ISSN 23499842(Online), Volume 1, Special Issue 1(NCFCSIT 15), March 2015

International Journal of Innovative Trends and Emerging Technologies

MINING TEXT DATA WITH SIDE INFORMATION

[1]K.VALARMATHI, [2]M.RAMESH KUMAR , M.E.,

[1]PG Scholar, Department of CSE, Agni College of Techmology, Chennai, valarmathisekar17@gmail.com

[2]Assistant Professor, Department of CSE, Agni College of Techmology, Chennai, rameshkumar.cse@act.edu.in

Abstract— **In numerous application domains such as the social networks, web, and other digital collections, side-information is obtainable with the text documents. Side-information are of different kinds, may be a document source information, users behavior from the web logs, links in the document, and other non-textual attributes. The attributes are enclosed into the text document and contain a huge amount of information for clustering and classification purposes. The side-information may be very hard to deal, especially when the side information is noisy. In such manner, it was risky to combine side-information to the mining process. So it needs a principled way to perform the mining process, and also to increase the advantages by using this side information. In this paper, designing an algorithm by which combining classical partitioning algorithm with probabilistic model to make an effective clustering and classification approach.**

Index Terms—**data mining, side information, clustering, classification.**

1. INTRODUCTION

Generally, data mining is the process of analyzing data from different perspectives and summarizing it into useful information. Data mining software is one of the analytical tools for analyzing data in large relational databases. It allows users to analyze the data from several different angles or exents, categorize and reduce it. Technically, it is the process of finding patterns among dozens of fields in large databases.

Text mining is the abstraction of data contained in natural language text(NLT). The technique of text mining application is to solve business problems named to be text analytics. The problem of text clustering emerges in the context of several application domains such as the social networks, web, and other digital collections. Rapidly increasing the amount of text data in context of large online collections which led to create an scalable and effective mining algorithm. Large amount of work has been done to overcome the problem of clustering in text collections of database and retrieval of information communities. Therefore, many application has been designed for the problem of pure text classification and clustering, without other kinds of attributes.

The side information is available with many text document may be provenance information which might be informative for mining. Sometimes side-information can be useful for improving the quality of the classification and clustering process, such a risky approach when the side-information is noise. Therefore, it worsen the process of text mining quality.

For an application user access behavior of web documents can be tracked, in the form of web logs. Each and every document, the meta-information corresponds to the user behavior of the different users. Therefore, logs can be used to improve the quality of the mining process. The logs can pick up exact correlations in content; it cannot be pickup by the raw text alone. The main aim of this approach is to determine the coherence between side information and text content, otherwise avoid the text content which does not provide similar hint.

2. RELATED WORK

The text clustering task is to group similar documents to be well-organized. Scatter gather technique is one of the popular known technique for unsupervised approach. D. Cutting, D. Karger, J. Pedersen, and J. Tukey, describe typically the documents is clustered either into disjoint, an intensive partition, else a hierarchical structure, uses a mixture of partitioned and agglomerative unsupervised approach. S. Guha, K. Shim, and R. Rastogi says that cluster which are generated by selecting the points are all scattered well from the cluster after that shrink by a specified fraction toward the center of the cluster.

An automated text categorization problem has been studied widely, in which desire to find the nearest matching for a given text document, the information from earlier existing division to supervise the new creation of a confirmed group of clusters. To perform classification, where cosine measure is used and a hierarchy has been created by continuously applying an agglomerative hierarchical clustering on the centered.

The Meta documents are created by hierarchy. M. Steinbach, G. Karypis, and V. Kumar K-means and agglomerative hierarchical approaches are combined so as to “obtain the best of these words”. The bisecting K-means approach is better than the standard K-means approach and as good or better than the hierarchical approaches that have to be tested for a variety of cluster. T. Liu, Z. Chen, S. Liu, and W.-Y. Ma, gave an empirical evidence to improve the performance and efficiency of text clustering algorithm using that feature selection methods. T. Zhang, R. Ramakrishnan, and M. Livny determine useful patterns in large datasets has attracted noticeable interest recently, and one of the problem which was studied most widely in this area is the description of clusters, or massively populated regions, in a multi-dimensional dataset. Former work does not adequately address the problem of heavy datasets and minimization of 1/0 costs. C. C. Aggarwal and H. Wang mainly focus on the scalable clustering of different types of multidimensional data.

3. SYSTEM DESIGN

In this section depicts the clustering and classification approach in detail. When an input to query as given in the form of text to retrieve suitable cluster with the use of side information. A query can be processed and extract the set of documents from the database, by considering each words as token and remove the presence of stop words(is,was,at,-,and,..) in the input.

