Scientific Journal Impact Factor (SJIF): 1.711
e-ISSN: 2349-9745
p-ISSN: 2393-8161
International Journal of Modern Trends in Engineering and Research
www.ijmter.com

Textual Data Partitioning with Relationship and Discriminative Analysis

Ms. P. Tamilarasi[1], Mrs.T. R. Vithya[2], Mr. R. Subramanian[3]
[1]*Research Scholar*
[2]*M.Sc., M.Phil., Assistant Professor, Department of Computer Science,*
[1,2]*Selvamm Arts And Science College (Autonomous), Namakkal, Tamilnadu, India*
[3]*Genius Systems, Erode*

Abstract-Data partitioning methods are used to partition the data values with similarity. Similarity measures are used to estimate transaction relationships. Hierarchical clustering model produces tree structured results. Partitioned clustering produces results 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.

Textual data elements are divided into two types' 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.

Document labels are used to estimate the discriminative word identification process. 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.

I. INTRODUCTION

Clustering is one of the most interesting and important topics in data mining. The aim of clustering is to find intrinsic structures in data and organize them into meaningful subgroups for further study and analysis. There have been many clustering algorithms published every year. They can be proposed for very distinct research fields and developed using totally different techniques and approaches. Nevertheless, more than half a century after it was introduced, the simple algorithm k-means still remains as one of the top 10 data mining algorithms nowadays. It is the most frequently used partitional clustering algorithm in practice. Another recent scientific discussion states that k-means is the favorite algorithm that practitioners in the related fields choose to use. Needless to mention, k-means has more than a few basic drawbacks, such as sensitiveness to initialization and to cluster size and its performance can be worse than other state-of-the-art algorithms in many domains [9]. In spite of that, its simplicity, understandability and scalability are the reasons for its tremendous popularity. An algorithm with adequate performance and usability in most of application scenarios could be preferable to one with

better performance in some cases but limited usage due to high complexity. While offering reasonable results, k-means is fast and easy to combine with other methods in larger systems.

A common approach to the clustering problem is to treat it as an optimization process. An optimal partition is found by optimizing a particular function of similarity among data. Basically, there is an implicit assumption that the true intrinsic structure of data could be correctly described by the similarity formula defined and embedded in the clustering criterion function. Hence, effectiveness of clustering algorithms under this approach depends on the appropriateness of the similarity measure to the data at hand. For instance, the original k-means has sum-of-squared-error objective function that uses Euclidean distance. In a very sparse and high-dimensional domain like text documents, spherical k-means, which uses cosine similarity (CS) instead of euclidean distance as the measure, is deemed to be more suitable.

Banerjee et al. that euclidean distance was indeed one particular form of a class of distance measures called Bregman divergences. They proposed Bregman hardclustering algorithm, in which any kind of the Bregman divergences could be applied. Kullback-Leibler divergence was a special case of Bregman divergences that was said to give good clustering results on document data sets. Kullback-Leibler divergence is a good example of nonsymmetric measure. Also on the topic of capturing dissimilarity in data, Pakalska et al. found that the discriminative power of some distance measures could increase when their non-Euclidean and nonmetric attributes were increased. They concluded that noneuclidean and nonmetric measures could be informative for statistical learning of data. Pelillo even argued that the symmetry and nonnegativity assumption of similarity measures was actually a limitation of current state-of-the-art clustering approaches. Simultaneously, clustering still requires more robust dissimilarity or similarity measures; recent works illustrate this need.

II. RELATED WORK

Document clustering methods can be categorized based on whether the number of clusters is required as the input parameter. If the number of clusters is predefined, many algorithms based on the probabilistic finite mixture model have been provided in the literature. Nigam et al. proposed a multinomial mixture model. It applies the EM algorithm for document clustering assuming that document topics follow multinomial distribution. Deterministic annealing procedures are proposed to allow this algorithm to find better local optima of the likelihood function. Though multinomial distribution is often used to model text document, it fails to account for the burstiness phenomenon that if a word occurs once in a document, it is likely to occur repeatedly [10]. Madsen et al. [2] used the DCM model to capture burstiness well. Its experiments that the performance of DCM was comparable to that obtained with multiple heuristic changes to the multinomial model. DCM model lacks intuitiveness and the parameters in that model cannot be estimated quickly. Elkan [1] derived the EDCM distribution which belongs to the exponential family. It is a good approximation to the DCM distribution. The EM algorithm with the EDCM distributions is much faster than the corresponding algorithm with DCM distributions proposed in [2]. It also attains high clustering accuracy. In recent years, EM algorithm with EDCM distribution is the most competitive algorithm for document clustering if the number of clusters is predefined.

