IOSR Journal of Computer Engineering (IOSR-JCE)
e-ISSN: 2278-0661, p- ISSN: 2278-8727Volume 16, Issue 2, Ver. IX (Mar-Apr. 2014), PP 28-34
www.iosrjournals.org

A Novel Approach for Semi Supervised Document Clustering with Constraint Score based Feature Supervision

[1]S. Princiya, [2]M. Prabakaran *Assistant Professor*,
[1,2]RatnaVel Subramaniam College of Engineering and Technology,Dindigul.

Abstract: *Text document clustering provides an effective technique to manage a huge amount of retrieval outcome by grouping documents in a small number of meaningful classes. In unsupervised clustering method the unlabeled input data is used to estimate the parameter values. In a semi supervised document clustering both labeled and unlabeled input data is used for document clustering. A semi supervised clustering with feature supervision and constraint score is proposed in this paper. This proposed system which handles document clustering and feature Supervision simultaneously and this system finds the number of clusters automatically. Feature supervision uses pairwise constraints that performs supervision between the each documents. The semi-supervised constraint score that uses both pairwise constraints and the constraint score is to compute relevant features and irrelevant feature on document data set. A variational inference algorithm uses the Dirichlet Process Mixture model for the document clustering.*

Index Terms: *variational inference algorithm, semi supervised clustering, feature supervision, Feature Selection, Filter method, Constraint score, Dirichlet Polynomial Allocation (DPA) model, Dirichlet Process Mixture Model (DPM)*

I. Introduction

Document clustering is the process of collecting similar documents into meaningful clusters.Clustering analysis is traditionally considered as an unsupervised learning method. The Semi Supervisd Documents produces better performance when compare to Unsupervised documents. In a semi supervised document clustering both labeled and unlabeled input data is used for document clustering.The Feature Supervision is the concept of discovering supervised Documents.Documents within the same clusters are more similar than those in different clusters. The Must-Link and Cannot-Link constraints between instances are considered previous to performing the clustering. Different users may want to organize the document collection in their own point of view instead of a universal one. Consider a collection of news articles about books. One user may like to arrange the collection by author while another user may want to organize it by alphabet order. This is not possible in unsupervised clustering.

The Constraints score concept is combined with Feature supervision.The purpose of performing this for feature subset selection.The Constraints score is used for computing relevance between features of each documents.This concept is taken from Feature Selection method. The feature selection is very effective for reducing dimensionality, removing irrelevant and redundant features. Typically, feature selection methods can be categorized into two groups, i.e., filter methods and wrapper methods. Filter methods measuring score form a subset of feature and this method providing way for using constraint score on pairwise constraints.

Another issue in document clustering is to determine the number of document clusters K automatically. In most of the previous document clustering method, K is predefined before the clustering process. However users need to search the entire document to estimate the K. This takes much more time when the size of the document is large. Further more erroneous evaluation of K might easily betray the clustering process.

Grouping of documents into an optimal number of clusters is done by Dirichlet Process Mixture Model (DPM). It shows the auspicious results by determining the number of clusters automatically when the number of clusters is unknown. When a new data point enters, it either increases from previous clusters or initiate a new cluster. DPM model based clustering process considers both the data possibility and the clustering property of the DP(Dirichlet Process) preceding that data points are more likely to be related to popular and large clusters. This feature makes the DPM model particularly adapt for document clustering. However, there is no work inspecting the DPM model for document clustering due to the high dimensional representation of the text documents. Each document consists of huge amounts of words containing both relevant words and irrelevant words. Only relevant words are used for clustering. The involvement of irrelevant words complicates the clustering process. Therefore during document clustering, it is necessary to separate the irrelevant words specifically when the K is unknown.

Variational inference algorithm is used to deduce DPM parameters. When the document data set is large, it is harder to apply the Gibbs sampling algorithm. It also needs long time to converge. Variational inference algorithm is used to deduce the document collection structure in a quick manner. In DPMFP approach, we need to deduce both document collection structure and supervision of document words at the same time. Therefore a traditional variational inference algorithm cannot apply in DPM model.

To overcome the problem of document clustering a new approach called Dirichlet Process Mixture Model with Feature Supervision and Constraint score (DPMFS & CS) is designed in this paper. To simplify the process of parameter estimation, a Dirichlet Polynomial Allocation with Feature Supervision namely DPAFS, used to approximate the DPMFS model. The supervision algorithm performs two types of supervision. The first document supervision algorithm involves in labeling the documents. It specifies the document constraints as must-link or cannot-link. Then the feature supervision method involves in labeling the features.Then the Constraint score algorithm ranking features by computing scores using pairwise constraints. Variational inference algorithm is derived from the DPAFS model.The variational inference algorithm uses the DPM model for the clustering process. The documents which are labeled are used for clustering. It finds the number of clusters and forms the cluster automatically.