Once the group of document is extracted, then analyze the side information for every document. For, each and every document source information is available get that information, then get the meta data for each document and the links associated with the text document .After analysis, we get some set of text documents named as corpus then perform transformation for choosing which are the text document related to input and which are not. By using binary transformation the text document which is related to query is treated as 1 or otherwise treated as 0.

After getting these documents performs stemming in order to reduce index size, the fundamental idea is to index each word in the text document. The document contains set of words how these set of words are related to the input with the count of appearance of word. To perform ranking assess the similarity between various document and there is no predefined classes, so by using set of documents that involve in the group to obtain the document scoring which are closely related to input. Verify this cluster purity by using probability clustering approach.

Once the document term frequency is identified by dividing the number of term t in the document with the total number of words in the document. Based on these such documents are ranked. Similarly when new documents are added to the cluster, the classification needs to be performed. The document which is placed at the top has higher ranking for the given query and also appropriate to each keyword present in the query. This depicts in the System design for mining text data with side information

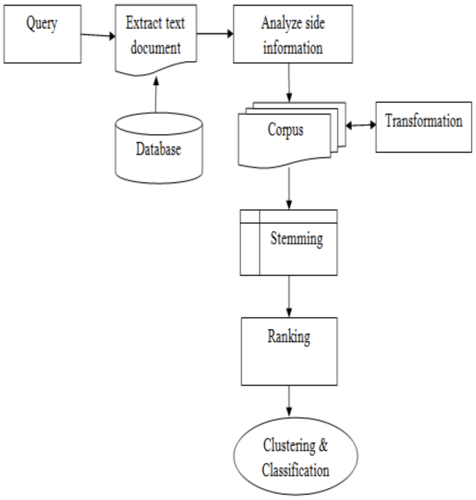


Figure 3.1: System design for mining text data with side information

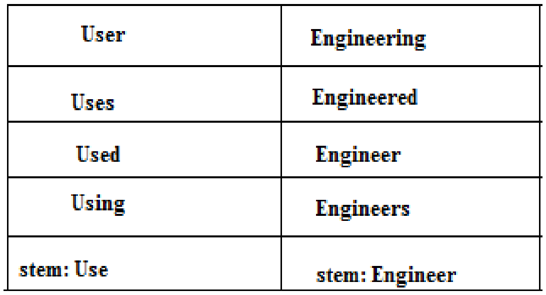
3.1TRANSFORMATION

Raw text document may contain noisy information for mining, in order to improve the clustering purity and classification process by using transformation process in initial set of documents. Sometimes side information is in different types, by using transformation it can be transformed into binary attributes by the use of transformation process. Whereas, to remove noise in text document smoothing process is taken place.

For an example, an attribute may be quantitative, numerical data and categorical data can be discretized into binary attributes. In normalization the attributes ranges are 0 and 1. Within these range the transformation is taken place if it is 1 related to the document or otherwise not related to that. After that aggregate all the document which are related to the query.

3.2 STEMMING

Stemming is a significant feature for an indexing and searching system in mining, with the help of stemming improves the information retrieval process. In order to improve the process by reducing the suffixes and prefixes associated with the root word and also to determine the stem or root word. Then the stem is treated as an index for the document. Not only using stem as index simply but also consider the semantic for the stem then it is treated as an index for text document. Several stemming algorithm is there to identify the semantic of word because sometimes the word can vary but the meaning is same. Stemming system is used to reduce the size of indexing as shown in stemming table. By removing the characters(s,ing,ed,ies,….) associated with the words we get stem word.



User	Engineering
Uses	Engineered
Used	Engineer
Using	Engineers
stem: Use	stem: Engineer

Table 3.1: stemming

3.3 RANKING

Frequency count plays a major role in ranking data, it has to counts the frequency of word occurrence in text document and also determine which are all the document contain that a word. With the help of frequency count compute the importance of document for that word, suppose the word is repeated for more number of times then the word is related to the subject and it play a major role in that document. By using support vector machine measure the distance between the word and the document with the help of cosine similarity method.

Let the document be considered as vector such as (d1, d2, … … , dn). The availability of word is represented in binary if dj= 1 then the respective term j is present in the document. If dj= 0 then the term j is not present in the document. The term frequency is computed as dj= tcj where tci is the number of times the word presented in the document. Let us compute the document frequency by multiplying term frequency dj =tcj*idfj=tcj*log(N/dfj)) where N the total number of documents in the set of cluster and dfj is the number of documents contains word j.

4. PROPOSED ALGORITHM

In order to access every word that are present in a disposed text document, a tokenization method is needed, and a query given is separate into a branch of words by restoring tabs, by eliminating every punctuation marks and some white spaces along with some non-text character. By combining collection of words from all text documents we obtain a dictionary of words named as bag of words. To minimize the size of dictionary by using the concept of stemming to remove the words like conjunctions, articles etc., and also to map the singular word to plural that is depicts as stripping the words in the text document.

