Available Online at www.ijcsmc.com

International Journal of Computer Science and Mobile Computing

A Monthly Journal of Computer Science and Information Technology

ISSN 2320–088X

IJCSMC, Vol. 5, Issue. 1, January 2016, pg.70 – 73

SURVEY on an EFFICIENT APPROACH for TEXT MINING USING SIDE INFORMATION

Kiran V. Gaidhane[1], Leena H. Patil[2], Chandrapal U. Chauhan[3]

[1]Research Scholar, Department of Computer Science and Engineering, Priyadarshini Institute of Engineering and Technology Nagpur (MS), India
[2]Asst.Professor, Department of Computer Science and Engineering, Priyadarshini Institute of Engineering and Technology Nagpur (MS), India
[3]Asst.Professor, Department of Computer Science and Engineering, Priyadarshini Institute of Engineering and Technology Nagpur (MS), India
[1] kirangdhne@gmail.com; [2] lhpatil10@gmail.com; [3] chanderpaul1@gmail.com

Abstract*— *In text mining application, text document having the side information along with it. Such side information may be in the form of links attached with text file, provenance information, title in the document, web logs contain user access behavior or other non-textual attributes contained in the text document. All these attributes having a large amount of information for mining purpose. But it is quite difficult to count the relative importance of this side information when document contains noisy data. The risk is associated with mining process by merging the side information, because it can raise the quality of representation for the mining purpose or can add noise in the process. Thus, for maximizing the advantage make use of side information in the mining procedure. In this paper using the classical partitioning algorithm combines with the probabilistic models for creating the effective clustering approach, extends to the classification problem.

Keywords*— *Data mining, Side information, Clustering, Classification

I. INTRODUCTION

Text mining is a discovery of new and previously unknown information by automatically extracting information from a usually large amount of different unstructured textual data. The main reason for designing the effective text mining algorithm is the increasing amount of textual data in surrounding. In text mining technique, many problems are raised due to several application domains like web information, digital data, and different networks. In these domains, a large amount of side information is associated with the documents. But it

is quite difficult to count the importance of side information because the merging of side information may raise the quality of the mining process. Therefore, will use an approach which ascertains the coherence of clustering characteristics of the side information with the text content. This will magnify the clustering effects of both kinds of data. The core of the approach is to determine the clustering in which the text attributes and side information provide similar hints about the nature of the underlying clustering, and ignores those aspects in which conflicting hints provides. For achieving this goal, using the classical partitioning algorithm in combination with probabilistic model. On the side information probabilistic evaluation process uses the partitioning information for evaluating the different attachments in the document.

The primary goal in this paper is to study the clustering problem; such approach can also be extended in principle to other data mining problems in which auxiliary information is present with text. Here will also extend the approach to the problem of classification, which shows the superior result because incorporation of side information.

II. LITERATURE REVIEW

Paper presented by Michele Ceccarelli, Antonio Maratea demonstrated [4] a metric learning approach as way to improve the classical fuzzy C-means clustering through a two step procedure; first a series of metrics, one for each cluster and then generalized version of fuzzy C-means is executed. In this paper, introduced both the classical fuzzy C-means and semi-supervised C-means (SSFC) algorithm. Optimization in the learning step is done through a heuristic algorithm. By using fuzzy clustering it can improve the performance, and quantified the advantages of using side information through a generalized version of partitioning entropy index. The side information points choose with care, as a wrong choice or a blind generation may not produce a stable solution.

Paper presented by Yuchen Zhao, Philip S. Yu demonstrated a unified distance measure on both link structure and side attributes for clustering, In this paper uses gradient descent algorithm i.e.; Dynamic Multi-Distance Optimization(DMO) for optimizing the weights of graph distance and side information distance metric. Then further introduced a designed statistics Sketch Based Compression framework SGS(C) which consumes with stream progression. *Gssclu* is a clustering method designed for graph stream with side information. The massive size of incoming stream of data and its increasing nature, the data must be stored in hard disk to avoid the out of memory problem. Sometimes side information are quite noisy, thus assigning arbitrary weights to links and side attribute may even degrade the clustering quality.

Paper presented by authors Eric P. Xing, Andrew Y. Ng, Michael I. Jorden and Stuart Russell demonstrated the distance metric learning used to significantly improve the clustering performance. In this paper, having the example of distance metric learned on artificial data, and consider four algorithms for clustering that are K-means using the default Euclidean metric, constrained K-means, K-means + metric and constrained K-means + metric using the distance metric. As the result of algorithm shows, K-means and constrained K-Means failed to find good clustering. But by first learning a distance metric and then clustering according to the true clusters from each other.