The rest of the paper is organized as follows. Section II presents a description about the previous research which is relevant to document clustering. Section III involves the detailed description about the proposed method.Section IV presents the performance analysis. This paper concludes in Section V.

II. Related Work

This section analysis the previous work performed in document clustering algorithms. Document clustering is used for retrieval of information from the text document and web documents. R. Huang, et alpresented aDirichlet process for finding the number of clusters automatically. Documents were automatically partitioned into discriminating words and nondiscriminative words.An Inference algorithm wasexploredto deduce the document collection structure and seperation of document words at the same time [1]. Mariam Kalakech et al presented feature selection scores using pairwise constraints (must-link and cannot-link) have shown better performances than the unsupervised methods and comparable to the supervised ones[2] .Nott, et al explored amodified method on the sequential updating and greedy search (SUGS) algorithm Wang and Dunson for proper Dirichlet process mixture models. The SUGS algorithm was mutated within a variational Bayes framework which was used to collect different approximations of the posterior distribution. It provided a probability distribution to allocate data to a cluster [3]. Luis Talavera et al explored optimal implementation of filters are methods that employ some criterion to score each feature and provide a ranking and the Filter methods compromise for feature selection problems [4]. Mohammed Hindawi et al presented a feature selection approach based on an efficient selection of pairwise constraints. It proposed a framework for feature selection based on constraint selection for semi-supervised dimensionality reduction. A new score function was developed to evaluate the relevance of features based on both, the locally geometrical structure of unlabeled data and the constraints preserving ability of labeled data[5].Chen, et al proposed an effective Fuzzy Frequent Itemset-based Document Clustering (F^2IDC). This approach combines the fuzzy association rule mining with the background knowledge embedded in WordNet, which improve the quality of document clustering[6]. Gil-García and Pons-Porratapresented two clustering algorithms particularly dynamic hierarchical compact and dynamic hierarchical star. The first algorithm creates disjoint hierarchies of clusters, while the second algorithm obtains overlapped hierarchies. This method offer a solution for hierarchical clustering in dynamic environment effectively and offer hierarchies easier to browse than traditional algorithms[7].Muhammad, et alcompared two approaches to document clustering based on suffix tree data model and quality of the results are analysed. First model excerpt phrases from documents and uses a similarity measure based on common suffix tree to cluster the documents.Second model extracts the similar word sequence from the document and uses the common word meaning sequence to perform the compact representation.Finally document clustering method is used to cluster the compact documents. To perform the actual clustering steps, agglomerative hierarchical document clusteringwas used [8].Jayabharathy, et alproposed a modified semantic-based model and topic detection method. Bisecting K-means algorithm was used to excerpt the related terms as concepts for concept based document cluster. Tester theory was used for determining the meaningful labels for the document clusters[9].Zhao, et alproposed aconstrained DBSCAN algorithm, which includes instant-level constraints, to obtained the better clustering performance.To accessthe user feedback, an effective learning approach is applied by selecting descriptive document pairs[10].Chen, et alformulated an efficient Fuzzy Frequent Itemset-Based Hierarchical Clustering (F^2IHC) approach. To upgrade the clustering accuracy, fuzzy association rule mining algorithm is used [11]. Mahdavi and Abolhassani presented a Harmony K-means algorithm (HKA) that deals with the work related to document clustering based on Harmony Search (HS) optimization method. It also compare the HKA with other meta-heurastic and model based document clustering approaches[12].Taiping, et al proposed a correlation preserving indexing (CPI), which is achieved in the correlation similarity measure space. In this method, correlations between the documents in the local patches are maximized and the correlations between the documents outside the patches are minimized together. The explored CPI method can efficiently determine the intrinsic structures embedded in high-dimensional document space [13].Forsati, et alproposed a novel document clustering algorithms based on the Harmony Search (HS) optimization method. HS based clustering algorithm obtain the nearby optimal clusters within a acceptable time. To achieve better clustering harmony clustering is unified with the K-means algorithm in three ways[14].Deng, et al presented a new approach called Locally Concept Factorization (LCCF) which captures the local geometry of the document submanifold. The documents associated with the same concept can be well clustered in this method[15]. Sun, et alproposed a new method for document clustering by grouping consequential topics from the document set and weighted acceptable features.It uses cluster correlationto remove the dissimilarity documents in clusters[16].

