A Review on Various Document Clustering Techniques with Feature Partitioning

Nikitha Gopal
Central College of Engineering and Management
Dept. of Computer Science and Engineering
Raipur, Chhattisgarh, India
nikitagopal57@gmail.com

Vaibhav Chandrakar
Central College of Engineering and Management
Dept. of Computer Science and Engineering
Raipur, Chhattisgarh, India
vaibhavchandrakar@gmail.com

Abstract

Study of this paper is immersed with the effective clustering and mining approach with the help of information. There are wide range of text mining applications, having information with them. This information may be of various types, such as the links in the document, web logs which contains user-access behavior, provenance information of the documents or other text document which are embedded into the non-textual attributes. These attributes may include lots of information for clustering intentions. However, attributes involved with the importance of this information may be hard to count, especially when some of the information is noisy. In such cases, it can be hazardous to merge the information into the mining process, because it can either improve the quality of the representation or can add more noise in the system. Therefore, this review suggests the 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 information.

Keywords— *Information, database applications-text mining, data mining, Clustering, document clustering, text classification.*

I. INTRODUCTION

The main issue of text clustering upraises in the surround of various application areas such as the web, social networking sites, and other digital data. The fast increasing amounts of text information in the surround of these huge online collections has governed to be an interesting in making scalable and effective mining algorithms. A Lot of work has been done in present days on the issue of clustering in text collections in the database and information retrieval communities. Despite the fact that this work is basically designed for the issue of pure text clustering in the nonexistence of other varieties of attributes. Some examples of such side-information are as follows

We captured the web documents which includes Meta data information related to browsing behavior of variety of users, this kind of data can be used to get better the quality of the text mining. This is because the documents which can often hold sharp correlations in content, which cannot be hold by the raw text by itself.

Different text documents having links in them and which can also be acted as an attributes, such links having lots of worthwhile information for mining intentions. As in the previous case, this type of attributes may often generate insights about the correlations among the documents in an approach which may not be easily within reach from raw content.

Meta data information related with many web documents complement to various kinds of attributes such as the origin or other information in other cases, data such as dominion, even temporal, or position Information may be informative for mining intentions. In a various number of applications, documents may be related with user-tags, which may also be truly educational.

At the same time as such side information can every now and then be beneficial in improving the quality of the clustering process, it can be so dangerous when the side information is noisy. In that situation, in point of fact that it can damage the quality of the mining mechanism. Therefore, we will use a method which can carefully as certain the coherence of the clustering parameters of the side information with that of the text content.

A. Clustering

Clustering can be measured as a very important unsupervised learning problem. It come down with searching a structure in a collection of unlabeled data. In general, the definition of clustering could be “the process

of organizing given objects into certain number of groups whose members are similar in some way". Therefore a cluster is a collection of objects which are "similar" and are "dissimilar" between them to the objects belonging to other clusters. If the general query is given then it is extremely difficult to recognize the specific document which the user is interested in. The users are required to explore through a long list of off topics from the documents. Furthermore, internal relationships among all the documents in the search result are hardly ever presented and are left for the user.

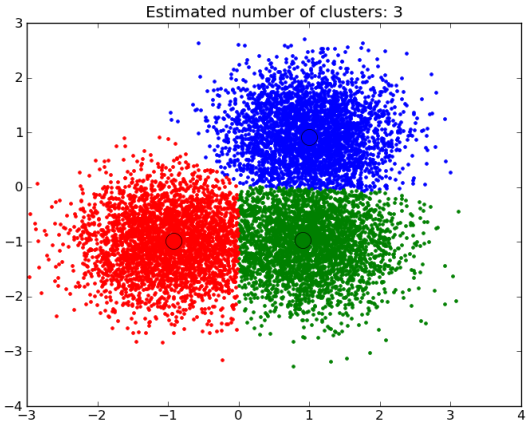

Fig.1. Clustering Overview, Cluster Size: 3

B. Document Clustering

It has been investigated that the Document clustering is use for in a number of various areas of text mining and information retrieval. At the beginning, the document clustering was used for enhancing the precision or recall in information retrieval applications as well as an efficient way of searching the nearest neighbors of a document so that system will response the maximum relevant document in return to user's query. 'The document clustering has also been used to generate the automatic hierarchical clusters of the documents. It is very closely related with data clustering. The document clustering includes the function of descriptors and descriptor extraction. The descriptors are the sets of words that explain the contents within the cluster which contains the "n" number of objects.

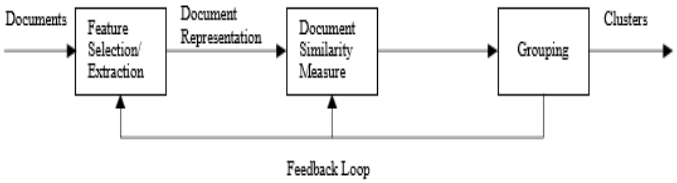

Fig. 2. Stages in Document Clustering

II. LITERATURE SURVEY

Guan Yu [1] proposed DPMFS approach handles document clustering and feature selection simultaneously. We constrain the DPM model only to define the cluster structure of the data with discriminative features which are identified by a latent binary vector. The Gibbs Sampling technique is used to infer both the cluster structure and the latent discriminative word subset. Our experiment shows that DPMFS approach group's document dataset into meaningful clusters without requiring the number of clusters known in advance.

Ruizhang Huang [2] 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.

Shady Shehata [3] 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.

In This paper Jian Ma [4] 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.

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.

