International Journal of Science and Research (IJSR)
ISSN (Online): 2319-7064
Impact Factor (2012): 3.358

Text Clustering and Classification on the Use of Side Information

Shilpa S. Raut[1], Prof. V. B. Maral[2]

Pune University, India

Abstract: ***Side-information is present with the text document in many text mining applications. An user-access behavior from web logs, or other non-textual attributes embedded into the text document, the links in the document, document provenance information etc are nothing but side information. These attributes contains a vast amount of information for clustering purposes. But it is difficult to estimate the relative importance when some information is noisy. In that case, it will be risky to incorporate side-information into mining process as there is possibility that it will increase the quality of the representation for the mining process or may add a noise to process. Thus a proper way to carry out the mining process is needed such that it will maximize the advantages form using side information. So in this project, an algorithm is designed, in order to give an effective clustering algorithm. This algorithm combines classical partitioning algorithms with probabilistic models. We then show how to extend the approach to the classification problem.***

Keywords: clustering, classifiers information, text mining, text collection, clustering methods

1. Introduction

An issue of text clustering arises in the context of many application domains like as the web, social networks and other digital collections. Scalable and effective mining algorithms are invented because of rapidly increasing amount of text data regarding to large online collections. Vast research and work is on the problem of clustering in text collection [1] [2].

But the work done is designed for the issue of pure text clustering when other kinds of attributes are absent. In many applications side information is associated along with the document. Few examples of such side-information are;

- In an application in which we track user access behavior of web documents, the user-access behavior may be captured in the form of web logs. For each document, the 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 to the user, and also application-sensitive. This is because the logs can often pick up subtle correlations in content, which cannot be picked up by the raw text alone.
- Many text documents contain links among them, which can also be treated as attributes. Such links contain a lot of useful information for mining purposes. As in the previous case, such attributes may often provide insights about the correlations among documents in a way which may not be easily accessible from raw content.

For upgrading the superiority of the clustering process the method of side information is useful, however, sometimes it may be risky approach when the side information is contained noisy data. At this case the method is decreased the quality of the mining process. Since, there is a need to develop the technique which manages the consistency of the process of clustering of the side information with that of the text context, in the proposed work we focus on this work.

2. The Coates Algorithm

We have designed COATES Algorithm for text clustering with side-information.

The name COATES is given by following way. It is a algorithm which corresponds to the fact that it is a COntent and Auxiliary attribute based Text cluStering algorithm.

Input: the number of clusters k
Assumption:
1) Stop-words have been removed
2) Stemming has been performed to improve the discriminatory power of the attributes.

The algorithm requires two phases:

- **Initialization**: Lightweight initialization is used. It is tender for clustering approach without any side information. We have used this algorithm because it is very efficient and provides reasonable initial starting point. In 1st phase centroid and partitions are outputs. This phase uses only text. No auxiliary information is used. Main aim of this initialization phase to construct an initialization and providing a good starting point for the clustering process based on text content.

- **Main phase**: Output of initial phase is an input for main phase. Main phase starts with the initial groups. Then clusters are used iteratively by using both the text content and the auxiliary information. As it uses alternating iterations it help to improve the quality of clustering. Content iterations and Auxiliary iterations are two main types of iterations. Combination of these two is called as major iteration. Each major iteration has two minor iterations corresponding to the auxiliary and text-based methods respectively.

The overall algorithm makes use of alternating minor iterations of content-based and auxiliary attribute-based clustering. These phases are called as content-based and

Licensed Under Creative Commons Attribution CC BY

auxiliary attribute-based iterations respectively. In different iterations set of seed centroid is refined. In every content-based phase a document is assigned to its closest seed centroid by using a text similarity function.

In every auxiliary phase, probabilistic model is created. It 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. It examines the coherence of the text clustering with the side information.

3. Literature Survey

Database community has studied lots about the problem of text-clustering [3], [4], and [5]. In [3] they represent the novel algorithm termed as CURE which is more robust to outliers, and identifiers clusters having non spherical shape and variance in size. In [4] proposed the method CLARANS which is based on randomized search. They also developed two spatial data mining algorithm. In [5] proposed a method termed as BIRCH, which demonstrate especially for very large databases. Scalable clustering of multidimensional data of different types is discussed in [6], [7], and [5]. Scatter-gather technique is the most popular technique for text-clustering [8].It uses a combination of agglomerative and partitional clustering. [9], [10] stated co-clustering methods for text data.[11] studied An Expectation Maximization (EM) method for text clustering. Matrix-factorization techniques for text clustering are stated in [12]. In this technique words from the document based on their relevance to the clustering process are selected and to refine the clusters an iterative EM method is used. Topic-modeling, event tracking, and text-categorization are the areas which are related to text clustering [13]. In [14], [15], [16] focus on the problem of text clustering. The method discussed in the above is focus on the pure text data, these method does not work for the text data which united with the other form of data.

4. Conclusion

Methods for mining text data with the use of side-information are stated in the paper. Side-information or meta-information is present in many forms of databases. It can be used to improve the clustering process. To design the advance clustering method we have combined iterative partitioning technique and a probability estimation process. It computes the importance of different kinds of side-information. For designing the clustering and classification algorithms a general approach is used. COATES Algorithm proves to be very effective.

Reference

[1] C. C. Aggarwal and P. S. Yu, “A framework for clustering massive text and categorical data streams,” inProc. SIAM Conf. Data Mining, 2006, pp. 477–481.
[2] 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.
[3] S. Guha, R. Rastogi, and K. Shim, “CURE: An efficient clustering algorithm for large databases,” inProc. ACM SIGMOD Conf.,New York, NY, USA, 1998, pp. 73–84.
[4] 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.
[5] T. Zhang, R. Ramakrishnan, and M. Livny, “BIRCH: An efficient data clustering method for very large databases,” inProc. ACM SIGMOD Conf., New York, NY, USA, 1996, pp. 103–114.
[6] S. Guha, R. Rastogi, and K. Shim, “CURE: An efficient clustering algorithm for large databases,” inProc. ACM SIGMOD Conf.,New York, NY, USA, 1998, pp. 73–84.
[7] 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.
[8] 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.
[9] Dhillon, “Co-clustering documents and words using bipartite spectral graph partitioning,” inProc. ACM KDD Conf.,NewYork, NY, USA, 2001, pp. 269–274.
[10] Dhillon, S. Mallela, and D. Modha, “Information-theoretic coclustering,” inProc. ACM KDD Conf., New York, NY, USA, 2003,pp. 89–98.
[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.
[12] W. Xu, X. Liu, and Y. Gong, “Document clustering based on nonnegative matrix factorization,” inProc. ACM SIGIR Conf.,New York, NY, USA, 2003, pp. 267–273.
[13] Banerjee and S. Basu, “Topic models over text streams: A study of batch and online unsupervised learning,” inProc. SDM Conf., 2007, pp. 437–442.
[14] C. Aggarwal and P. S. Yu, “A framework for clustering massive text and categorical data streams,” in *Proc. SIAM Conf. Data Mining*, 2006, pp. 477–481.
[15] Q. He, K. Chang, E.-P. Lim, and J. Zhang, “Bursty feature representation for clustering text streams,” in *Proc. SDM Conf.*, 2007, pp. 491–496.
[16] S. Zhong, “Efficient streaming text clustering,” *Neural Netw.*, vol. 18, no. 5–6, pp. 790–798, 2005.

Licensed Under Creative Commons Attribution CC BY