III. Dirichlet Polynomial Allocation With Feature Supervision (Dpafs)

A. Dirichlet Process Mixture Model

DPM model is a flexible mixture model which is used to count infinite mixture model. This paper proposes the infinite mixture model by first characterize the simple finite mixture model.

Each data point is derived from one of K fixed unknown distributions in the fixed mixture model. Let us take $L_k$be the parameter from which the document S_k is generated. Since the number of K is always unknown, we assume that the data point S_k follows a general mixture model in which L_k is generated from a distribution B. The comparison is as follows.

$$L_k \mid B \sim B, d = 1,2,3 \ldots\ldots D_p \qquad (1)$$
$$S_k \mid L_k \sim F(S_k \mid L_k), d = 1,2,3 \ldots\ldots D_p \qquad (2)$$

where D_p is the number of data points and $F(S_k \mid L_k)$ is the distribution of S_k given L_k.

The probability distribution B described above is always unknown. The generative mixture model is reduced to finite mixture model if B is a discrete distribution on a finite set of values. Dirichlet process mixture model places a Dirichlet process prior on the unknown distribution B in the nonparametric Bayesian analysis. Bis considered as a mixture distribution with a random number of components.

The hierarchical Bayesian requirement of the DPM modelis described below.

$B \mid \sigma, B_o \sim DP(\sigma, B_o)$ (3)

$L_k \mid B \sim B, k = 1,2,3 \ldots\ldots D_p$ (4)

$S_k \mid L_k \sim F(S_k \mid L_k), k = 1,2,3 \ldots\ldots D_p$ (5)

where$DP(\sigma, B_o)$ represents a DP with a base distribution B_o and a positive scaling parameter σ. Possibly, B_o is the mean of the DP and σ is the inverse variance.

The pictorial representation of the DPM model is shown in Fig.1.

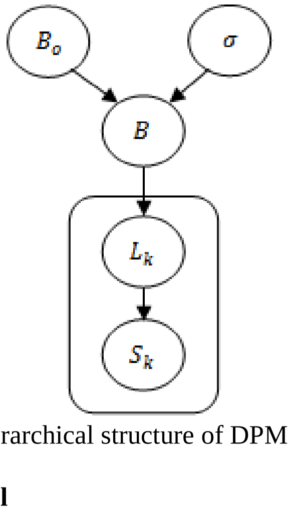

Fig.1. Hierarchical structure of DPM model

B. Dirichlet Polynomial Allocation Model

When the number of mixture components is taken as infinity, the DPM model can be derived as the limit of a sequence of finite mixture models.Dirichlet Polynomial Allocation (DPA) model is one of the eminent approximations of the DPM model. The common model for the DPA model is as follows:

$$R \sim \text{Dirichlet}\ (\sigma|M, \ldots\ldots\ldots\ldots. \sigma|M)$$

$$L_i \sim B_o, i = 1,2,3 \ldots\ldots M$$

$$X_k|R \sim \text{Discrete}(r_1, r_2, r_3 \ldots\ldots r_M), k = 1,2,3 \ldots\ldots D_p$$

$$S_k|X_k , L_1, L_2 \ldots\ldots L_M \sim F(S_k|L_{X_k}), k = 1,2,3 \ldots\ldots D_p$$

where M is the number of mixture components. R is an M-dimensional vector indicating the mixing proportions for components given a Dirichlet prior with symmetric parameter $\sigma|M$. X_k is an integer illustrate the latent component allocation of the data point S_k . For every component, the parameter L_i indicates the distribution of data points from that component.

C. Feature Supervision and Constraint Score

A document k can be measured as a list of words in the document.words take place in the document, i.e., <w1;w2; : : : ;wk >, where k is the length of the document. To label a document, the user need to understand the document.While understanding a document, the user is assumed to be able to label words he encounters. The labeled words are included in the labeled feature set W. The user labels a feature if it is a good description of the topic of a cluster and discriminates among the clusters.

M= {(wi,wj), such as wi and wj must be connected}
C = {(wi,wj), such as wi and wj cannot be connected}

Constraint score algorithm evaluates features on pairwise constraint subsets M and C.It discovers relevant features from supervision applied documents.After this the Constraint score algorithm ranking the features.

