Available Online at www.ijcsmc.com

International Journal of Computer Science and Mobile Computing

A Monthly Journal of Computer Science and Information Technology

ISSN 2320–088X

IJCSMC, Vol. 2, Issue. 10, October 2013, pg.131 – 137

RESEARCH ARTICLE

Text Document Clustering Using DPM with Concept and Feature Analysis

S Kajapriya[1], K.N Vimal Shankar[2]

PG Scholar[1], Asst. Professor[2]
Department of Computer Science & Engineering[1, 2]
V.S.B. Engineering College, Karur, India[1, 2]
Kpriya24san@gmail.com [1], yvsinformation@yahoo.in [2]

Abstract*— Clustering is one of the most important techniques in machine learning and data mining tasks. Similar documents are grouped by performing clustering techniques. Similarity measuring is used to determine transaction relationships. Hierarchical clustering model produces tree structured results. Partitioned based clustering produces the outcome in grid format. Text documents are unstructured data values with high dimensional attributes. Document clustering group ups unlabeled text documents into meaningful clusters. Traditional clustering methods require cluster count (K) for the document grouping process. Clustering accuracy degrades drastically with reference to the unsuitable cluster count. Document features are automatically partitioned into two groups' discriminative words and nondiscriminative words. Only discriminative words are useful for grouping documents. The involvement of nondiscriminative words confuses the clustering process and leads to poor clustering solution in return. A variation inference algorithm is used to infer the document collection structure and partition of document words at the same time. Dirichlet Process Mixture (DPM) model is used to partition documents. DPM clustering model uses both the data likelihood and the clustering property of the Dirichlet Process (DP). Dirichlet Process Mixture Model for Feature Partition (DPMFP) is used to discover the latent cluster structure based on the DPM model. DPMFP clustering is performed without requiring the number of clusters as input. Discriminative word identification process is improved with the labeled document analysis mechanism. Concept relationships are analyzed with Ontology support. Semantic weight model is used for the document similarity analysis. The system improves the scalability with the support of labels and concept relations for dimensionality reduction process. The system development is planned with Java language and Oracle relational database.

Keywords*— Database management; Dirichlet Process Mixture Model; Document Clustering; Feature Partition; Semi-Supervised; Text mining

I. INTRODUCTION

Clustering is especially useful for organizing documents to improve retrieval and support browsing. Document clustering aims to automatically group the related documents into clusters. Based on the various distance metrics, a number of methods have been developed to handle document clustering. With speedy growth of Internet and wide availability of news documents, one of the most useful tasks in text mining is Document clustering, which has received more and more interest in the recent years. The traditional document clustering approaches, the assumption that *K* is a pre-defined parameter determined by users and provided before the document clustering process. Moreover, an improper estimation of *K* might easily mislead the clustering process and result in bad clustering outcome. Therefore it is useful, if the document clustering approach could be designed relaxing the assumption of the pre-defined *K*. To group documents into an optimal number of clusters while the number of clusters K is discovered automatically.

II. PREVIOUS WORK

The Dirichlet Process Mixture (DPM) is to develop a model for partitioning the documents. The basic concept of DPM model is to consider both the data likelihood and the clustering property of the Dirichlet Process. The flexibility of the DPM model makes it promising for document clustering. Each document is represented by the large amount of words: the discriminative words and the non discriminative words. Here only the discriminative words are considered useful for grouping documents, because the involvement of non discriminative words confuses the clustering process and leads to poor clustering. To address this issue, a DPMFP model is introduced which extends the traditional DPM model by conducting feature partition with the unlabeled data. Each document words are partitioned into a mixture of two components, discriminative words and non discriminative words. Only discriminative words are used to imply the latent cluster structure. A Dirichlet Multinomial Allocation model [1], is used to approximate the DPMFP model to simplify the process of parameter estimation.

