ISSN:0975-9646

A Review on Side Information Entangling For Effective Clustering Of Text Documents in Data Mining

Mr. Y.R.Gurav[1], Assoc. Prof. P.B.Kumbharkar[2]

[1,2]Compuetr Engineering,CAYMET'S
Siddhant College of Engineering,
Sudumbare, Pune, Maharashtra, India

Abstract: The study of this paper is immersed in effective clustering and mining approach with help of side information. Number of text mining applications, having side-information with them. This information may be of various forms, such as provenance information of the documents, the links in the document, web logs which contains user-access behaviour, or other text document which are embedded into the non-textual attributes. These attributes may contain a lot of information for clustering purposes. However, the concerned importance of this side-information may be hard to count, especially when some of the information is noisy. In such cases, it can be hazardous to merge side-information into the mining process, because it can either enhance the quality of the representation or can add noise in the system. Therefore, Discussion suggests way to design efficient algorithm which combines classical partitioning algorithm with probabilistic model for effective clustering approach, so as to maximize the benefits from using side information.

***Keywords*—Side information, data mining, data clustering, text classification.**

I. INTRODUCTION

The issue of text clustering uprises in the surround of many application domains such as the web, social networks, and other digital data. The rapidly increasing amounts of text data in the surround of these large online collections has led to an interest in creating scalable and effective mining algorithms. Lots of work has been done in recent years on the issue of clustering in text collections in the database and information retrieval communities. despite, this work is basically designed for issue of pure text clustering in the absence of other kinds of attributes. Some examples of such side-information are as follows

- We captured the web logs which contains Meta information related to browsing behaviour of various users, this kind of information can be used to improve the quality of the text mining. This is because the logs can often grasp sharp correlations in content, which cannot be grasped by the raw text alone.
- Various text documents having links in them, which can also be acted as attributes, such links having a lot of useful information for mining purpose. As in the former case, such attributes may often provide insights about the correlations among documents in a way which may not be easily accessible from raw content.
- Meta-data associated with many web documents correspond to different kinds of attributes such as the provenance or other information in other cases, data such as dominion, position, or even temporal Information may be informative for mining purposes. In a number of applications, documents may be associated with user-tags, which may also be truly educational.

While such side-information can sometimes be beneficial in enhancing the quality of the clustering process, it can be a dangerous when the side-information is noisy. In such cases, it can actually damage the quality of the mining process. hence, we will use an approach which carefully ascertains the coherence of the clustering characteristics of the side information with that of the text content

The main idea from the study of paper presented by Charu C. Aggarwal [8] is that, to adjudicate a clustering in which the text attributes and side-information delivers similar clue about the nature of the primal clusters, and at the same time reject those aspects in which incompatible clues are provided. To achieve this, we will merge a partitioning approach with a probabilistic evaluation method that, decides the attachment of the side-attributes in the clustering process. A probabilistic evaluation process on the side information uses the partitioning information for the purpose of evaluating the attachment of different clusters with side attributes. This helps in removing the noise in the membership behaviour of different attributes. Although our main aim in this paper is to study the clustering problem, which also be continued in principle to other data mining issues in which supplementary information is present with text.

II. LITERATURE REVIEW

Chris Clifton's [1] Top Cat (Topic Categories) is a technique for finding topics that repeat in articles in a text collection. For finding key entities in particular articles Natural language processing techniques are applied. This paper introduces a new method for finding related items based on traditional data mining approaches. From the groups of items frequent itemsets are achieved which, comes from clusters formed with a hypergraph partitioning scheme. The main issue is: what topics are repeatedly discussed in given collection of the documents? The target is to assist human understanding, so a good answer must find topics in a way that makes sense to a person. Author applies data mining technology to this issue by considering a document as a collection of entities, allowing user to trace this into a market basket problem. Author select named entities from a document by using natural language technology. Then keeping watch for frequent itemset, groups of named entities that commonly happens together. Next, clusters the groups of named entities, catching closely related entities that may not actually occur in the same document .The result are exact set of clusters. Every cluster is showing as a set of named entities and relation to a current topic in the collection. The origination of TopCat lies in how these distinct technologies are merged, plus a few specific additions such as the frequent-itemset filtering, the hypergraph-based clustering mechanism, and a generalization of the mechanism proposed. This technique also having some limitations like Performance: incrementally revising the topics without looking at all the past data, new knowledge: How to activate the user when something has changed, either new topics or new information, Other challenge for frequent itemsets, to trace when a new document results in a new group of items but, conducting this with the help of hypergraph partitioning and clustering is a difficult problem, another issue is using other extra types of information.