If the number of clusters K is unknown before the clustering process, one solution is to estimate K first and use this estimation as the input parameter for those document clustering algorithms requiring K predefined. Many methods have been introduced to find an estimation of K. The most straightforward method is the likelihood cross-validation technique, which trains the model with different values of K and picks the one with the highest likelihood on some held-out data. Another method is to assign a prior to K and then calculate the posterior distribution of K to determine its value. In the literature, there are also many information criteria proposed to choose K, e.g., Minimum Description Length (MDL),

Minimum Message Length (MML), Akaike Information Criterion (AIC) and Bayesian Information Criteria (BIC). The basic idea of all these criteria is to penalize complicated models in order to come up with an appropriate K to tradeoff data likelihood and model complexity.

An alternative solution is to use the DPM model which infers the number of clusters and the latent clustering structure simultaneously. The number of clusters is determined in the clustering process rather than preestimated. In our preliminary work, we proposed the DPMFS approach [3] using the DPM model to model the documents. A Gibbs Sampling algorithm was provided to infer the cluster structure. However, as the other MCMC methods, the Gibbs sampling method for the DPMFS model is slow to converge and its convergence is difficult to diagnose. Furthermore, it's difficult for us to develop effective variational inference method for the DPMFS model. Our proposed new model and the associated variational inference method in this paper solves these problems successfully.

III. PROBLEM STATEMENT

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. The following problems are identified from the existing system. Discriminative words set identification is not optimized. Labeled documents are not considered. The system supports clustering with low scalability. Concept relationships are not considered.

IV. DOCUMENT CLUSTERING WITH DIRICHLET PROCESS MIXTURE MODEL

In this paper, we attempt to group documents into an optimal number of clusters while the number of clusters K is discovered automatically. The first contribution of our approach is to develop a Dirichlet Process Mixture (DPM) model to partition documents. The DPM model has been studied in nonparametric Bayesian for a long time [5]. The basic idea of DPM model is to jointly consider both the data likelihood and the clustering property of the Dirichlet Process (DP) prior that data points are more likely to be related to popular and large clusters. When a new data point arrives, it either rises from existing cluster or starts a new cluster. This flexibility of the DPM model makes it particularly promising for document clustering. In the literature, there is little work investigating DPM model for document clustering due to the high-dimensional representation of text documents. In the problem of document clustering, each document is represented by a large amount of words including discriminative words and nondiscriminative words [8]. 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 [7]. When the number of clusters is unknown, the affect of nondiscriminative words is aggravated.

The second contribution of our approach is to address this issue and design a DPM model to tackle the problem of document clustering. A novel model, namely DPMFP, is investigated which extends the traditional DPM model by conducting feature partition. Words in documents set are partitioned into two groups, in particular, discriminative words and nondiscriminative words. Each document is regarded as a mixture of two components. The first component, discriminative words are generated from the specific cluster to which document belongs. The second component,

nondiscriminative words, are generated from a general background shared by all documents. Only discriminative words are used to infer the latent cluster structure.

IV.I. Dirichlet Process Mixture Model

The DPM model is a flexible mixture model in which the number of mixture components grows as new data are observed. It is one kind of countably infinite mixture model [4]. We introduce 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. Let η_d be the parameter of the distribution from which the document x_d is generated.

IV.II. Mean Field Variational Inference

Mean field variational inference is a particular class of variational methods [6]. Consider a model with a hyperparameter θ, latent variables $W = \{v_1, v_2, \ldots, v_S\}$ and data points $x = \{x_1, x_2, \ldots, x_D\}$. In many situations, the posterior distribution $p(W|x,\theta)$ is not available in a closed form. The mean field method approximates the posterior distribution $p(W|x,\theta)$ with a simplified distribution. In order to yield a computationally effective inference method, it's very necessary and important to choose a reasonable family of distributions Q. A common and practical method to construct such a family often breaks some of the dependencies between the latent variables. In this paper, we use the fully factorized variational distributions which break all of the dependencies between latent variables.