4.1 CLUSTERING TECHNIQUES

Clustering or cluster approach is the process of extracting a set of clusters. The objects which are present in the same cluster are more or less similar and the other objects are dissimilar. The task of clustering is relative to data mining. In data mining for upcoming clustering process is done by using already processed text document. Feature selection process is used for an exact document retrieval approach; one of the main features for clustering text document is determining the distance between documents for the query. Once the clusters are generated which is assessed by cluster purity and the cluster quality is measured by statistical measures.

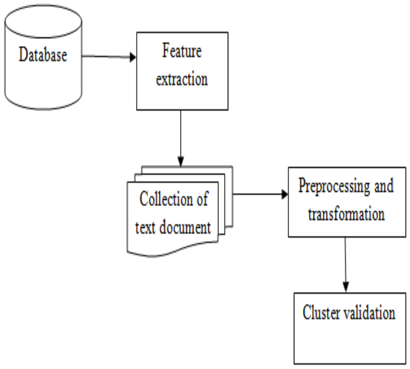


Figure 4.1 Process for clustering

During clustering text document, the most essential thing is to measure document coherence. In order to determine document coherence by using word coherence concept, using k-means clustering cluster the text document which are provide similar hint or coherence otherwise ignore those document in the cluster. The k-means clustering approach is explained by using Figure 4.2: Document for clustering, now select a value for K which is nothing but number of cluster totally or otherwise select k value by taking means, consider it as an initial value for clustering.

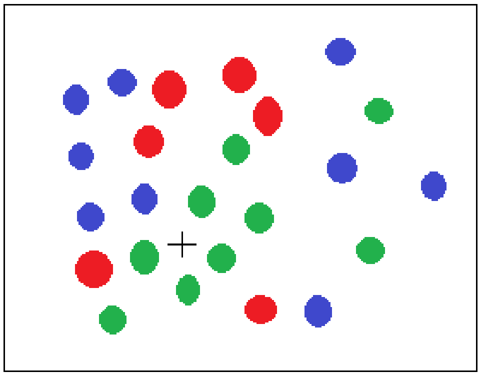


Figure 4.2: Document for clustering

From the k value determine the cluster centroid and measure the distance between the text documents which are closely related to the cluster centroid. The distance measure is repeated for various times until the pure cluster is formed, so cluster centroid is recalculated for every time. At last we obtain the set of cluster by using k-means approach and the clutering process is scalable and effective. Thus the clustering process is possible for large set of text documents.

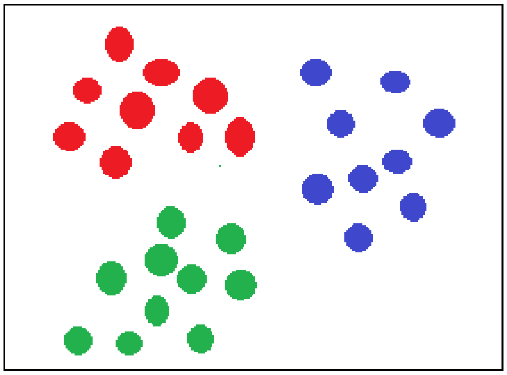


Figure 4.3: Clustered document

4.2 CLASSIFICATION TECHNIQUES

Once the cluster is formed make use this cluster to develop a model for each category. With the help of features available in text document let choose the document which is suitable for which category of cluster when the new document enter into cluster category if that document belongs to more than one category based on likelihood determine which category is suitable for new text document

In supervised clustering set of classes or objects is known earlier. Data preparation is done by using feature selection process and eliminates those text documents which are not related to the index or class label. Supervised approach is to analyze the text document by using available feature in the text document and to create an accurate depicts or model for every class. Once the classification training model is created by classifying all the documents in the database let filter the stop words (but, is, at, was, were, how, an, etc)and also filter the word with low frequency. Classification of text documents into some categories so it is easy to determine and retrieve for its efficient use and effective mining approach.

V. RESULTS AND DISCUSSIONS

In our experiment search engine is made to be most effective with the help of attributes associated with the text document. The process of clustering is carried out in efficient manner by removing stop words in the index and analyzing the side information for the cluster. Next to perform an transformation process for the cluster and stemming to get an effective cluster for the query in that cluster choose the significant text document for the input based on the ranking which has to measure the cosine similarity between the text documents.

By using the weblogs to track the user access behavior, make the information retrieval to be effective and also make the process to be scalable for mining the text document. Even sub links in the document and metadata make this approach to in order to increase the unsupervised process.