Paper presented by Shady Shehata [2] considered that, in text mining majority of the methods are based on the statistical study of a term or word. This statistical study gives term frequency which shows the significance of the term within a document. However, one of the terms contributes more to the meaning of its sentences than the other when two or more terms have the same frequency in their documents. Thus, the basic 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, this helps to identify of the topic of the document. A latest concept-based mining design that evaluate terms on the sentence, document, and collection stage and this mining model can distinguish between non important terms with respect to sentence semantics and terms which hold the concepts that show the sentence meaning. The proposed model can efficiently find significant matching concepts within documents, according to the semantics of their sentences. Similarity between documents is calculated by the new concept-based similarity count. This similarity count makes full use of the concept analysis measures on the sentence, document, and collection levels in calculating the similarity between documents. This work minimizes the gap between natural language processing and text mining. This concept based mining model mixture of four parts, is designed to enhance the quality of text clustering. The first part is the sentence-based concept analysis which examines the semantic structure of each sentence to obtain the sentence concepts using the proposed conceptual term frequency measure. Then, the second part, document-based concept analysis, uses concept-based term frequency to examine each concept at the document level. The third part analyzes concepts on the collection level using the document frequency global measure. The fourth part is the concept-based similarity count which allows counting the importance of each concept with respect to the semantics of the sentence, the topic of the document, and distinguishes among documents in a collection. Some limitations to these techniques are this work is not linked to web document clustering, and same model not applied to text classification. Andrew Skabar [3] studied that, fuzzy clustering algorithms allow pattern to belong to all clusters with differing degrees of membership in comparison with hard clustering methods, where a pattern corresponds to a cluster. This is valuable in sentence clustering domains since a sentence is likely to be related to more than one theme or topic present within a document. However, because most sentence similarity measures do not shows sentences in a common measurable area, normal fuzzy clustering techniques based on mixtures of Gaussians are generally not applicable to sentence clustering. Fuzzy clustering algorithm presented in this paper operates on relational input data; that is, data in the shape of a square matrix of pairwise similarities between data objects. The algorithm uses a graphical format of the data, and act in an Expectation-Maximization format. For sentence clustering task use of this algorithm shows that, the algorithm is capable of discovering overlapping clusters of semantically related sentences, and it is applicable in a variety of text mining tasks. The algorithm is able to achieve best results to benchmark Spectral Clustering and k-Medoids algorithms when externally examined in hard clustering mode on a challenging data set. The drawback of this work is that it identifies only flat clusters but the concepts present in natural language documents usually display some type of hierarchical structure.

Paper presented by Longbing Cao [4] says that, business data mining applications often involve complicated data such as multiple large different data sources, user choices, and business impact. In such conditions, a single method or one-step mining is often restricted in identifying informative knowledge. It takes too much time and space, so it is necessity to develop effective approaches for mining patterns combining necessary information from many related business fields. The recent years have seen number of efforts on mining more informative patterns, e.g., combining frequent pattern mining with categorization to achieve frequent pattern-based classifiers. Instead of presenting a particular algorithm, this paper builds on existing works and suggests combined mining as a general approach to mining for informative patterns combining components from either multiple data sets or multiple features or by multiple methods on demand. This paper has presented a comprehensive and general approach named combined mining for identifying informational knowledge in complicated data. It focuses on discussing the frameworks for handling multifeature-, multisource, and multimethod-related issues. It addressed challenging problems in combined mining and proposed effective pattern merging and interaction paradigms. This work requires further development for effective models, combined pattern types, combined mining methods, and multiple sources of data available in industrial projects such as government agencies, share market, insurance companies, and banking field.

