www.ijecs.in
International Journal Of Engineering And Computer Science ISSN:2319-7242
Volume 3, Issue 10 October, 2014 Page No. 9002-9006

Survey on User Behavioral Search using the Auxiliary Information Mining

Pooja Awandkar, Amit Pimpalkar

Department of Computer science & Engineering,
G.H.R.A.E.T , Nagpur University, India.
puja.awandkar@gmail.com

Department of Computer science & Engineering,
G.H.R.A.E.T , Nagpur University, India.
amitpimpalkar@gmail.com

Abstract—Many text mining applications contains side-information along with the text documents. Many web documents consist of meta-data with them which correspond to 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. Such side-information may contain a huge amount of information. This huge amount of information may be used for performing clustering.

However, it may be difficult to compute the importance of this side-information, especially when some of the information from it is noisy. When the information is noisy it can be a risky approach for performing mining process along with the side information, because it can actually worsen the quality of mining process. This is why we need a principled way for performing the mining process, so that the advantages from using this side information can be maximized. We will do mining and clustering using the side information and iterative clustering and clusters will be formed. From these clusters we will search the desired keyword using user behavior, localization, personalization.

Keywords—Text mining, Side information, Mining.

Introduction:

The use of digital information is increasing day-by-day. This information in the digital world is increasing to the extent that the extraction of some relevant information from this huge amount of data is becoming quite tedious. This causes an interest in creating scalable and efficient mining algorithms. Till now the clustering of data in the pure form is done. But to handle such huge quantity of data we need to index the data according to the users need. For this we will use meta-data that is the side information that is present on most of the text documents. Many web documents consist of meta-data with them. These meta-data correspond to 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 associated with user-tags in many network and user-sharing applications, which may also be quite informative. For doing effective text mining we are making use of this side information for clustering the data. Text data mining is the process of deriving high quality information from text. 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 to 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. By using this user behavior search we will get the more relevant searching results and the desired output.

Literature review:

One of the most popular techniques for text-clustering was introduced by D. Cutting, etal [1] the scatter-gather technique, which uses a fusion of agglomerative and partitional clustering. Scatter-gather technique is particularly helpful in situations where it is difficult or undesirable to specify query formally. Y. Gong, etal [2] 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 merits of building text categorization systems were discussed by S. C. Gates, etal [3] by using supervised clustering techniques. They also discussed the new technique which helps the classifier to distinguish better among closely related clusters. The important differences between two styles of document clustering in the context of Topic Detection and Tracking were investigated by M. Franz, etal [5]. G. P. C. Fung, etal [6] 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. In this context, topic-driven clustering for text data method has been proposed by H. Wang etal [11]. Amit Pimpalkar [7] 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.

P. S. Yu, etal [8] 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. New temporal representations for text streams based on bursty features were introduced by J. Zhang, etal [9]. It was introduced for highlighting the temporally important features present in the text streams. A fast and adaptive clustering of text streams were studied by S. Zhong [10]. 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. Data centric views of online social networks are provided by C. C. Aggarwal [12].

He also focused on content-based mining issues in social networks. The issues of naturally structuring linked document collections by using clustering were addressed by R. Angelova etal [13]. They provided techniques which results in higher cluster purity and better overall accuracy.

A model of documents and the links between them, the Relational Topic Model (RTM), was proposed by J. Chang etal [14]. The problems of topic modeling with network structure (TMN) were defined by Q. Mei, etal [15]. 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. A new topic modeling structure for document networks, which examines both text and structure information for documents was proposed by Y. Sun, etal [16]. The problem of combining link and content analysis for community detection from networked data was considered by T. Yang, etal [17]. The problem with graph clustering based on structural and also attribute similarities were solved by Y. Zhou, etal [18], 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. Ting Yuan, etal [19] proposed a new recommendation model named Group-Sparse Matrix

Factorization (GSMF) that integrates various multiple types of user behaviors by performing modeling of the shared and private factors among them. W. White, etal [20] proposed methods for modeling user's on-task search behavior. These models were used to improve the personalization methods.

Problem definition:

In the existing system the method of performing clustering was very simple. There was no use of any side information for performing the clustering. The clusters were formed 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 previous methods used for clustering were designed only for the cases of pure text data. These methods do not work 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.