Paper presented by Charu C. Aggarwal, Yuchen Zhao, Philip S. Yu demonstrated [7] the content based clustering and extend it to the classification problem by using supervised K-means approach. In this paper a COntent and Auxiliary attribute based TExt clustering (COATES) algorithm is used for clustering the data and COntent and auxiLiary aTtribute based Classification (COLT) algorithm is used for classification, by which the level of efficiency increased but they used only preprocessed data for clustering and classification.

III. PROPOSED WORK

Here will use the Hierarchical clustering which builds a cluster hierarchy or, in other words, a tree of clusters, also known as a dendogram. Every cluster node contains child clusters; sibling clusters partition the points covered by their common parent. Such an approach allows exploring data on different levels of granularity. Hierarchical clustering methods are categorized into agglomerative (bottom - up) and divisive (top - down).An agglomerative clustering starts with one point clusters and recursively merges two or more most appropriate clusters. A divisive clustering starts with one cluster of all data points and recursively splits the most appropriate cluster. The process continues until a stopping criterion (frequently, the requested number k of clusters) is achieved. It having the advantages are, embedded flexibility regarding the level of granularity, ease of handling of any forms of similarity or distance, consequently, applicability to any attribute types.

The Content and Auxiliary attribute based Text clustering (COATES) algorithm used for purifying the clusters with the help of side information. The COATES algorithm having two phases in which first phase carried out with contents only and the second phase carried out with both content and side information. Later, from incoming text document data objects will classify on the basis of K-Nearest Neighbor. Classification may refer to categorization, the process in which objects are recognized, and understood. It is a data mining technique used to predict group membership for data instances. K-NN is a non-parametric method used for classification and regression. In k-NN classification the output is a class membership. An object is classified by votes of its neighbor, with the object being assigned to the class most common among its k nearest neighbors. By using this method the document later comes need not go through the total clustering process.

IV. CONCLUSION

In this paper will give the brief introduction about the broad field of document clustering and classification. The techniques which are used for clustering like hierarchical and classification like K-Nearest Neighbour algorithm. This paper also presented methods for mining text data with the use of side-information. Many forms of text databases contain a large amount of side information or meta-information, which may be used in order to improve the clustering process. In order to design the clustering method, combination of an iterative partitioning technique with a probability estimation process which computes the importance of different kinds of side information takes place. COATES approach can greatly enhance the quality of text clustering while maintaining a high level of efficiency.

REFERENCES

[1] 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.

[2] W. Xu, X. Liu, and Y. Gong, “Document clustering based on nonnegative matrix factorization,” *inProc. ACM SIGIR Conf.*,NewYork, NY, USA, 2003, pp. 267–273.

[3] 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.

[4] Michele Ceccarelli and Antonio Maratea,”Improving fuzzy clustering of biological data bye metric learning with side information”,*Elsevier*,2007.

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

[6] G. P. C. Fung, J. X. Yu, and H. Lu, "Classifying text streams in the presence of concept drifts," *inProc. PAKDD Conf.*, Sydney, NSW, Australia, 2004, pp. 373–383.

[7] C. C. Aggarwal, Yuchen Zhao and Philip S. Yu, "On the Use of Side Information for Mining Text Data," *IEEE trans.knowlege and data engineering*, vol. 26, no. 6, june 2014.

[8] Q. He, K. Chang, E.-P. Lim, and J. Zhang, "Bursty feature representation for clustering text streams," *inProc. SDM Conf.*, 2007,pp. 491–496.

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

[10] C. C. Aggarwal and H. Wang, Managing and Mining Graph Data. New York, NY, USA: *Springer*, 2010.

[11] C.C.Aggarwal, Social Network Data Analytics.NewYork,NY,USA:*Springer*, 2011.

[12] R. Angelova and S. Siersdorfer, "A neighborhood-based approach for clustering of linked document collections," *in Proc. CIKMConf.*, New York, NY, USA, 2006, pp. 778–779.

[13] J. Chang and D. Blei, "Relational topic models for document networks," *inProc. AISTASIS, Clearwater, FL*, USA, 2009, pp. 81–88.

[14] Y. Sun, J. Han, J. Gao, and Y. Yu, "iTopicModel: Information network integrated topic modeling," *inProc. ICDM Conf.,Miami,FL*,USA, 2009, pp. 493–502.

[15] C. C. Aggarwal and C.-X. Zhai, Mining Text Data, *Springer*, 2012.

[16] Y. Zhou, H. Cheng, and J. X. Yu, "Graph clustering based on structural/attribute similarities," *PVLDB*, vol. 2, no. 1, pp. 718–729,2009.