5. CONCLUSION

In this paper, I explained about side information which is presented in several application domains in their text documents. Sometimes side-information may contain erroneous information in order to avoid that combining an algorithm with the probabilistic model to show that cluster purity computes the posterior probability for set of cluster document. This approach makes use of side information otherwise called as meta –attributes available in many text document in the databases to be built in effective manner.

While combining algorithm with probabilistic approach computes the significant of side information for the clustering and classification techniques. Application of this project is while searching for text document related to the query in many fields for an example an IT industry, when we get a project to retrieve the text document from their database that are related to the project. Using side information maintains a greater level of efficiency.

6. FUTURE WORK

In addition to this concept some side information may be included to make this clustering and classification approach to be effective. The time frequency and feedback concept to be consider side information for the future upcoming work in this project and use some other algorithm for these processes.

REFERENCES

[1]] Charu C. Aggarwal, Fellow, IEEE Yuchen Zhao, and Philip S. Yu, Fellow “On the Use of Side Information for Mining Text Data”, IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, VOL. 26, NO. 6, JUNE 2014 1415

[2] C. C. Aggarwal and P. S. Yu, “On text clustering with side information,” in Proc. IEEE ICDE Conf., Washington, DC, USA, 2012.

[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] 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. Guha, R. Rastogi, and K. Shim, "CURE: An efficient clustering algorithm for large databases," in Proc. ACM SIGMOD Conf., New York, NY, USA, 1998, pp. 73–84.

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

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

[8] C. C. Aggarwal and C.-X. Zhai, "A survey of text classification algorithms," in Mining Text Data. New York, NY, USA: Springer,2012.

[8] C. 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.

[9] 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.

[10] M. Franz, T. Ward, J. S. McCarley, and W. J. Zhu, "Unsupervised and supervised clustering for topic tracking," in Proc. ACM SIGIR Conf., New York, NY, USA, 2001, pp. 310–317.

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

[12] H. Frigui and O. Nasraoui, "Simultaneous clustering and dynamic keyword weighting for text documents," in Survey of Text Mining, M. Berry, Ed. New York, NY, USA: Springer, 2004, pp. 45–70.

[13] 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.

[14] F. Sebastiani, "Machine learning for automated text categorization," ACM CSUR, vol. 34, no. 1, pp. 1–47, 2002.

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

[16] W. Xu, X. Liu, and Y. Gong, "Document clustering based on nonnegative matrix factorization," in Proc. ACM SIGIR Conf., New York, NY, USA, 2003, pp. 267–273.

[17] Y. Zhao and G. Karypis, "Topic-driven clustering for document datasets," in Proc. SIAM Conf. Data Mining, 2005, pp. 358–369.

[18] "A Brief Survey of Text Mining" Andreas Hotho, Andreas N¨urnberger and Gerhard Paaß,May 13, 2005.

[19] http://www.textminingnews.com/, 2005.

[20] "Natural language processing and text mning", Anne Kao Stephen R.Poteet(Eds),springer,text book.

[21] Nahm, U.Y. and Mooney, R.J. (2002) "Text mining with information extraction." Proc AAAI-2002 Spring Symposium on Mining Answers from Texts and Knowledge Bases. Stanford, CA.

[22] Dan Klein, Sepandar D. Kamvar, and Christopher D. Manning. From instance-level constraints to space-level constraints: Making the most of prior knowledge in data clustering. In Proceedings of the 19th International Conference on Machine Learning, pages 307–314. Morgan Kaufmann Publishers Inc., 2002.

[23] F. Sebastiani, "Machine learning for automated text categorization,"ACM CSUR, vol. 34, no. 1, pp. 1–47, 2002.

[24] C. Silverstein and J. Pedersen, "Almost-constant time clustering of arbitrary corpus sets," in Proc.CM SIGIR Conf., New York, NY, USA, 1997, pp. 60–66.

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

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

[27] W. Xu, X. Liu, and Y. Gong, "Document clustering based on nonnegative matrix factorization," in Proc. ACM SIGIR Conf.,NewYork, NY, USA, 2003, pp. 267–273.

[28] T. Yang, R. Jin, Y. Chi, and S. Zhu, "Combining link and content for community detection: A discriminative approach," in Proc. ACM KDD Conf., New York, NY, USA, 2009, pp. 927–936.

[29] T. Zhang, R. Ramakrishnan, and M. Livny, "BIRCH: An efficient data clustering method for very large databases," in Proc. ACM SIGMOD Conf., New York, NY, USA, 1996, pp. 103–114.

[30] Y. Zhao and G. Karypis, "Topic-driven clustering for document datasets," in Proc. SIAM Conf. Data Mining, 2005, pp. 358–369.

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

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