Proposed work:

In the proposed work our objective will be firstly to collect information i.e. retrieving the different kinds of attributes for text clustering (side information of a particular page). After this Text based clustering will be performed. In the text based clustering we will cluster all the retrieved information 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. Collection of Information
   In collection of information we will retrieve the different kinds of attributes for text clustering. These different attributes are the side information of a particular page.
b. Text based clustering
   We will cluster all the retrieved information using the COATES algorithm
c. Text classification
   Classifying the clustered text for generating the optimized result according to the user behavior (localization, personalization). User behavior can be predicted by using the cookies stored in the browser and the user's login details i.e. the location, interests, etc.
d. Outcome
   We will show the graph that will show the relevant data mined from the particular page by removing the irrelevant information.

In our proposed system first of all we will take any document as input and then we will apply Stop-word removal algorithm and Stemming algorithm on the document which was taken as input. Also we will extract the side information from the document. After this we will apply COATES algorithm on it which is iterative clustering process. After the COATES algorithm is applied we will get the output in the form of clusters. Once the clusters are formed we can do searching on the mined data with the help of user behavior (localization, personalization). By user behavior we mean that what the user wants to search exactly. User behavior can be predicted by using the cookies stored in the browser and the user's login details i.e. the location, interests, etc. Figure shows the system flow of the proposed system.

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:

- **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.
- **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.

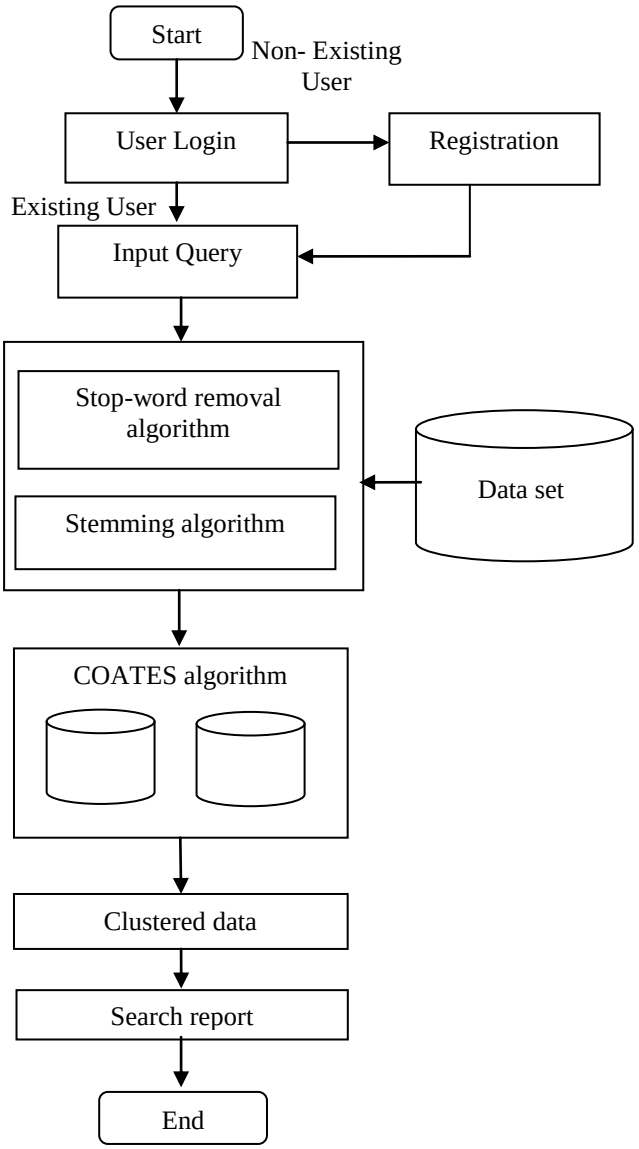

Fig (1): System flow diagram

Conclusion: In this paper mining and clustering using the side information and iterative clustering is performed and clusters will be formed. From these clusters we will search the desired keyword using user behavior (localization, personalization). By user behavioral search the user will get the desired output and the searching results will be more relevant.

References:

[1] 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.

[2] 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.

[3] 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.

[4] 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.

[5] 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.

[6] 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.

[7] 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

[8] 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.

[9] 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.

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

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

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

[13] 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.

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

[15] 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.

[16] 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.

[17] 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.

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

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

[20] 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.