Table I. shows various existing methods and its limitations

Ref. No.	Method Used	Data Source	Approach	Strength	Limitation
1	Dirichlet Process	Text Document	Proposed DPMFS approach handles document clustering and feature selection simultaneously. Author constrain the DPM model only to define the cluster structure of the data with discriminative features which are identified by a latent binary vector	Proposed approach works on both a synthetic dataset and several realistic document datasets.	Document clustering need to be more and more labeled document. .
2	Dirichlet Process	Text Document	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.	Author evaluate method that performs well on the synthetic data set as well as real data sets	Need to use of additional information to improve the performance of approach.
3	Concept-based mining model	Text Document	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	Results demonstrate the substantial enhancement of the clustering quality using the sentence-based, document-based, corpus-based, and combined approach concept analysis.	Does not did experiment with text classification.
4	Clustering analysis	Social Network Data	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.	Results can be used to improve the efficiency and effectiveness of research project selection processes in other government and private research funding agencies.	Extra work is needed to cluster external reviewers based on their research areas and to assign grouped research proposals to reviewers systematically.
5	Heterogeneous data co-clustering	Multimedia Document	Author proposed a generalized form of Heterogeneous Fusion Adaptive Resonance Theory, named GHF-ART, which perform co-clustering of huge web multimedia topics.	Shown that GHF-ART achieves significantly better clustering performance and is much faster than many existing state-of-the-art algorithms.	Need to develop the effective criteria for learning the desired vigilance parameters values.

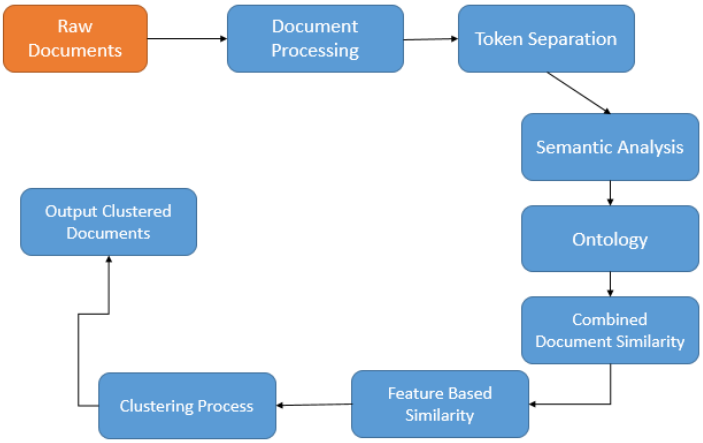

Fig. 3. Methodology for Document Clustering Method

The clustering of documents are based on weight of the node tree. The tree represent weights hence setting up weight are based on below formula. The clustering formula empirically expressed by:

$$w_{k,i} = \frac{height\_M_i - j + 1}{height\_M_i + 1}$$

where $w_{k,i}$ is the weight of an attribute k on the model i, $height\_M_i$ is height of the model i, and j is the level of attribute node k.

III. TOOLS USED

There are many tools available for processing data and extracting user information from collection of datasets. Various tools can be used are described below.

a. Text Tokenization: Used for tokenization documents.
b. Jargs: Pre-packaged and comprehensively documented suite of command line option parsers for the use of Java programmers.
c. Args4j: args4j is a small Java class library that makes it easy to parse command line options / arguments in your CUI application.

IV. CONCLUSION

This paper help to discuss the significant role of document for effective clustering and mining. There are variety of text mining applications, which contains information within them, this information may be of number of kinds, such as origin of information of the documents, web logs, the links in the documents which contains user-access behavior. A lot of work has been done in present days on the issue of clustering in text collections in the database and information retrieval. Still, this work is usually designed for issue of pure text clustering in the lack of various kind of attributes. These attributes may also having a lot of data for clustering intentions.

In this paper, we studied various different techniques, algorithm for effective text clustering and mining, after studying these techniques we came to the conclusion that, considering information for text data clustering and mining is a very excellent option because if the information is related then it provide extremely wonderful results and if the information is noisy it can be hazardous to merge 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 helps to combine the classical partitioning algorithm with DPMM model for effective clustering approach, so as to maximize the benefits from using information.

REFERENCES

[1] C. J. Van Rijsbergen, The Geometry of Information Retrieval. Cambridge, U.K.: Cambridge Univ. Press, 2004.

[2] X. Wang, H. Fang, and C. Zhai, "A study of methods for negative relevance feedback," in Proc. 31st Annu. Int. ACM SIGIR Conf. Res. Develop. Inf. Retrieval, 2008, pp. 219–226.

[3] R. W. White and R. A. Roth, Exploratory Search: Beyond the QueryResponse Paradigm. San Rafael, CA, USA: Morgan & Claypool,2009.

[4] S. Wong and V. Raghavan, "Vector space model of information retrieval: A reevaluation," in Proc.

7th Annu. Int. ACM SIGIR Conf. Res. Develop. Inf. Retrieval, 1984, pp. 167–185.

[5] Yuanhua Lv, ChengXiang Zhai, "Adaptive Relevance Feedback in Information Retrieval", ACM November 2–6, 2009.M. Young, The Technical Writer's Handbook. Mill Valley, CA: University Science, 1989.

[6] Kevyn Collins-Thompson, Sebastian de la chica and David Sontag," Personalizing Web Search Results by Reading Level", 2011 ACM.

[7] Donna Harman, "Relevance Feedback Revisited", 2010 ACM.

[8] Ingo Frommholz, Benjamin Piwowarski , Mounia Lalmas and Keith van Rijsbergen. "Processing Queries in Session in a Quantum-inspired IR Framework", 2011 ACM.

[9] Yuanhua Lv, ChengXiang Zhai, "Adaptive Relevance Feedback in Information Retrieval", ACM November 2–6, 2009.M. Young, The Technical Writer's Handbook. Mill Valley, CA: University Science, 1989.