vidyalankar Polytechnic, Wadala Mumbai (India)

Vidyalankar Institute of technology, Wadala Mumbai (India)

ABSTRACT

Clustering is a most widely studied data mining problem in the text mining. This problem finds various applications in classification, document organization, collaborative filtering and indexing. Information is collected from document in large quantity and present in the form of text. Data is not available in the pure form text form, which contains a lot of Side Information that can be different kinds of link present in the document, user-access behavior, and document source information from web - logs or other non-textual attributes. These attributes may be containing and available large amount of information for clustering purposes. Though, it is difficult to determine the relative information, when information contains noisy data. In such situation, it will be insecure to integrate this side-information into the mining process, because it adds noise to the mining process or improve the quality of the illustration for the mining process. An ethical way is required to perform the mining process, and to maximize the advantages of available side information. Therefore, it is suggested to design an efficient algorithm which makes combination of classical portioning algorithm with probabilistic models in order to create an effective clustering approach. Afterwards, extension to the classification problem is also shown.

Index Terms—Clustering, Information Retrieval, Side information, Text Mining.

I. INTRODUCTIoN

The problem of text clustering emerged in the context of many domain mainly application domains such as the web page, social networks sites, and other digital collections/contents. Additional information is known as Side Information or data which describe other data that is available with text document in various texts mining application. Links present in the document, Document source information, and other non-textual attributes which are contained in the document and web logs are the also known as different kind of side information. In this case, it is risky to combine side-information into the mining process because it may improve the standard of the depiction for the mining process or it can append noise in the process/system. Thus, there should be a proper system for mining process so that it will maximize their advantages by using side information. Therefore, it is suggested to design an efficient algorithm which is combination of classical portioning algorithm (i.e. bisecting K-means) with probabilistic models that create an approach which is effective for clustering purpose.

A. Information Retrieval

Information retrieval (IR) is the activity of obtaining information resources relevant to an information need from a collection of information resources. Automated information retrieval systems are used to reduce what has been called "information overload". Web search engines are the most visible IR applications. An information retrieval process begins when a user enters a query into the system. Queries are formal statements of information needs, for example search strings in web search engines.

B. Text Mining

The process of obtaining high-quality information from text is normally referred as Text mining. Text mining involves the process of the input text structuring, patterns obtaining within the structured data, and finally evaluation/ranking and understanding of the output. The main job of text mining typically includes categorization of text, text clustering, extraction ideas/concept, summarization of document, and entity relation modelling. Whereas main application of text mining is to examine a set of documents written in a natural language and used these set of document for predictive classification purposes or grouping in different clusters and populate a database or search index with the information extracted.

C. Side Information

A massive amount of side information is also associated along with the main documents in various application domains. Text documents typically occur again and again in the context of a diversity of applications that is required and useful for clustering process, large amount of information is available for clustering process (e.g. database attributes or meta-data). Following are some examples of side information:

- In a web application we track user access behaviour of web documents; the user-access behaviour is captured in the form of web logs. Quality of the mining process can be enhancing by using users web logs, because the logs can often pick up subtle relation in content, which cannot be collected alone by the raw text.
- Many text documents contain links associated with them, which can also be considered as side information. Such attributes may often provide intuition about the correlations among documents in such way that which cannot be easily obtainable from raw data.
- Many web documents have massive amount of data associated with main document such as ownership, location, or even temporal information etc. is also important and informative for clustering purpose in mining purposes.

D. Clustering

Cluster analysis or clustering process is the task of grouping similar set of objects in one same group called as cluster, and these similar objects are more similar to each other than to those in other clusters. It is a main job of exploratory data mining, and a common technique for statistical study of data which is used in many fields. For clustering process various algorithms are used which is actually significantly differ in their in conception. Clustering can be prepared for the multi-objective optimization problem. Clustering algorithm and parameter settings are selected by individual data set and intended use of the results.

II. LITERATURE REVIEW

A lot of work has been carried out and studied widely in the database and statistics literature in the context of a wide variety of data mining tasks. The problem of clustering has also been studied thoroughly in the context of text data to forming groups of similar object. A survey of text clustering methods may be found in. One of the most well-known techniques for text clustering is the scatter gather techniques for text clustering is pro-posed in [1]. Clustering is applied on text as well as additional attributes by using COATES algorithm and classification methods over many baseline techniques on real and scientific datasets. However, such algorithms notably increase the time and space complexity without any significant result in the output [1], [2]. For proper clustering of text data, Natural Language Processing techniques are used for proper clustering which increase the complexity further. The amount of data to be processed, on the internet is nearly too infinite.

Paper presented by Douglass Michael Steinbach George KarypisVipin Kumar demonstrates [3] that results of some common document clustering techniques are presented with the help of an experimental study. Here two main clustering techniques are compared that agglomerative hierarchical clustering and K-means. Hierarchical clustering is gives the better quality clustering, but it has quadratic time complexity. Whereas K-means and its variants have a linear time complexity but are thought to produce inferior clusters. Many time combination of K-means and agglomerative hierarchical approaches is used together to get the best result. Here, the results stated that the bisecting K-means clustering method is much better than the standard K means approach and as good or better than the hierarchical approaches that we tested for a variety of cluster evaluation metrics. An explanation for these results that is based on an analysis of the specifics of the clustering algorithms and the nature of document data is proposed here.