Let 'tri' indicate the r-th feature of the i-th sample , i = 1,…,m; r = 1,…,n. To evaluate the score of the r-th feature using the pairwise constraints in C and M, and describe two score functions S1 and S2.

The Constraint score algorithm in the form of Constraint score function described below.

$$S1=\frac{\sum_{xi,xj\in M}(tri-trj)^2}{\sum_{xi,xj\in C}(tri-trj)^2} \qquad (6)$$

$$S2 \; = \sum_{xi,xj\in M}(tri - trj)^2 - \mu \sum_{xi,xj\in C}(tri - trj)^2 \qquad (7)$$

So name the feature selection algorithms based on the score functions in Eqs (6) and (7) as **Constraint Score**.Features on documents sampled to compute the constraint scores.if there is a must-link constraint between two data samples, a 'good' feature occurred, where two samples are close to each other; on the other hand, if there is a cannot-link constraint between two data samples, a 'bad' feature occurred, where two samples are far away from each other.

D. Algorithm Description

```
Algorithm:    Constraint Score

Input: Data set X , pairwise constraints set M and C ,λ (for store the Constraint Score only)
Output: The ranked feature list
Step 1: For each of the n features, compute its constraint score (comparing sampled each two features)
Step 2: Store constraint score for each features
Step 3: Rank the features according to their constraint scores in ascending order.
```

The procedure of semi supervised constraint score shown in algorithm 1

```
Algorithm: Semi-supervised Clustering with Feature Supervision and Constraint Score

Input: Set of data points X
Output: K clusters $\{X_l\}_l^k = 1$
Method:
Step 1: Perform dual supervision, i.e., document supervision andfeature supervision
Step 2: Obtain the labeled feature set WL and the documentseed set S or must-link set M and cannot-link set C
Step 3: Compute constraint scores on Feature subsets M and C
Step 4: Ranking Feature based on Constraint Scores
Step 5: if Combine connected Documents
Step 6: Perform basic Clustering using the learnedweights
Step 7: else
Step 8: Perform feature reweighting based on labeled featureset WL.
Step 9: Cluster the documents using semi-supervised clustering algorithm.
Step 10: end if
```

The procedure of semi supervised clustering with feature supervision and Constraint Scores is shown in algorithm 2.

E. Semi supervised clustering with Feature Supervision

Traditionalsemi-supervised method employs user supervision in the form of pairwiseconstraints. Adding feature supervision to semi supervised clustering provides dual supervision for clustering. Both supervision takes place together before the clustering algorithm begins.The Supervision improved by constraint score algorithm. The variational inference algorithm is derived from the DPAFS model to infer the document collection structure and the separation of document words at the same time. To find the cluster structure variational inference algorithm is used. The probability distributions are calculated by using the term frequencies. The conditional probability distribution for each document is calculated by using the variational inference algorithm. After finding the approximation for all the documents the document is decided in which cluster it belongs. In the document where it achieved highest approximation value the document belongs to that cluster. So the user can give the input as the words to get the relevant clusters. The architecture of the system is shown in the Fig.2.

In data mining data preprocessing is the first step for document clustering. Here the 20 newsgroups dataset considered for the clustering. For the labeling of documents semi supervised method with feature supervision algorithm is used. It differentiates the document constraints as must-link and cannot-link. constraint score algorithm computes relevant features on feature subsets M and C. Ranking relevant features based on constraint score.documents are filtered by the constraint score then combine the connected documents.Data preprocessing is used to remove unwanted data and noise in data sets. Two steps mainly stemming and removal of stopword. The stemming process removes the suffixes appended with each word.Stopword process removes the article and prepositions.To find the term frequencies we should handle the words which are repeated.The clusters form based upon the frequencies of the terms. To find the term frequency first of all the words in the documents is collected. The document can contain the repeated words throughout the document. So to handle the repeated words we should find the distinct words in the documents. The distinct words are the words which are not repeated comes out in the documents. So the term frequency is constructed from these distinct words. Finally Variational inference algorithm is used for cluster identification.