III. PROPOSED WORK

A proposed Approach is with the semi-supervised document clustering [2], since more and more labeled documents or documents with constraints [3], are available in real-life world situation. The additional information could improve the clustering quality .The first goal is the reasonable model parameters and initial value can be chosen from this additional information. The second goal is we can use this information guide our sampling process. In this approach, Discriminative word identification process is improved over the labeled document clustering. Here the considerations of the Concept relationships are analyzed with Ontology support. This enhanced system improves the scalability with the support of labels and constraints. Discriminative word identification process is improved with the labeled document analysis mechanism. Concept relationships are analyzed with Ontology support. Semantic weight model is used for the document similarity analysis. The system improves the scalability with the support of labels and concept relations for dimensionality reduction process. The system development is planned with Java language and Oracle relational database.

IV. SCENARIO

A. Dirichlet Process Mixture (DPM) Model

The DPM model is a flexible mixture model, were the number of mixture components grows as new data are observed. It is a kind of countable infinite mixture model [4]. We are introducing this infinite mixture model by first describing the simple finite mixture model. In the finite mixture model, each data point is drawn from one of K fixed unknown distributions. For example, the multinomial mixture model for document clustering assumes that each document x_d is drawn from one of K multinomial distributions [5]. Let η_d be the parameter of the distribution from which the document x_d is generated. Since the number of clusters is always unknown, to allow it to grow with data, we assume that the data point x_d follows a general mixture model in which η_d is generated from a distribution G.

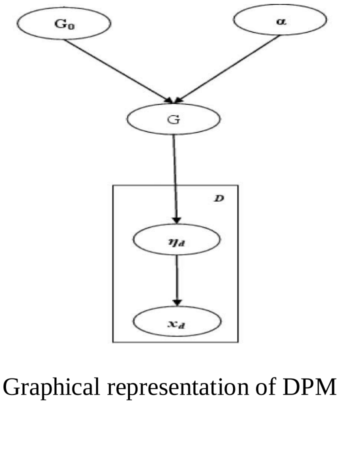


Fig.1. Graphical representation of DPM model

In the nonparametric Bayesian analysis, the Dirichlet process mixture model places a Dirichlet process prior on the unknown distribution G. In this way G can be considered as a mixture distribution with a random number of components. G is viewed as a random probability distribution in the DPM model As mentioned above the hierarchical representation of the DPM model is shown in Fig. 1 which gives better idea about this DPM Model.

B. Dirichlet Multinomial Allocation (DMA) Model

It has been shown that the DPM model can be derived as the limit of a sequence of finite mixture models when the number of mixture components is taken to infinity. One famous approximation to the DPM model is the Dirichlet Multinomial Allocation (DMA) model[6].
The generative model for the DMA model is as follows:

$P \sim \text{Dirichlet}\ (\alpha/N,\ldots,\alpha/N)$,
$\eta_i \sim G_0, i=1,2,\ldots,N$,
$Z_d \mid P \sim \text{Discrete}\ (p_1, p_2\ldots, p_N)$, $d=1,2,..,D$,
$x_d \mid z_d,\eta_1,\eta_2,\ldots,\eta_N \sim F(x_d|\eta_{zd})$, $d=1,2,\ldots,D$.

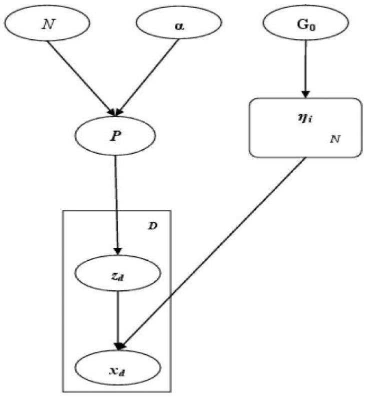


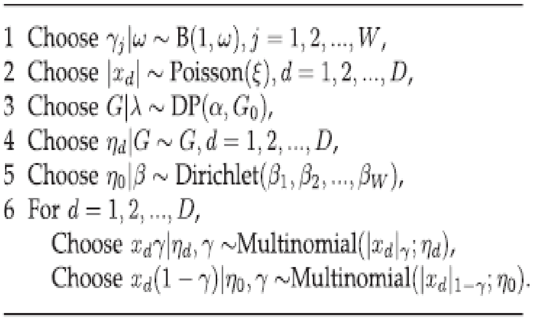
Fig.2. Graphical representation of the DMA model

