Scientific Journal of Impact Factor (SJIF): 4.72

e-ISSN (O): 2348-4470
p-ISSN (P): 2348-6406

International Journal of Advance Engineering and Research Development

Volume 4, Issue 3, March -2017

ON THE USE OF SIDE INFORMATION FOR MINING TEXT DATA

Laxmi Mehetre[1], Durgesh Patil [2], Manish Pimple [3] , Akshay Satkar[4]

[1]Computer Engineering, D. Y. Patil College Of Engineering, Ambi, Pune
[2]Computer Engineering, D. Y. Patil College Of Engineering, Ambi, Pune
[3]Computer Engineering, D. Y. Patil College Of Engineering, Ambi, Pune
[4]Computer Engineering, D. Y. Patil College Of Engineering, Ambi, Pune

Abstract — *Side information is available along with text document several text mining application. This side information can be the link in the documents, web logs which contains user access behavior, provenance information, the link for ant document or any other non-textual attributes which are embedded in text document. All these attributes may contain huge amount of information for clustering purposes. Sometimes clustering more difficult when some of the information is noisy. In this matter it is inconvenient to merge side-information into the mining process because either it can upgrade the quality of the representation for mining process or can add noise in this system. Thus, there should be a right way to do this mining process so that it will make use of side information to maximize their advantage. Therefore, it suggests to design an efficient algorithm which makes combination of classical portioning algorithm with probabilistic models in order to create an effective clustering approach. Then the clustering approach will extend to classification approach for real data set which shows advantages of using such an approach.*

Keywords*- Text Mining, Side Information, COATES, Clustering, Data Mining.*

I. INTRODUCTION

Text Mining is the discovery by computer of new previously unknown information, by automatically extracting information from different written resources. A key element is the linking together of the extracted side information together to form new facts or new hypotheses to be explored further by more conventional means of experimentation. Text mining is different from what are familiar within web search. In search, the user is typically looking for something that is already known and has been written by someone else. The problem is pushing aside all the material that currently is not relevant to your needs in order to find relevant information. In text mining the goal is to discover unknown information, something that no one yet knows and so could not have yet written down.

Text mining is a variation on a field called data mining that tries to find interesting patterns from large databases. Text mining, also known as Intelligent Text Analysis, Text Data Mining or Knowledge-Discovery in Text (KDT), refers generally to the process of extracting interesting and non-trivial information and knowledge from unstructured text. Text mining is a young interdisciplinary field which draws on information retrieval, data mining, machine learning, statistics and computational linguistics. As most information (over 80%) is stored as text, text mining is believed to have a high commercial potential value. Knowledge may be discovered from many sources of information, yet, unstructured texts remain the largest readily available source of knowledge. The problem of Knowledge Discovery from Text (KDT) is to extract explicit and implicit concepts and semantic relations between concepts using Natural Language Processing (NLP) techniques. Its aim is to get insights into large quantities of text data. KDT, while deeply rooted in NLP, draws on methods from statistics, machine learning, reasoning, information extraction, knowledge management, and others for its discovery process. KDT plays an increasingly significant role in emerging applications, such as Text Understanding. Text mining is similar to data mining, except that data mining tools are designed to handle structured data from database, but text mining can work with unstructured or semi-structured data sets such as emails, full-text documents and HTML files etc. As a result, text mining is a much better solution for companies. To date, however, most research and development efforts have centered on data mining efforts using structured data. The problem introduced by text mining is obvious: natural language was developed for humans to communicate with one another and to record information, and computers are a long way from comprehending natural language.

II. SIDE INFORMATION

In many application domains, a tremendous amount of side information is also associated along with the documents. This is because text documents typically occur in the context of a variety of applications in which there may be a large amount of other kinds of database attributes or misinformation which may be useful to the clustering process. Some examples of such side-information are as follows:

- In an application in which we track user access behavior of web documents, the user-access behavior may be captured in the form of web logs. For each document, the meta- information may correspond to the browsing behavior of the different users. Such mining process in a way which is more meaningful to the user, and also application-sensitive. This is because the logs can often pick up subtle correlations in content, which cannot be picked up by the raw text alone.

- Many text documents contain links among them, which can also be treated as attributes. Such links contain a lot of useful information for mining purposes. As in the previous case, such attributes may often provide insights about the correlations among documents in a way which may not be easily accessible from raw content.
- Many web documents have meta-data associated with them which correspond to different kinds of attributes such as the provenance or other information about the origin of the document. In other cases, data such as ownership, location, or even temporal information may be informative for mining purposes. In a number of network and user-sharing applications, documents may be associated with user-tags, which may also be quite informative.