IV.III. Ontologies

Ontologies belong to the knowledge representation approaches that have been discussed above and they aim to provide a shared understanding of a domain both for the computers and for the humans. Thereby, ontology describes a domain of interest in such a formal way that computers can process it. The outcome is that the computer system knows about this domain. Ontology is a formal classification schema, which has a hierarchical order and which is related to some domain. An ontology comprises the logical component of a "Knowledge Base". Typically, a knowledge base consists of ontology, some data and also an inference mechanism. Ontology, comprising the logical component of the knowledge base, defines rules that formally describe how the field of interest looks like. The data can be any data related to this field of interest that is extracted from various resources such as databases, document collections, the Web etc. The inference mechanism would deploy rules in form of axioms, restrictions, logical consequences and other various methods based on the formal definition in the ontology over the actual data to produce more information out of the existing one.

V. RELATIONSHIP AND DISCRIMINATIVE ANALYSIS SCHEME FOR DOCUMENT CLUSTERING

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 [11]. The system improves the scalability with the support of labels and concept relations for dimensionality reduction process. The DPMFP model is enhanced to perform the clustering with semantic analysis mechanism. Word categorization process is improved with label values. Label and concept details are used to identify the optimal cluster count value. The system is divided into five major modules. They are document preprocess, discrimination identification, concept analysis, feature analysis and clustering process. The document preprocess module is designed to

convert the documents into tokens. Discrimination identification module is designed to fetch word category values. Concept analysis module is designed to fetch semantic relationship. Feature selection process is used to identify the document features. Clustering process is used to partition the document collection.

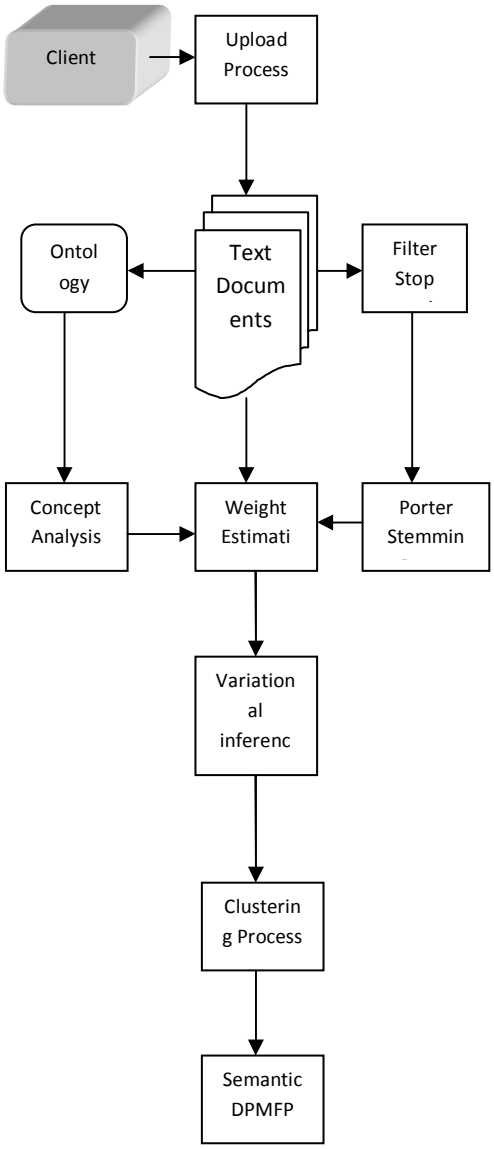

Fig. 5.1. Relationship and Discriminative Analysis Scheme for Document Clustering

V.I. Document Preprocess

The document preprocess is performed to parse the documents into tokens. Stop word elimination process is applied to remove irrelevant terms. Stemming process is applied to carry term suffix analysis. Document vector is constructed with terms and their count values.

