ISSN: 2277-9655
Impact Factor: 4.116

IJESRT

INTERNATIONAL JOURNAL OF ENGINEERING SCIENCES & RESEARCH TECHNOLOGY

AN EFFICIENT APPROACH FOR TEXT MINING USING SIDE INFORMATION

Kiran V. Gaidhane*, Prof. L. H. Patil, Prof. C. U. Chouhan

DOI: 10.5281/zenodo.58632

ABSTRACT

Nowadays, in many text mining applications, information is present in the form of text documents. Text document contains various types of information such as side information or metadata. The different types of information such as document provenance information, title of the document, links in the document, user-access behavior from web logs, or other non-textual attributes treated as side information contained into the text document. Such attributes contains a large amount of information for clustering purposes. It is difficult to estimate the importance of this side-information when text document contains some of the information is noisy. In such cases, to avoid the low quality of mining process we need a principled way to perform the text mining, to maximize the advantages from using this side information. Conformation to that, this paper represents solution to the use of side information for clustering by hierarchical algorithm which then extends to the classification problem on real data sets.

KEYWORDS: Text mining, Side information, Preprocessing, Clustering, Classification.

INTRODUCTION

Text mining is a new area of computer science which strongly connected with natural language processing, data mining, information retrieval, knowledge management and machine learning. Text mining is a process of extracting useful information from unstructured or semi-structured textual data by identification and exploration of interesting patterns from different sources. Each document contains side information along with it. Side information may be in the form of web logs contain Meta information which give information about the browsing behavior of different users. A lot of text document having connections among them are also called as attributes, such links posses a lot of information for mining purpose. Sometimes it is difficult to count the importance of side information because the merging of side information may raise the quality of mining process or may decrease the quality because of noisy data in the documents. At that time it can actually degrade the quality of mining process.

The clustering is the process of combining the set of objects in such a way that objects in the same group are more similar than that of different group of objects, this group of objects are called as clusters. We proposed the method which shows the advantages of using side information. We will also present the technique to extend to the classification problem.

LITERATURE REVIEW

Various methods in text mining are best on statistical study of objects in a number of documents. Thus, the text mining model should signify terms that capture the semantics of text. In this case, the mining model can catches term that shows the concepts of the sentence. There are lots of clustering problems explained by database community [1][3][8]. A general survey of clustering algorithms found in[4]. In text mining various methods are based on the statistical study of a word or term. This study gives term frequency to show the significance of the term inside a document. One of the term having more meaning of its sentences than the other when other terms have the same frequency in their documents. An Expectation Maximization (EM) method for text clustering proposed in [5].

Paper presented by auther E. P.Xing, A. Y. Ng, M. I. Jorden and S. Russell in [6] demonstrated the distance learning used to significantly improve the clustering problems. Here, they learned a distance metric using similarity

information and cluster data using that metric. For that they consider four algorithms that are k-means algorithm using default Euclidean metric, constrained k-means, k-means+metric, constrained k-means+metric.

Authors Y. Zhao, P. S. Yu in [13] demonstrated on graph stream clustering with side information, in which a unified distance measure on both link structure and side attributes for clustering, In this paper introduced gradient descent algorithm i.e.; Dynamic Multi-Distance Optimization(DMO) for optimization of weights of graph distance and side information distance metric. They designed statistics Sketch Based Compression framework SGS(C) which consumes with stream progression. Then further designed clustering method Gssclu for graph stream with side information. The massive size of incoming stream of data and its increasing nature, the data has to stored in the 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 C. C. Aggarwal, Y. Zhao and P. S. Yu demonstrates [10] the content based clustering by using supervised K-means approach and then extend it to the classification problem. Each time the number of cluster (K) has to define and centroids are must choose at each iteration.

M. Ceccarelli and A. Marateain [7] demonstrated a metric learning approach to improve classical fuzzy C-means clustering in a two step procedure, first with Euclidean metric and the second with learned metric. The classical fuzzy C- means and semi-supervised C-means [SSFC] algorithm both are introduced fuzzy clustering can improved the performance and quantifies the advantages of using side information through a generalize version of partitioning entropy index.

PROPOSED WORK

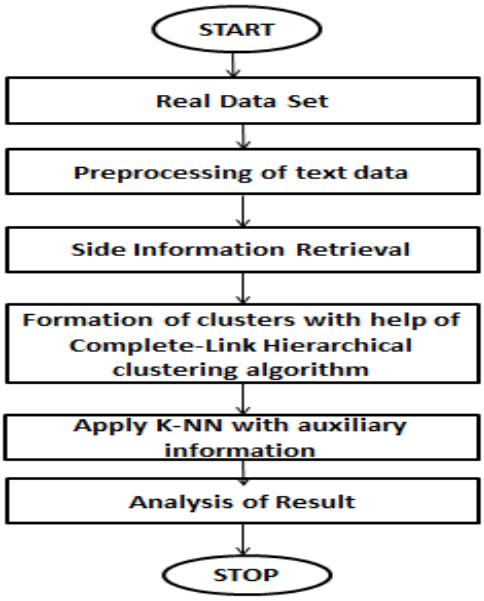

Fig 1: Flowchart for cluster formation with the help of Side information.