Where N is the number of mixture components. P is a N dimensional vector indicating the mixing proportions for components given a Dirichlet prior with symmetric parameters α/N. z_d is an integer indicating the latent component allocation of the data point x_d. For each component, the Parameter η_i determines the distribution of data points from that component. The graphical representation of the DMA model is shown in Fig. 2.

C. Dirichlet Process Mixture with Feature Partition (DPMFP)

Formally, we define the following terms:

- A word w is an item from a vocabulary indexed by {1, 2…,W}.
- A cluster is characterized by a multinomial distribution over words. It is represented by a multinomial parameter.
- A document x is represented as a W dimensional vector
  $x_d=\{x_{d1}, x_{d2},\ldots x_{dw}\}$ where x_{dj} is the number of appearance of the word w_j of the document x_d.
- A document data set χ is a collection of D documents denoted by $\chi=\{x_1,x_2,\ldots,x_D\}$.



```
1 Choose $\gamma_j|\omega \sim \mathrm{B}(1,\omega), j = 1,2,...,W$,
2 Choose $|x_d| \sim \mathrm{Poisson}(\xi), d = 1,2,...,D$,
3 Choose $G|\lambda \sim \mathrm{DP}(\alpha, G_0)$,
4 Choose $\eta_d|G \sim G, d = 1,2,...,D$,
5 Choose $\eta_0|\beta \sim \mathrm{Dirichlet}(\beta_1,\beta_2,...,\beta_W)$,
6 For $d = 1,2,...,D$,
    Choose $x_d\gamma|\eta_d,\gamma$ ~Multinomial$(|x_d|_\gamma;\eta_d)$,
    Choose $x_d(1-\gamma)|\eta_0,\gamma$ ~Multinomial$(|x_d|_{1-\gamma};\eta_0)$.
```

Fig.3. The Generative process for the DPMFP model

Our model assumes the generative process for the document data set χ is as shown in Fig. 3. G_0 is a Dirichlet distribution with parameter $\lambda = (\lambda_1,\lambda_{2,...,}\lambda_w)$; $|x_{dj}|$ is the total appearance of the words in the document x_d; the multinomial parameter η_d represents the specific cluster to which the document x_d belongs; the multinomial parameter η_0 represents the general background sharing by all the documents in the document data set χ.

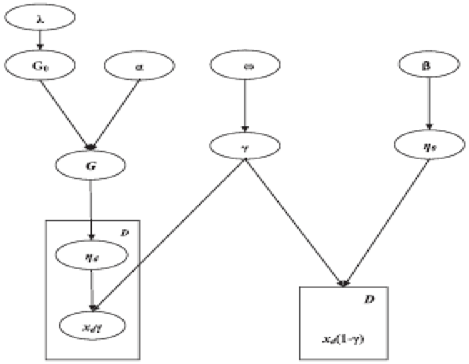


Fig.4. Graphical representation of the DPMFP model

In our model, the DP prior is only used for the specific cluster η_d. Note that $|x_{dj}|$ is an ancillary variable as it is independent of all the other data generating parameters [7]. Therefore, we ignore its randomness in the following development. The graphical representation of the DPMFP model is shown in Fig. 4.

V. IMPLEMENTATION

A. SYSTEM DESIGN

Ontology is a conceptual framework which defines entities and their hierarchical relationships. Ontology's can be used to represent documents at a semantic level. Semantic Based Analysis architecture is shown in the Fig.5.