V.II. Discrimination Identification

Term and its importance is estimated by the system. Statistical weight estimation process is applied with term and its count values. Term weight estimation is performed with Term Frequency (TF) and Inverse Document Frequency (IDF) values. Variational inference algorithm is used to perform partition of document words.

V.III. Concept Analysis

Concept analysis is performed to measure the term relationships. Ontology repository is used for the concept relationship identification. Concept weight is assigned for each document elemeVnt. Element type and frequency values are used in the concept weight estimation process.

V.IV. Feature Analysis

Feature analysis is performed to identify the feature subspace in the documents. Term features and semantic features are extracted in the feature analysis. Statistical weight values are used in the term feature extraction process. Concept weights are used in the concept features extraction process.

V.V. Clustering Process

Dirichlet Process Mixture Model for Feature Partition (DPMFP) mechanism is used to group up the documents. Cluster count estimation is carried out to find optimal partitions. Term features and concept features are used in the clustering process. Term and concept similarity analysis is used to measure the document relationships.

VI. RESULT AND DISCUSSION

The text document clustering system is developed to partition the text document collections with reference to the similarity relationship and discriminatory levels. The system uses the statistical weights and semantic weights for similarity analysis process. Discriminatory analysis is carried out using the variabtional inference algorithm. Ontology is used for the concept relationship analysis. The Dirichlet Process Mixture model with Feature Partition (DPMFP) and Dirichlet Process Mixture model with Feature Partition and Concept analysis (DPMFPC) techniques are used in the system. The system performance is measured using three different metrics. They are Fitness measure (F-Measure), Purity and Entropy measures. Cluster quality is measured and compared for both DPMFP and DPMFPC techniques. The system is tested with 1000 text documents collected from the IEEE web site. Preprocessing and stemming process are applied to extract the features. Data mining domain related Ontology is used in the system.

VI.I. Datasets

Experiments were performed on document data sets with various characteristics and sizes. Table 6.1 lists the data set properties used for evaluation. IEEE Abstracts, 20NG and RCV1 are standard text mining data sets, while IEEE Abstracts was manually collected from IEEE web site. Below is a brief description of each data set. IEEE Abstracts is a collection of 1,000 articles from IEEE Journal. It contains 16 categories, which have rather unbalanced distribution. It has been used in document clustering process.

S. No.	Property Name	Value
1	Dataset	IEEE Abstracts
2	No. of Document	1,000
3	No. of Classes	16
4	Minimum Class Size	21
5	Maximum Class Size	74

6	No. of Unique Terms	3,543
7	Average terms/Document	104

Table 6.1. Description for IEEE Abstracts Dataset

VI.II. Performance Evaluation

Cluster quality is measured with different performance metrics. The system uses the Fitness Measure (F-measure), purity and entropy to evaluate the accuracy of the clustering algorithms. Inter cluster and intra cluster accuracy levels are analyzed in the performance analysis.

VI.II.I. Fitness Measure (F-measure)

The F-measure is a harmonic combination of the precision and recall values used in information retrieval. Each cluster obtained can be considered as the result of a query, whereas each preclassified set of documents can be considered as the desired set of documents for that query. Thus the system can calculate the precision P(i, j) and recall R(I, j) of each cluster j for each class i. If n_i is the number of members of the class i, n_j is the number of members of the cluster j and n_{ij} is the number of members of the class i in the cluster j, then P(i, j) and R(i, j) can be defined as