Preprocessing of Textual Data

Before applying the text mining process, preprocessing of data is important task during text mining process. Text document having large amount of words on which different text mining methods can be applied. The special methods such as filtering and stemming are applied on the set of keywords from different resource. Symbols and numerical values are removed from the set of words. Words which do not provide relevant information like prepositions, conjunctions, articles etc., removed by filtering methods, here stop word filtering is used as standard filtering method. The additional benefit of elimination of stop word is in the reduction of indexing size. The remaining set of words is use as input to the stemming algorithm. The porter stemming algorithm is one of the most popular stemming method,

having six steps and within each steps, rules are applied until one of them passes the conditions. The suffix is removed if a rule is accepted and then performs next step. After end of the sixth step the resultant stem is returned.

Porter stemmer produces less error rate and best output as compared to other stemmers. The list of stemmed word is used to partition the side information by calculating the term frequency of each word.

Information Retrieval

As we know, documents having side information along with it, to extract this side information from the set of keywords, we perform information retrieval. In information retrieval TF-IDF, short for term, Term Frequency-Inverse Document Frequency, is a numerical statistic that is intended to reflect how important a word is to a document in a collection. It is often used as a weighting factor in information retrieval.

The TF is nothing but the number of words occurs in a document and IDF is the number of times a word occurred in different documents. We use TF for extracting side information to its randomly selected value. Here, we introduced Jaccard Coefficient Similarity algorithm to find the distance between two entities. It is commonly used measure of overlap of two sets, it always assign a number 0 and 1.After calculating distance function between two sets, the values are considered to form the clusters.

Clustering algorithm

The Complete-link Hierarchical clustering is one of several methods of agglomerative hierarchical clustering. Complete Link clustering avoids a drawback, called Chaining phenomenon. Complete link tends to find compact clusters of approximately equal diameters.

Mathematically, the Complete Linkage function-the distance $D(X,Y)$ between clusters and – is described by the following expression:

$$D(X,Y) = \max_{x \in X, y \in Y} d(x,y)$$

- Where $d(x,y)$ is the distance between element $x \in X$ and $y \in Y$.
- x and y are two sets of clusters.

At the beginning of the process, each element is in a cluster of its own. The clusters are then sequentially combined into larger clusters until all elements end up being in the same cluster. At each step, two clusters are separated by the shortest distance, are combined.

In Compete-Linkage Clustering, the link between two clusters contains all element pairs, and the distance between clusters equals the distance between those two elements that are farthest away from each other. The shortest of these links that remains at any step cause the fusion of two clusters whose elements are involved.

Graphical User Interface

Fig 2: Preprocessing of Data Set

In this section, we will preprocess the textual data by applying filtering and stemming methods, the output shown in Fig. 2. Data set is an input to the methods. Symbols, numbers, stop words are removed from the number of files, list

ISSN: 2277-9655
Impact Factor: 4.116

of stopwords are maintained in a separate dictionary. Symbols and numbers which will extract from the text documents maintained in the source code. If they are present at the time of scan then that are removed.

Here, we use a porter stemmer in which the stem need not be identical to the morphological root of the word; it is usually sufficient that related words map to the same step even if this stem is not in itself a valid root. A porter stemmer having 60 rules in six steps with no recursion.

A stemming algorithm reduces the words "fishing", "fished" and "fisher" to the root word "fish". On the other hand, "argue", "argued", "argues", "arguing", "argus" reduce to the stem "argu" illustrating the case where the stem is not itself a word or root but "argument" and "arguments" reduce to the stem "argument" .

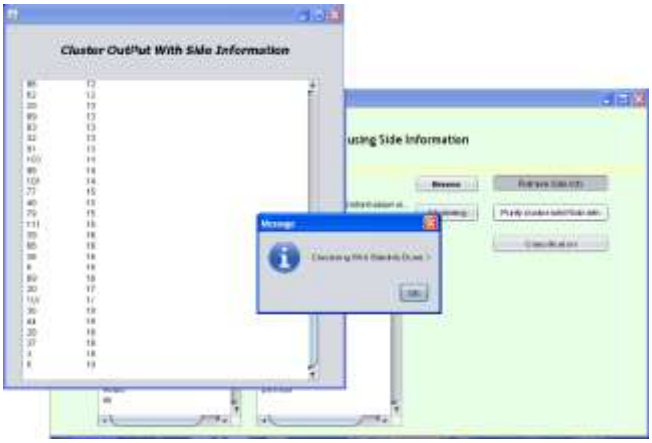

Fig 3: Formation of Clusters using side information.