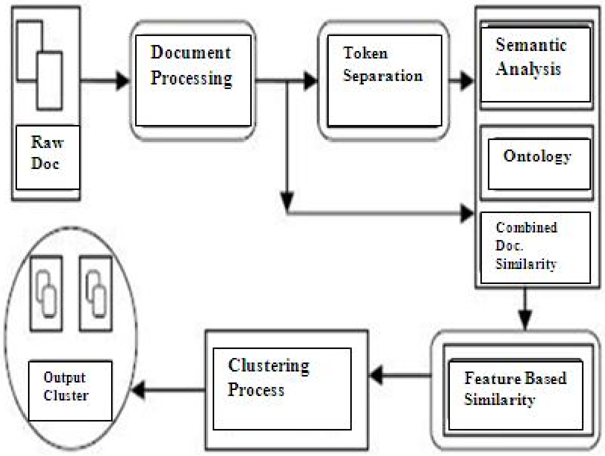


Fig.5. Architecture

In Document clustering, the traditional method extends the DPM model for the semi supervised documents. The raw documents are preprocessed and then the discrimination and nondiscrimination words are separated. With Semantic mechanism, ontology is determined with the feature partition which includes the concept and term relationships based Synonym, Meronym and Hypernym. Finally the similarity is analyzed to perform the partitioning. However, the current methods with the term weight identification do not provide an integrated solution for feature selection and clustering.

B. SYSTEM MODULES

1). Document Preprocessing: Document preprocessing describes any type of processing performed on raw data to prepare it for another processing procedure. Commonly used as a preparatory data mining practice, document preprocessing transforms the data into a format that will be more easily and effectively processed for the purpose of the user. Document pre-processing is an often neglected but it is the important step in the data mining process. The phrase "Garbage In, Garbage Out" is particularly applicable to data mining and machine learning projects. Data preparation and filtering steps can take considerable amount of processing time. Document pre-processing includes stop word elimination, stemming, etc.

2). Identifying Discrimination: This module is designed to fetch word category values based on the topics covered. Each document is represented by a large amount of words including discriminative words and non discriminative words. Only discriminative words are useful for grouping the documents. The discriminative words are the related to the concepts of the documents available in the process. Whereas the non discriminative words are the additional words which are deviating the concepts, i.e. non relevant words of the document. The discriminating and non discriminating words are determined with the help of the Variational Inference Algorithm. For the algorithm of variational inference, it could be applied to infer the document collection structure in a much quicker manner.

3). Concept Analysis: A concept-based similarity measure depends on matching concept at sentence, document, and corpus instead of individual terms. This similarity measure based on three main aspects. First is analyzed label terms that capture semantic structure of each sentence. Second is concept frequency that is used to measure participation of concept in sentence as well as document. Last is the concepts measured from number of documents. A raw document with well defined sentence boundaries is given as input to the proposed system. According to the Semantic analysis, each of the sentences in the document is labeled automatically. The sentences in the document may have one or more labeled structures.The objective behind the concept-based mechanism is to achieve an accurate analysis of concepts on thE sentence, document, and corpus levels rather than document only.

4). Feature Analysis: The Semantic Mechanism process is used to identify the document features based on the Ontology analysis which is determined based on the Dictionary View of the document words. The Semantic representation of the features are made by meaning extraction. Building Ontology's can be used to represent documents at a semantic level but this concept based model needs a well defined database or a gold standard set for mapping words to pre-defined concepts. Viewing semi-supervised learning from a clustering angle is useful in practical situations when the set of labels that are available in labeled data are not complete, i.e., untrained data contain new classes that are not present in labeled data. This proposed system analyzes several multinomial model based semi-supervised document clustering methods under a principle Dirichlet Process Mixture (DPM) model framework. Text document clustering gives an important role in providing worthy document retrieval, document browsing, and text document mining. Traditionally, clustering techniques do not consider the semantic relationships between words, such as synonymy, meronym and hypernymy. To utilize semantic relationships, ontology's have been used to improve clustering results.