$$P(i, j) = \frac{n_{ij}}{n_j}, \qquad (1)$$

$$R(i, j) = \frac{n_{ij}}{n_i} \qquad (2)$$

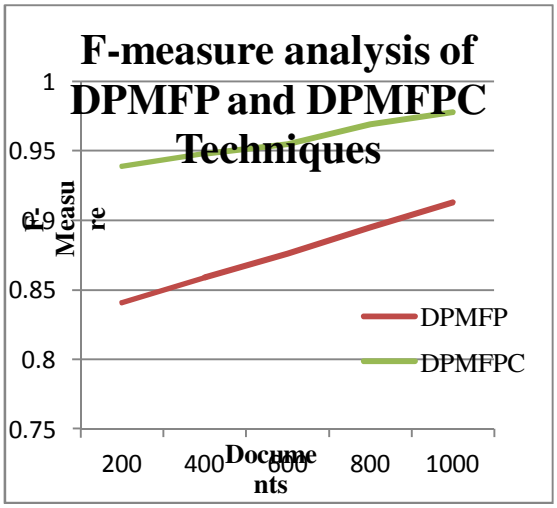

Figure 6.1: F-measure Analysis of DPMFP and DPMFPC Techniques

The corresponding F-measure F(i, j) is defined as

$$F(i, j) = \frac{2 * P(i, j) * R(i, j)}{P(i, j) + R(i, j)} \qquad (3)$$

Then, the F-measure of the whole clustering result is defined as

$$F = \sum_i \frac{n_i}{n} \max_j (F(i, j)), \qquad (4)$$

where n is the total number of documents in the data set. In general, the larger the F-measure is, the better the clustering result is (2).

The Fitness measure (F-Measure) analysis is performed on Dirichlet Process Mixture model with Feature Partition (DPMFP) technique and Dirichlet Process Mixture model with Feature Partition and Concept analysis (DPMFPC) techniques. Figure 6.1. the comparative analysis between the DPMFP and DPMFPC techniques. The F-measure analysis show that the Dirichlet Process Mixture model with Feature Partition and Concept analysis (DPMFPC) technique achieves 10% accuracy level higher than the Dirichlet Process Mixture model with Feature Partition (DPMFP) technique.

VI.II.II. Purity

The purity of a cluster represents the fraction of the cluster corresponding to the largest class of documents assigned to that cluster; thus, the purity of the cluster j is defined as

$$Purity\ (j) = \frac{1}{n_j} \max_i (n_{ij}) \qquad (5)$$

The overall purity of the clustering result is a weighted sum of the purity values of the clusters as follows:

$$Purity = \sum_j \frac{n_j}{n} Purity(j) \qquad (6)$$

In general, the larger the purity value is, the better the clustering result is (6).

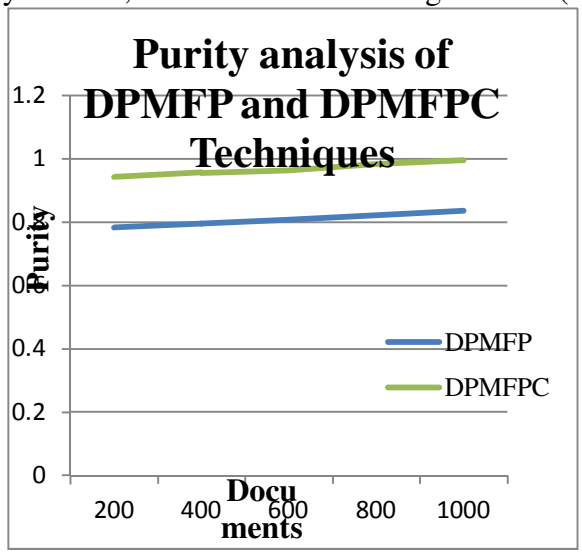

Figure 6.2: Purity Analysis of DPMFP and DPMFPC Techniques

The Purity measure analysis is performed on Dirichlet Process Mixture model with Feature Partition (DPMFP) technique and Dirichlet Process Mixture model with Feature Partition and Concept analysis (DPMFPC) techniques. Figure 6.2. shows the comparative analysis between the DPMFP and DPMFPC techniques. The purity analysis show that the Dirichlet Process Mixture model with Feature Partition and Concept analysis (DPMFPC) technique increases 15% accuracy level than the Dirichlet Process Mixture model with Feature Partition (DPMFP) technique.

VI.II.III. Entropy

Entropy reflects the homogeneity of a set of objects and thus can be used to indicate the homogeneity of a cluster. This is referred to cluster entropy. Lower cluster entropy indicates more

homogeneous clusters. On the other hand the system can also measure the entropy of a prelabeled class of objects, which indicates the homogeneity of a class with respect to the generated clusters. The less fragmented a class across clusters, the higher its entropy and vice versa. This is referred to as class entropy.