Lei Meng [5] considered that Co-clustering is a commonly used technique for knocking the rich meta-information of multimedia web documents, including category, annotation, and explanation, for relative discovery. However, most co-clustering methods proposed for different data do not consider the representation issue of short and noisy text and their performance is bounded by the empirical weighting of the multi-modal features. In this paper, author proposed a generalized form of Heterogeneous Fusion Adaptive Resonance Theory, named GHF-ART, which perform co-clustering of huge web multimedia topics. This approach is proposed to treat multimedia data with an immediately rich level of meta-information. For treating short and noisy data, GHF-ART does not acquire information exactly from the text. Rather, it discovers key tags by learning the probabilistic distribution of tag occurrences. In this paper, author proposed a novel heterogeneous data co-clustering algorithm termed Generalized Heterogeneous Fusion ART, targeting at fast and robust clustering of web multimedia data. GHF-ART extends the Heterogeneous Fusion ART from two channels to multiple channels so that it can be applied to the clustering of more than two modalities wherein each channel may receive different types of data patterns. GHF-ART has the advantages in comparison with existing co clustering system1) Strong noise immunity 2) adaptive channel weighting method 3) Low computational complexity 4) incremental clustering manner, this work also have some limitations first tag ranking methods can be applied in the textual feature construction stage to filter noisy tags so as to further degrade the effect of noisy tags, second since the learning function for meta-information is designed to mark the probabilistic distribution of the data set in an incremental manner, there is no assurance of convergence in response to the changing data characteristics Third, as the current method for fixing vigilance parameters still cannot fully solve the problem of category proliferation.

Finding the suitable number of clusters to which documents should be separated is crucial in document clustering. In this paper, Ruizhang Huang [6] proposed a rare approach, namely DPMFP, to discover the latent cluster structure based on the DPM model without requiring the number of clusters as input. Document features are automatically separated into two groups, in particular, selective words and nonselective words, and participate differently to document clustering. A variational inference algorithm is examined to infer the document collection structure as well as the separation of document words at the same time. This approach is robust and effective for document clustering as compare to advance document clustering approaches. This approach treats document clustering and feature partition at a same time. A document clustering approach is examined based on the DPM model which collects documents into a suitable number clusters. Document words are separated according to their usefulness to select the document clusters. The selective words are used to decide the document collection structure. Nonselective words are regarded to be generated from a general background shared by whole documents. The pair of the variational inference algorithm and the blocked Gibbs Sampling method is designed to infer the cluster structure as well as the latent selective word subset. This approach shows it acquires high clustering accuracy and reasonable partition of document words, and also shows that the DPM model with automatic feature partition method could effectively discover word partitions. Challenges for this approach are adaption with the semisupervised document clustering and use of additional information to improve the performance of approach.

Research project selection is a valuable task for government and private research funding organizations. When plenty of research proposals are collected, it is common to group them according to their similarities in research field. The grouped proposals are then assigned to the appropriate experts for review. Current methods for grouping proposals are based on manual matching of similar research fields or keywords. Still, the correct research discipline field of the proposals cannot often be accurately labeled by the applicants due to their personal views and possible misinterpretations. Text-mining methods have been proposed to solve the problem by automatically classifying text documents, basically in English. Still, these approaches have drawbacks when dealing with other than English language texts, e.g., Chinese research proposals. In This paper Jian Ma [7] presented a novel ontology-based text-mining approach to cluster research proposals based on their similarities in research fields. The method is capable and competent for clustering research proposals with both English and Chinese texts. This paper has presented technique for grouping of research proposals. Research ontology is created to separate the concept terms in different fields and to form association among them. It facilitates text-mining and optimization techniques to cluster research proposals based on their similarities and then to balance them according to the applicants' characteristics. It also provides a proper procedure that enables similar proposals to be grouped together in a professional and ethical manner. The proposed method can also be used in other government research funding agencies that face information overload issues. Extra work is needed to cluster external reviewers based on their research areas and to assign grouped research proposals to reviewers systematically.