5). Clustering Process: Clustering is a automatic document organization, topic extraction and fast information retrieval or filtering. It is closely related to data clustering. A web search engine often returns thousands of pages in response to a broad query, making it difficult for users to browse or to identify relevant information.

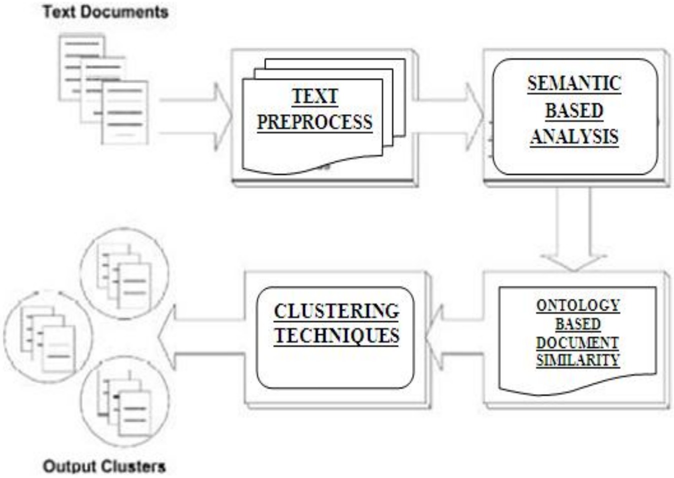


Fig.6. Semantic Based Mining Model System

Clustering methods can be used to automatically group the retrieved documents into a list of meaningful categories. Document clustering involves the use of descriptors and descriptor extraction. Descriptors are sets of words that describe the contents within the cluster. Document clustering is generally considered to be a centralized process. Examples of document clustering include web document clustering for search users. After applying the clustering techniques we get clustered document. The general format is as shown in the diagram Fig.6.That will help to find out main concepts from the several documents. The Clustering process is used to grouping the document collection for performing the word clustering.

IV. CONCLUSION

An innovative approach for document clustering, which handles both document clustering and feature partition simultaneously. A document clustering approach is investigated based on the DPM model which groups documents into an optimal number of clusters. Document words are partitioned based on their usefulness to discriminate the document clusters. The DPMFP approach in the semi-supervised document clustering is used with the Semantic Analysis, since more and more labeled documents or constraints are available in real-world. The additional information could increase the clustering quality from at least two aspects. The first one is, reasonable model parameters and initial value can be chosen from this additional information. The second one is, we can use this information for our sampling process.

REFERENCES

[1] C. Elkan, "Clustering Documents with an Exponential-Family Approximation of the Dirichlet Compound Multinomial Distribution," Proc. Int'l Conf. Machine Learning, pp. 289-296, 2006.

[2] S. Zhong, "Semi-Supervised Model-Based Document Clustering: A Comparative Study," J. Machine Learning, vol. 65, no. 1, pp. 3-29, 2006.

[3] K. Nigam, A.K. McCallum, S. Thrun, and T.M. Mitchel, “Text Classification from Labeled and Unlabeled Documents Using Em,” J. Machine Learning, vol. 39, no. 2, pp. 103-134, 2000.

[4] G. Yu, R. Huang, and Z. Wang, “Document Clustering via Dirichlet Process Mixture Model with Feature Selection,” Proc. ACM Int’l Conf. Knowledge Discovery and Data Mining, pp. 763-772, 2010.

[5] D. Blei and M. Jordan, “Variational Inference for Dirichlet Process Mixtures,” Bayesian Analysis, vol. 1, no. 1, pp. 121-144, 2006.

[6] D.M. Blei, A.Y. Ng, and M.I. Jordan, “Latent Dirichlet Allocation,” J. Machine Learning Research, vol. 3, pp. 993-1022, 2003.

[7] M.H.C. Law, M.A.T. Figueiredo, and A.K. Jain, “Simultaneous Feature Selection and Clustering Using Mixture Models,” IEEE Trans. Pattern Analysis and Machine Intelligence, Sept.2004.