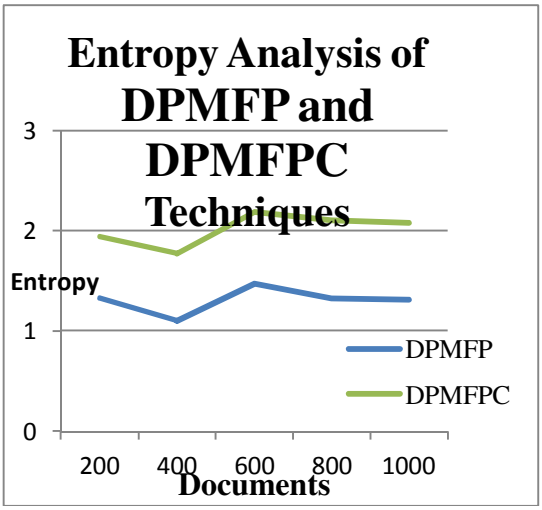

Figure 6.3: Entropy Analysis of DPMFP and DPMFPC Techniques

The Entropy measure analysis is performed on Dirichlet Process Mixture model with Feature Partition (DPMFP) technique and Dirichlet Process Mixture model with Feature Partition and Concept analysis (DPMFPC) techniques. Figure 6.3. shows the comparative analysis between the DPMFP and DPMFPC techniques. The entropy analysis show that the Dirichlet Process Mixture model with Feature Partition and Concept analysis (DPMFPC) technique increases 25% accuracy level than the Dirichlet Process Mixture model with Feature Partition (DPMFP) technique.

VII. CONCLUSION AND FUTURE WORK

The text document clustering system is developed to partition the text documents with relationship and discriminative analysis. Documents are grouped into an optimal number of clusters with automatic K estimate mechanism. Automatic cluster count estimate mechanism is used for optimal cluster count selection requirements. Statistical and semantic weight models are used in the system for the similarity analysis process. Ontology is used to performed concept analysis. Labeled documents are used for the clustering process. Clustering accuracy is improved in the system. The system reduces the process time and memory requirements for the document clustering process. The system can be enhanced with the following features. The system can be enhanced to support distributed document clustering process. The system can be adapted to support hierarchical document clustering process. The text document clustering scheme can be improved to cluster the XML documents and web documents.

REFERENCES

[1] C. Elkan, "Clustering Documents with an Exponential-Family Approximation of the Dirichlet Compound Multinomial Distribution," Proc. Int'l Conf. Machine Learning, 2006.
[2] R. Madsen, D. Kauchak and C. Elkan, "Modeling Word Burstiness Using the Dirichlet Distribution," Proc. Int'l Conf. Machine Learning, pp. 545-552, 2005.
[3] 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.
[4] Y. Teh, M. Jordan, M. Beal and D. Blei, "Hierarchical Dirichlet Processes," J. Am. Statistical Assoc., vol. 101, no. 476, pp. 1566-1581, 2007.
[5] Eduardo J. Ruiz and Vagelis Hristidis, "Facilitating Document Annotation Using Content and Querying Value", IEEE Transactions On Knowledge And Data Engineering, Vol. 26, No. 2, February 2014.

[6] D. Blei and M. Jordan, "Variational Inference for Dirichlet Process Mixtures," Bayesian Analysis, vol. 1, no. 1, pp. 121-144, 2006.
[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, vol. 26, no. 9, pp. 1154-1166, Sept. 2004.
[8] 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, August 2013.
[9] Joel Coffman and Alfred C. Weaver, "An Empirical Performance Evaluation of Relational Keyword Search Techniques", IEEE Transactions On Knowledge And Data Engineering, January 2014.
[10] Mikel Larranaga, Angel Conde and Ana Arruarte, "Automatic Generation of the Domain Module from Electronic Textbooks: Method and Validation", IEEE Transactions On Knowledge And Data Engineering, January 2014.
[11] Yung-Shen Lin, Jung-Yi Jiang and Shie-Jue Lee, "A Similarity Measure for Text Classification and Clustering", IEEE Transactions On Knowledge And Data Engineering, Vol. 26, No. 7, July 2014.

