ABHIYANTRIKI
An International Journal of Engineering & Technology
(AIJET)
Vol. 1, No. 2 (December, 2014) http://www.aijet.in/ eISSN: 2394-627X

Extracting Side Information from Unstructured Data using Data Mining Technique

Puja Awandkar*
Deptt. of Computer Science & Engineering,
G.H.R.A.E.T, Nagpur University, INDIA

Amit Pimpalkar
Deptt. of Computer Science & Engineering,
G.H.R.A.E.T, Nagpur University, INDIA

Abstract

There are many text mining applications which consist of side information along with the text documents. Also there are many web documents which include meta-data in them. This meta-data corresponds to different kinds, such as document provenance information, or the information related to the document's origin. This meta-data may prove to be informative for performing the mining process as this data contains a large amount of information. This large amount of information may be used for performing clustering. However, it may prove to be a difficult process for computing the importance of this side-information, especially when it is noisy. And hence the quality of mining process may decrease. Therefore, we will provide a principled way for performing mining. In this paper, we will focus on how to extract the side-information from the text document.

Keywords*: Text mining, Metadata, Side information, Mining.*

****Author for Correspondence*** puja.awandkar@gmail.com

1. Introduction

Day-by-day the use of digital information is increasing. This information in the digital world is growing to the extent that the finding some of the relevant information from this huge amount of data is becoming quite tedious. This proves to a reason in creating scalable and efficient mining algorithms. The clustering of data in the pure form is done till now. But to manage such large quantity of data we require indexing the data according to the users need. For this we will be using the meta-data that is the side information which is present on almost all of the text documents. A large number of web documents consist of meta-data with them. These meta-data are exactly matching to the various different kinds of attributes such as the origin or other information related to the origin of the document. Data such as location, possession or even temporal information may prove to be informative for mining purposes in other cases. Documents may be linked with user-tags in many network and user-sharing applications. This may also be quite informative. We are making use of the side information for clustering the data for doing effective text mining. The process of deriving high quality information from text is known as text data mining. While this side-information can sometimes prove useful in improving the quality for the clustering process, but when the side-information is noisy it can be a risky approach and the quality of the mining process can actually become worse. Therefore, we will

proceed towards to use an approach which carefully finds the well-organized form of the clustering characteristics of side information with that of text content. The basic approach of the system is that it can form a clustering in which the text attributes along with the side-information provide similar hints about the character of the basic clusters, and at the same time fail to consider those features in which conflicting hints are provided. Also we will use an efficient searching method based on user behavior. With the help of the use of this user behavior search we will get the more relevant searching results and the desired output.

2. Related Work

Pooja Awandkar et al. [1] explain the concept of searching based on user behavior with the help of mining the side-information. They made the use of COATES algorithm for performing clustering. Ting Yuan et al. [2] proposed a new recommendation model named Group-Sparse Matrix Factorization (GSMF). This model integrates various multiple types of user behaviors by performing modeling of the shared and private factors among them. Amit Pimpalkar [3] proposed the system which collects the reviews from various online websites. Their proposed system also does the comparison between two products by taking the help of reviews identified from the online resources which leads to find the best one of it. W. White et al. [4] proposed methods for modeling user's on-task search behavior. These models were used to improve the personalization methods. Data centric views of online social networks are provided by C. C. Aggarwal [5]. Topic-driven clustering for text data method has been proposed by H. Wang et al. [6]. A model of documents and the links between them, the Relational Topic Model (RTM), was proposed by J. Chang et al. [7]. A new topic modeling structure for document networks, which examines both text and structure information for documents was proposed by Y. Sun et al. [8]. The problem of combining link and content analysis for community detection from networked data was considered by T. Yang et al. [9]. The problem with graph clustering based on structural and also attribute similarities were solved by Y. Zhou, et al. [10], though this work is not relevant to the case of usual side-information attributes. They also designed a learning algorithm which adjusts the degree of contributions of various different attributes in the random walk model. The problems of topic modeling with network structure (TMN) were defined by Q. Mei et al. [11]. They proposed method which combines both topic modeling and social network analysis. The proposed model was generalized which can be applied to any kind of text collections with a combination of topics and an associated networks structure. New temporal representations for text streams based on bursty features were introduced by J. Zhang et al. [13]. It was introduced for highlighting the temporally important features present in the text streams. P. S. Yu et al. [14] presented an online approach for clustering the massive text and categorical data streams by using the statistical summarization methodology. They proposed algorithm which can be used for both text and categorical data mining domain. Their experimental results showed that the algorithm was very effective in quickly adapting the temporal variations in the data stream. The issues of naturally structuring linked document collections by using clustering were addressed by R. Angelova et al. [15]. They provided techniques which results in higher cluster purity and better overall accuracy. Fast and adaptive clustering of text streams were studied by S. Zhong [16]. They combine an effective online spherical k-means (OSKM) algorithm with an existing expandable clustering strategy. This was done to achieve fast and adaptive clustering of the text streams. However, all of these methods were designed for the instance of pure text data, and these methods do not work for cases in which text-data is merged with other forms of data. The merits of building text categorization systems were discussed by S. C. Gates et al. [17] by using supervised clustering techniques. They also discussed the new