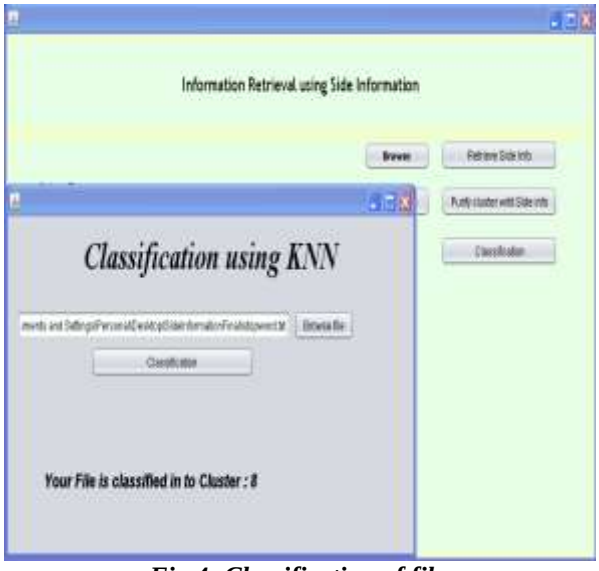

Fig 4: Classification of file

The words after filtering and stemming are used as input to the clustering process. Figure 3 shows the cluster formation by using only keywords. For this, term frequency of each word is calculated. If the term frequency is greater than or equals to randomly selected frequency that is 4 then that words are considered as Keywords and other words are declared as Side Information. The records for Keywords and Side Information are maintained in the database.
We used Jaccard Coefficient-Similarity algorithm for calculating the distance function.Each keyword having the a value between 0 and 1 which then desides the conentss of clusters

The clustering process with the help of side information is similar as that of formation of clusters with only keywords. But, for cluster formation using side information, we use both keywords and side information. Clusters are formed as output as cluster forms using only keywords. Later, we will introduce the classification method to group membership for data instances. We will extend the clustering algorithm to the K-Nearest Neighbors classification algorithm, the output is a cluster membership. The instance is classifies by a majority vote of its neighbors, with the instance being assigned to the class most common among its K nearest neighbors.

CONCLUSION

In this paper, we studied the methods for mining text data with the use of side information. We also studied the problem of improving data clustering by using both instance and attribute level side information. In order to design the clustering method, we proposed unsupervised algorithm in which we don't have to define the number of clusters and clusters are created accordingly, used the real data sets. We performed the classification to classify the instance to a specific cluster.

REFERENCES

[1] S. Guha, R. Rastogi, and K. Shim, "CURE: An Efficient Clustering Algorithm for Large Databases", in ACM SIGMOD Conf., pp. 73-84, 1998.
[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, pp. 318–329,1992.
[3] T. Zhang, R. Ramkrishnan, and M. Livny, "BIRCH: An Efficient Data Clustering method for Very Large Databases", in ACM SIGMOD Conf., pp. 103-114, 1996.
[4] C. C. Aggarwal and C.-X. Zhai, Mining Text Data, Springer, 2012.
[5] T. Liu, S. Liu, Z. Chen, and W.-Y. Ma, "An evaluation of feature selection for text clustering", in ICML Conf., pp. 488-495, 2003.
[6] E. P. Xing, A. Y. Ng, M. I. Jorden and S. Russell ,"Distance metric learning, with application to clustering with side-information",inAdvances in Neural Information Processing Systems 15, MIT PRESS, PP. 505-512, 2002.
[7] Michele Ceccarelli and Antonio Maratea, , "Improving fuzzy clustering of biological data bye metric learning with side information",Elsevier,2007.
[8] R. NG and J. Han, "Efficient and Effective Clustering methods for Special Data Mining", in VLDB Conf., pp. 144-155, 1994.
[9] M. Steinbach, G. Karypis, and V. Kumar, ―A comparison of document clustering techniques,‖ in Proc. Text Mining Workshop KDD, pp. 109–110, 2000.
[10] 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, 2014.
[11] S. Zhong, "Efficient streaming text clustering," Neural Netw.,vol. 18, no. 5–6, pp.790–798, 2005.
[12] 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, pp. 373–383, 2004.
[13] Y. Zhao, and Philip S. Yu, "On Graph Stream Clustering with Side Information",in Proceedings of the thirteenth SIAM International Conference on Data Mining,2013.
[14] 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, 2004.
[15] Q. He, K. Chang, E.-P. Lim, and J. Zhang, "Bursty feature representation for clustering text streams," inProc. SDM Conf., pp. 491–496, 2007.
[16] C. C. Aggarwal and H. Wang, Managing and Mining Graph Data. New York, NY, USA: Springer, 2010.