III. CONCLUSION

This paper discusses the significance of side information for effective clustering and mining. . Number of text mining applications, contains side-information with the them, this information may be of various kinds, such as provenance information of the documents, the links in the document, web logs which contains user-access behaviour. Lots of work has been done in recent years on the issue of clustering in text collections in the database and information retrieval socity. still, this work is basically designed for issue of pure text clustering in the lack of other kinds of attributes.These attributes may contain a lot of information for clustering purposes. In this paper we studied various technique, algorithm for effective text clustering and mining , after studding these techniques we comes to the conclusion that, considering side information for text data clustering and mining is excellent option because if the side information is related then it give extremely wonderful results and if the side information is noisy it can be hazardous to merge side-information into the mining process, because it can add noise to the process.so by removing this kind of noisy information we can improve the quality of clustering. Therefore, Discussion suggests way to design efficient algorithm which combines classical partitioning algorithm with probabilistic model for effective clustering approach, so as to maximize the benefits from using side information

ACKNOWLEDGE

I would like to express my sincere thanks to my guide Prof.P.B.Kumbharkar for his motivation and useful suggestions which truly helped me in improving the quality of this paper. I take this opportunity to express my thanks to my teacher, family and friends for their encouragement and support.

REFERENCES

[1] Chris Clifton, Robert Cooley"TopCat: Data Mining for Topic Identification in a Text Corpus" *IEEE Transactions on Knowledge and Data Engineering*, vol. 16, no. 8, pp.949-964, 2004.

[2] Shady Shehata, Fakhri Karray"An Efficient Concept based Mining Model for enhancing text clustering" *IEEE transaction on knowledge and data engineering*, vol.22, no.10, pp.1360-1371, 2010.

[3] Andrew Skabar, Khaled Abdalgader"Clustering Sentence-Level Text Using a Novel Fuzzy Relational Clustering Algorithm" *IEEE Transactions On Knowledge And Data Engineering*, vol. 25, no. 1, pp.62-75, 2013.

[4] Longbing Cao, Huaifeng Zhang, Dan Luo, Chengqi Zhang "Combined Mining: Discovering Informative Knowledge in Complex Data" *IEEE Transactions On Systems, Man, And Cybernetics—Part B: Cybernetics*, vol. 41, no. 3, pp.699-712, 2011

[5] Lei Meng, Ah-Hwee Tan, Dong Xu " Semi-Supervised Heterogeneous Fusion for Multimedia Data Co-Clustering" *IEEE Transactions On Knowledge And Data Engineering*, vol. 26, no. 9, pp.2293-2306, 2014

[6] Ruizhang Huang, Guan Yu, Zhaojun Wang, Jun Zhang, and Liangxing Shi " Dirichlet Process Mixture Model for Document Clustering with Feature Partition" *IEEE Transactions On Knowledge And Data Engineering*, vol. 25, no. 8, pp.1748-1759,2013

[7] Jian Ma, Wei Xu, Yong-hong Sun, Efraim Turban, Shouyang Wang, and Ou Liu " An Ontology-Based Text-Mining Method to Cluster Proposals for Research Project Selection" *IEEE Transactions On Systems, Man, And Cybernetics—Part A: Systems And Humans,* vol. 42,no. 3, pp.784-790, 2012

[8] Charu C. Aggarwal, Yuchen Zhao, and Philip S. Yu, Fello " On the Use of Side Information for Mining Text Data" *IEEE Transactions on knowledge and data engineering vol 26,no.6 pp 1415-1429,2014*