Paper presented by Douglass R. Cutting, David R. Karger, Jan O. Pedersen, John W. Tukey demonstrates [4] that for information retrieval, document clustering has not been well used. There are two main classes for its objection: first, for hug collaboration clustering is very slow and second, with the help of clustering information retrieval is not improved. Such problems are arriving when clustering is used to improve conventional method. Document clustering is given as primary operation in document browsing technique. Fast clustering algorithms are also presented which support this interactive browsing paradigm.

C. C. Aggarwal and C.-X.Zhai [5] have given an overview of all the methods for text clustering and text classification. They have defined text-specific algorithms for document representation and processing.

Tian Xia and Yanmei Chai [6] have used the Term Frequency Inverse Document Frequency (TF-IDF) to find out which words in the collection of documents are important to be used in query based on the weights assigned to words using TF-IDF.

M. Steinbach, G. Karypis, and V. Kumar [7] have compared different classification algorithms. They have come out with the conclusion that *K*-means clustering outperforms agglomerative and hierarchical clustering techniques.

Yiming Yang, Jan O. Perdersen [8] have presented different feature selection methods. They have found out that Information gain(IG) performs quite well by identifying unique features. IG is used with k-nearest neighbor classification algorithm on Reuter-22173 dataset which improved classification accuracy.

Ryan Prescott Adams, George E. Dahl and Iain Murray [9] have proposed a probabilistic matrix factorization model which makes the use of Gaussian process priors for incorporating side information. Authors have applied this method to estimate the scores of basketball games. The side information used here is venue and date of the game.

P. Domingos and M. J. Pazzani [10] have tested the optimality of the Bayesian classifier and verified that it performs quite well in many domains. They showed that Bayesian classifier has advantages in terms of simplicity, classification speed and is better classifier when the data size is small.

Paper presented by S. Zhong demonstrates [11] that clustering data streams has been a new research topic, also used by many real data mining applications, and has attracted a lot of research observation. However, there is not much work is carried on clustering of high-dimensional text streaming data. This paper combines an effective online spherical k-means algorithm with an existing scalable clustering algorithm to get desire result and achieve fast and adaptive clustering of text streams. The spherical k-means algorithm is modified and it becomes the OSKM algorithm, with the help of online update based on the well-known. It has been shown to be as efficient as SPKM, but much superior in clustering quality. The scalable clustering algorithm was previously developed for very large number of data bases which is not possible to that accommodate into a limited memory also not possible to read/scan multiple times. To make the proposed clustering algorithm adaptive to data streams, a forgetting factor is introduced here that applies exponential decay to the importance of history data. The older a set of text documents, the less weight they carry. The experimental results demonstrate the efficiency of the proposed algorithm and reveal an intuitive and an interesting fact for clustering text streams—one needs to forget to be adaptive.

III. CONCLUSION

In this paper, different methods are discussed for use of side information for mining text data. Side information may be presented in many forms of text database which are used to enhance the clustering process. Iterative portioning technique is combined with a estimation process to design the clustering method which gives the importance of different kinds of side information. This general method is used to design both clustering and classification algorithms. The process of text mining and retrieving information can be greatly enhanced by using an approach that can be effective as well as fast. Computer science is a field which builds and iterates over previously done work to make it more effective.

REFERENCES

[1] Charu C. Aggarwal, Yuchen Zhao, Philip S.Yu, “On the Use of Side Information for Mining Text Data” IEEE,2014.

[2] Vishal Gupta, Gurpreet S. Lehal, A Survey of Text Mining Techniques and Applications , in Journal of Emerging Technologies in Web Intelligence, Vol. 1, No. 1, August 2009, pp.60-76

[3] M. Steinbach, G. Karypis, and V. Kumar, ―A comparison of document clustering techniques,‖ in *Proc.*

Text Mining Workshop KDD,2000, pp. 109–110.

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

[6] Tian Xia, YanmeiChai , "An Improvement to TF-IDF: Term Distribution based Term Weight Algorithm" .

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

[8] Yiming Yang, Jan O. Perdersen "A comparative study of feature selection In text categorization"

[9] Ryan Prescott Adams, George E. Dahl, Iain Murray, "Incorporating Side Information in Probabilistic Matrix Factorization with Gaussian Processes", arXivpreprint , 2010.

[10] P. Domingos and M. J. Pazzani, "On the optimality of the simple Bayesian classifier under zero-one loss," Mach. Learn., vol. 29, no. 2–3, pp. 103–130, 1997.

[11] S. Zhong, ―Efficient streaming text clustering,‖ *Neural Netw.*,vol. 18, no. 5–6, pp. 790–798, 2005.