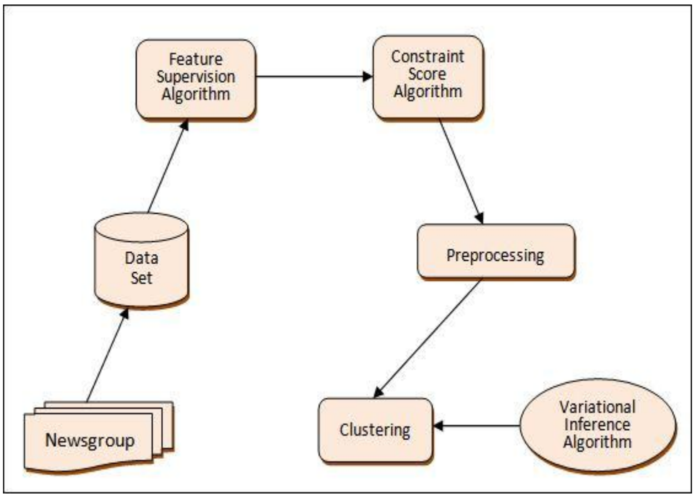
Feature Supervision Algorithm
Constraint Score Algorithm
Data Set
Preprocessing
Newsgroup
Clustering
Variational Inference Algorithm

Fig.2. System Architecture

IV. Performance Analysis

This section presents the performance evaluation of the proposed DPAFS method. The performance is evaluated based on the following measures:

F. Time Consumption

Fig.3. Illustrate the time taken for cluster formation in semi supervised method and unsupervised method. USDM refers to an Unsupervised document method. SSDM refers to a Semi supervised document

method. From the figure it is observed that the forming of clusters in the unsupervised method takes more time when compared to the semi supervised method.

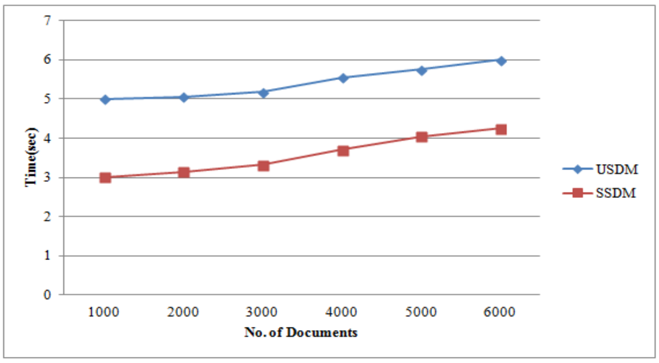

Fig.3. Time taken for cluster formation

G. Number of clusters

In the proposed method the number of clusters should be calculated previously. If the number of clusters is at a high level, then the document cluster provides efficient result. Fig.4. Illustrates the number of cluster formation in unsupervised document clustering and semi supervised document clustering. From the figure it is observed that the number of clustersis high in semi supervised document method.

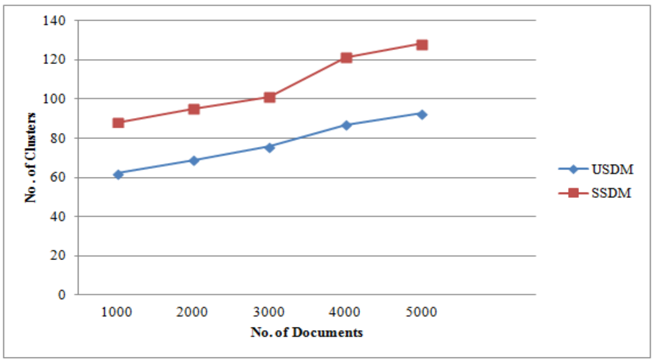

Fig.4. Number of Clusters in USDM and SSDM

H. Must-Link Feature

Fig.5 illustrates the comparison of number of must-link features in semi supervised document method and unsupervised document clustering.

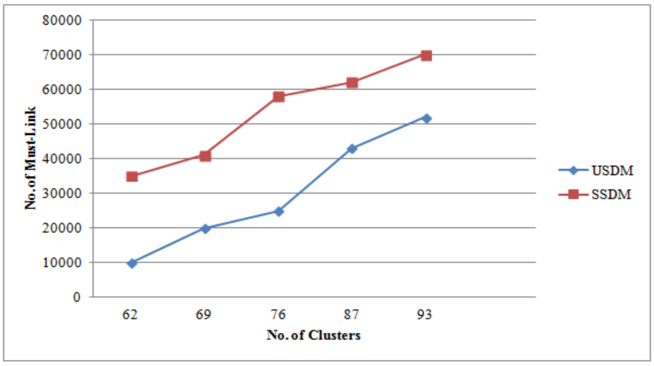

Fig.5. Number of must-link words in USDM and SSDM

I. Must-Link Feature with Constraint Score

Fig.6 illustrates the comparison of number of must-link features with Constraint Score in semi supervised document method and unsupervised document clustering.