technique which helps the classifier to distinguish better among closely related clusters. G. P. C. Fung et al. [18] focused on two issues of concept drifts, namely, concept drifts detection and model adaptation in text stream context. They used statistical control for detecting concept drifts, and proposed the new multi-classifier strategy for model adaptation. Y. Gong et al. [19] proposed Matrix-factorization techniques for text clustering. This technique selects words from the document which are based on their application of the clustering process, and also uses an iterative EM method. This method is used in order to refine the clusters. The important differences between two styles of document clustering in the context of Topic Detection and Tracking were investigated by M. Franz et al. [20]. One of the most popular techniques for text-clustering was introduced by D. Cutting et al. [21] the scatter-gather technique, which uses a fusion of agglomerative and partitioned clustering. Scatter-gather technique is particularly helpful in situations where it is difficult or undesirable to specify query formally.

3. Problem Definition

In the existing system clustering was performed by simple methods. No use of side information was done for performing clustering. The formation of clusters was done based on the data given as input to them. Later on we came to know that the side information present in the document can also prove useful and also help for enhancing the clustering techniques. The side information can also contain the data which can be useful for mining purpose. The earlier methods used for clustering were working only for the cases of pure text data. They were not working for the cases in which the text data is merged with other forms of data. There were no methods of performing searching operation in the clusters formed from the text data along with the side information or the auxiliary information.

4. Proposed Work

In the proposed work our objective will be firstly to collect information i.e. gathering the data set. After this extracting the keywords and the side information from those data sets will be our next objective. Once the keywords and the side information are extracted we will then perform the clustering using the COATES algorithm. Then the Text classification is carried out, means classifying the clustered text for generating the optimized result according to User Behavior (localization, personalization). Then the output will be shown in the form of graph. Graphical representation will show the relevant data mined from the particular page by removing the irrelevant information. Also the analytical mined reports will be generated. These reports will depend on the previous searching method and the method used in our proposed system.

a. Gathering data set: We will gather data set. From the large amount of data we will select our data set on which we will be working.
b. Extracting keywords and side-information: From the data set we will extract the keywords and the side-information.
c. Formation of clusters: Once we will extract the keywords and the side-information we will form the clusters based on the keywords and then on the side-information.
d. Outcome: We will show the graph that will show the relevant data mined from the particular page by removing the irrelevant information.

Our contribution in this proposed work is describe as follows: We will first of all take any text file as input from the user on which the mining has to be performed. From that file we will separate the keywords and the side-information. We have maintained two dictionaries for performing the stemming and stop-words removal on the text file. With the help of these

dictionaries stemming and stop-words are removed. By doing so all the unwanted noise will get removed. After the removal of stemming and stop-words we get the relevant words in the text document. From these relevant words we count the occurrence of a particular word and then separate the keywords and the side-information.

Stop Words: Stop words are the words which should be filtered out before or after the text processing. Stop words should be removed to support the phrase searching because they can cause problems while searching for phrases that include them. Some of the list of stop words includes, "is, the, at, which, on, are, but, onto, etc". Such stop words are removed in the preprocessing step of our project.

Stemming: Stemming words are the derived words from their root words. Most of the times we come across words which have similar semantic interpretation and instead of those words their root words can be used for the information retrieval process. Stemming words includes, "Relating/Related/" can be replaced by their root word "Relate". Such stemming words are replaced by their root words in our project.

Our project consists of 3 modules. We have developed 1 module and described it shortly as above. The further development is in progress. Until now we have done the data collection process and the preprocessing on the text document which will be given as input. In future we will do mining and clustering. Clustering will be performed by applying the COATES algorithm. After the clusters are formed searching will be done based on the user behavior.