While such side-information can sometimes be useful in improving the quality of the clustering process, it can be a risky approach when the side-information is noisy. In such cases, it can actually worsen the quality of the mining process. Therefore, we will use an approach which carefully ascertains the coherence of the clustering characteristics of the side information with that of the text content. This helps in magnifying the clustering effects of both kinds of data. The core of the approach is to determine a clustering in which the text attributes and side-information provide similar hints about the nature of the underlying clusters, and at the same time ignore those aspects in which conflicting hints are provided.

III. COATES ALGORITHM

In this section, we will describe our algorithm for text clustering with side-information. We refer to this algorithm as *COATES* throughout the paper, which corresponds to the fact that it is a *COntent and Auxiliary attribute based TExt cluStering* algorithm. We assume that an input to the algorithm is the number of clusters *k*. As in the case of all text-clustering algorithms, it is assumed that stop-words have been removed, and stemming has been performed in order to improve the discriminatory power of the attributes. The algorithm requires two phases:

- **Initialization**: We use a lightweight initialization phase in which a standard text clustering approach is used without any side-information. The reason that this algorithm is used, because it is a simple algorithm which can quickly and efficiently provide a reasonable initial starting point. The centroids and the partitioning created by the clusters formed in the first phase provide an initial starting point for the second phase. We note that the first phase is based on text only, and does not use the auxiliary information.
- **Main Phase**: The main phase of the algorithm is executed after the first phase. This phase starts off with these initial groups, and iteratively reconstructs these clusters with the use of *both* the text content and the auxiliary information. This phase performs alternating iterations which use the text content and auxiliary attribute information in order to improve the quality of the clustering. We call these iterations as *content* iterations and *auxiliary iterations* respectively. The combination of the two iterations is referred to as a *major iteration*. Each major iteration thus contains *two minor iterations*, corresponding to the auxiliary and text-based methods respectively.

The focus of the first phase is simply to construct an initialization, which provides a good starting point for the clustering process based on text content. Since the key techniques for content and auxiliary information integration are in the second phase, we will focus most of our subsequent discussion on the second phase of the algorithm. The first phase is simply a direct application of the text clustering algorithm proposed. The overall approach uses alternating minor iterations of content-based and auxiliary attribute-based clustering. These phases are referred to as *content-based* and *auxiliary attributebased* iterations respectively. The algorithm maintains a set of seed centroids, which are subsequently refined in the different iterations. In each content-based phase, we assign a document to its closest seed centroid based on a text similarity function. The centroids for the *k* clusters created during this phase are denoted by $L_1 ... L_k$. Specifically, the cosine similarity function is used for assignment purposes. In each auxiliary phase, we create a probabilistic model, which relates the attribute probabilities to the cluster-membership probabilities, based on the clusters which have already been created in the most recent text-based phase. The goal of this modeling is to examine the coherence of the text clustering with the sideinformation attributes.

III. CONCLUSION

In this Project, we 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, we used an iterative partitioning technique with a probability estimation process which computes the importance of different kinds of side-information. This general approach is used in order to design both clustering and classification algorithms.

IV. REFERENCES

[1] C. C. Aggarwal and P. S. Yu, “A Framework for Clustering Massive Text and Categorical Data Streams,” in SIAM Conf. on Data
Mining ,pp. 477-481 ,2006

[2] C. C. Aggarwal and P. S. Yu, “A Framework for Clustering Massive Text and Categorical Data Streams,” in *SIAM Conf. on Data* . *Data Mining*,pp. 477–481, 2006.

[3] R. Angelova and S. Siersdorfer, “A neighborhood-based approach for clustering of linked document collections,” in *CIKM Conf.*,
pp. 778–779, 2006.

[4] H. Frigui and O. Nasraoui, “Simultaneous clustering and dynamic keyword weighting for text documents, Survey of text mining,”
Michael Berry, Ed, *Springer*, pp. 45–70, 2004.

[5] S. Guha, R. Rastogi, and K. Shim, “CURE: An Efficient Clustering Algorithm for Large Databases,” in *ACM SIGMOD Conf.*, pp.
73–84, 1998.

[6] S. Guha, R. Rastogi, and K. Shim, “ROCK: A Robust Clustering Algorithm for Categorical Attributes,” in *Inf. Syst.*, vol. 25(5), pp.
345-366,2000.