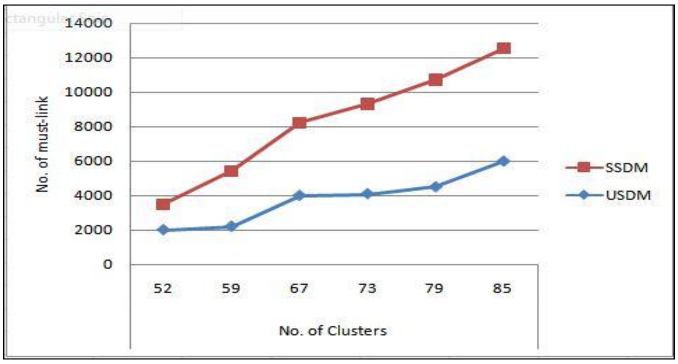

Fig.6. Number of must-link words in USDM and SSDM

V. Conclusion

In this paper, an approach is proposed that handles both document cluster and feature supervision with constraint score at the same time.Semi supervised document clustering with feature supervision, which asks the user to label the features by indicating whether they discriminate among clusters.The constraint score is computed to find the relevance between feature of each documents. A document cluster approach is investigated supported the DPM model that teams documents into capricious range clusters. Our experiment shows that our approach acquires high cluster accuracy and affordable supervision of document words. The comparison between our approach and progressive approaches indicates that our approach is strong and effective for document cluster. Our analysis of the experiment result also shows that the DPM model with automatic feature supervision with constraint score methodology may effectively discover word supervision and improve the document cluster quality.

References

[1]. R. Huang, et al., "Dirichlet Process Mixture Model for Document Clustering with Feature Partition,"2012.

[2]. M.Kalakech et al., Philippe Biela et al.,"Constraint scores for semi-supervised feature selection:A comparative study,"2011.

[3]. D. Nott, et al., "A sequential algorithm for fast fitting of Dirichlet process mixture models," arXiv preprint arXiv:1301.2897, 2013.

[4]. L. Talavera et al.,"An evaluation of filter and wrapper methods for feature selection in categorical clustering".

[5]. M.Hindawi et al., Kaïs Allab et al ., " Constraint Selection based Semi-supervised Feature Selection",2011

[6]. R. Gil-García and A. Pons-Porrata, "Dynamic hierarchical algorithms for document clustering," Pattern Recognition Letters, vol. 31, pp. 469-477, 2010.

[7]. C.-L. Chen, et al., "An integration of fuzzy association rules and WordNet for document clustering," Knowledge and information systems, vol. 28, pp. 687-708, 2011.

[8]. R. Muhammad, et al., "A comparison of two suffix tree-based document clustering algorithms," in Information and Emerging Technologies (ICIET), 2010 International Conference on, 2010, pp. 1-5.

[9]. J. Jayabharathy, et al., "Document clustering and topic discovery based on semantic similarity in scientific literature," in Communication Software and Networks (ICCSN), 2011 IEEE 3rd International Conference on, 2011, pp. 425-429.

[10]. W. Zhao, et al., "Effective semi-supervised document clustering via active learning with instance-level constraints," Knowledge and information systems, vol. 30, pp. 569-587, 2012.

[11]. C.-L. Chen, et al., "Mining fuzzy frequent itemsets for hierarchical document clustering," Information Processing & Management, vol. 46, pp. 193-211, 2010.

[12]. M. Mahdavi and H. Abolhassani, "Harmony K-means algorithm for document clustering," Data Mining and Knowledge Discovery, vol. 18, pp. 370-391, 2009/06/01 2009.

[13]. Z. Taiping, et al., "Document Clustering in Correlation Similarity Measure Space," Knowledge and Data Engineering, IEEE Transactions on, vol. 24, pp. 1002-1013, 2012.

[14]. R. Forsati, et al., "Efficient stochastic algorithms for document clustering," Information Sciences, vol. 220, pp. 269-291, 2013.

[15]. C. Deng, et al., "Locally Consistent Concept Factorization for Document Clustering," Knowledge and Data Engineering, IEEE Transactions on, vol. 23, pp. 902-913, 2011.

[16]. P. Sun, et al., "Document Clustering Method Using Weighted Semantic Features and Cluster Similarity," in Digital Game and Intelligent Toy Enhanced Learning (DIGITEL), 2010 Third IEEE International Conference on, 2010, pp. 185-187.