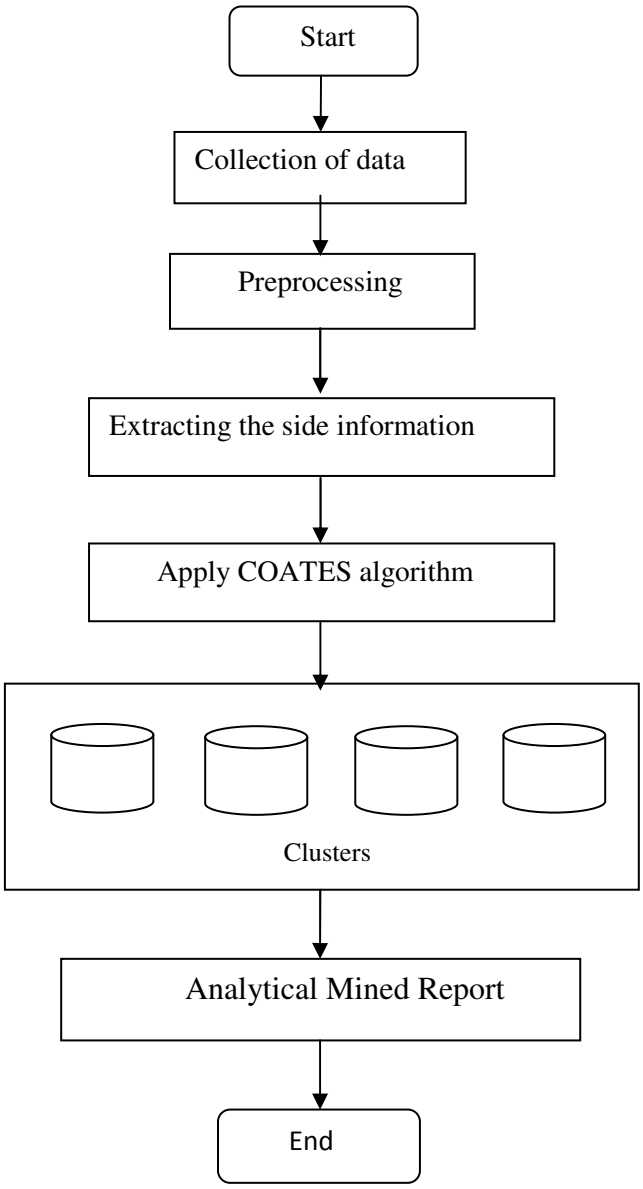

Fig.: System flow diagram

The COATES Algorithm

We will use the COATES algorithm for doing the text clustering with the help of side information. COATES is the abbreviation of COntent and Auxiliary attribute based Text clustering algorithm. The input to this algorithm will be the numbers of clusters say k. For applying the COATES algorithm it is necessary that the stop words are removed and stemming has been performed. The algorithm works in two phases:

a) Initialization: This is the first phase of the COATES algorithm. In the first phase pure text clustering is performed without using any kind of side information or the auxiliary information.

b) Main Phase: This phase is executed after the completion of the first phase. The work of the main phase is to perform the alternating iterations with the help of the text content and the auxiliary attribute information. Thus this will improve the quality of clustering.

5. Conclusion

There are many text mining applications which consist of side information along with the text documents. However, it may prove to be a difficult process for computing the importance of this side-information, especially when it is noisy. And hence the quality of mining process may decrease. Therefore, we provide a principled way for performing mining. We also focus on how to extract the side-information from the text document.

References

[1] Pooja Awandkar, Amit Pimpalkar, "Survey on User Behavioral Search using the Auxiliary Information Mining", in International Journal of Engineering And Computer Science ISSN:2319-7242 Volume 3, Issue 10 October, 2014 Page No. 9002-9006.

[2] Ting Yuan, Jian Cheng, Xi Zhang, ShuangQiu, Hanqing Lu, "Recommendation by Mining Multiple User Behaviors with Group Sparsity", in Proceedings of the Twenty-Eighth AAAI Conference on Artificial Intelligence, 2014.

[3] Amit Pimpalkar "Review of Online Product using Rule Based and Fuzzy Logic with Smiley's", International Journal of Computing and Technology, Volume 1, Issue 1, February 2014.

[4] Ryen W. White, Wei Chu, Xiaodong He, Ahmed Hassan1, Yang Song, Hongning Wang, "Enhancing Personalized Search by Mining and Modeling Task Behavior",in Proceedings of ACM International World Wide Web Conference (WWW), 2013.

[5] C. C. Aggarwal, Social Network Data Analytics. New York, USA: Springer, 2011.

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

[7] D. Blei and J.Chang, "Relational topic models for document networks", in Proceedings of AISTASIS, Clearwater, FL, USA, pp. 81–88, 2009.

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

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

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

[11] Q.Mei, D.Cai, D.Zhang, and C.X.Zhai, "Topic modeling with network regularization", in Proceedings of WWW Conference, NewYork,USA, pp. 101–110, 2008.

[12] S. Basu and Banerjee, "Topic models over text streams: A study of batch and online unsupervised learning", in Proceedings of SDM Conference, pp. 437–442, 2007.

[13] J. Zhang, Q. He, K. Chang, and E. P. Lim, "Bursty feature representation for clustering text streams", in Proceedings of SDM Conference, pp. 491–496, 2007.

[14] P. S. Yu, and C. C. Aggarwal, "A framework for clustering massive text and categorical data streams", in Proceedings of SIAM Conference Data Mining, pp. 477–481, 2006.

[15] S. Siersdorfer, and R. Angelova, "A neighborhood-based approach for clustering of linked document collections", in Proceedings of CIKM Conference, New York, USA, pp. 778–779, 2006.

[16] S. Zhong, "Efficient streaming text clustering", Neural Network., volume 18, no. 5–6, pp.790–798, 2005.

[17] S. C. Gates, P. S.Yu, and C. C. Aggarwal, "On using partial supervision for text categorization", IEEE Transaction Knowledge and Data Engineering, volume 16, no.2, pp. 245–255, February 2004.

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

[19] Y. Gong , W. Xu, and X. Liu, "Document clustering based on nonnegative matrix factorization", in Proceedings of ACM SIGIR Conference, New York, USA, pp. 267–273, 2003.

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

[21] D. Cutting, J. Pedersen, J. Tukey, and D. Karger, "Scatter/Gather: A cluster-based approach to browsing large document collections", in Proceedings of ACM SIGIR Conference, New York, USA, pp. 318–329, 1992.