An Empirical Selection Method for Document Clustering

P.Perumal
Sri Ramakrishna Engg College
Department of CSE
Coimbatore

R. Nedunchezhian
Sri Ramakrishna Engg. College
Department of IT
Coimbatore

D.Brindha
Sri Ramakrishna Engg. College
Department of CSE
Coimbatore

ABSTRACT
Model Selection is a task selecting set of potential models. This method is capable of establishing hidden semantic relations among the observed features, using a number of latent variables. In this paper, the selection of the correct number of latent variables is critical. In the most of the previous researches, the number of latent topics was selected based on the number of invoked classes. This paper presents a method, based on backward elimination approach, which is capable of unsupervised order selection in PLSA. During the elimination process, proper selection of some latent variables which must be deleted is the most essential problem, and its relation to the final performance of the pruned model is straightforward. To treat this problem, we introduce a new combined pruning method which selects the best options for removal, has been used. The obtained results show that this algorithm leads to an optimized number of latent variables. In this paper, we propose a novel approach, namely DPMFS, to address this issue.

Keywords:
Document clustering, Model selection, EM algorithm, Dirichlet Process Mixture Model, Feature Selection.

1. INTRODUCTION
Document clustering is a key issue in information retrieval, which groups documents in an unsupervised manner. Information retrieval (IR) is the science of searching for documents, for information within documents, and for metadata Document clustering is an automatic grouping of text documents into clusters so that documents within a cluster have high similarity in comparison to one another, but are dissimilar to documents in other clusters. Document clustering also referred to as Text clustering is closely related to the concept of data clustering. Document clustering is the task of automatically organizing text document into meaning full cluster or group, such that the document in the same cluster are similar, and are dissimilar from the one in other clusters. It is one of the most important tasks in text mining. There are several number of technique launched for clustering documents since there is rapid growth in the field of Internet and computational technologies, the field of text mining have an abrupt growth, so that simple document clustering to more demanding task such as production of granular taxonomies, sentiment analysis, and document summarization for the scope of developing higher quality information from text. Document clustering algorithms mainly uses features like words, phrases, and sequences from the documents to perform clustering. Document clustering has been studied intensively because of its wide applicability in areas such as web mining and information retrieval. Document clustering has long been an important problem in text processing systems. The goal in most of document clustering systems is to automatically discover, in the absence of metadata or a pre-existing categorization, sensible topical organizations of the document.

Document clustering is a more specific technique for unsupervised document organization, automatic topic extraction and fast information retrieval or filtering. Document clustering techniques can be divided basically into two main groups: Similarity-based and generative-based approaches. Our approach for realizing the model selection capability is based on the hypothesis that, if we search for solutions in an incorrect solution space. The Problem is to estimate the model order in the application of PLSA. First issue is the probable stopping in local optima during the learning process. The second is choosing proper criteria to evaluate obtained models and to present some solutions for these problems.

The remainder of this paper is organized as follows. Section II discusses some of the related work on document clustering. Section III provides an earlier research work. Section IV focuses on proposed work. Section V focuses on experimental results. Section VI concludes the paper and future enhancement.

2. RELATED WORK
2.1 Model Selection Method
In basic forms, model selection is one of the fundamental tasks of scientific inquiry. Determining the principle that explains a series of observations is often linked directly to a mathematical model predicting those observations. Model selection techniques can be considered as estimators of some physical quantity, such as the probability of the model producing the given data.The problem of picking among different mathematical models which all purport to describe the same data set. The goal of model selection is estimation when bs bm (Dn) (Dn) is used for estimating and the goal is to minimize its loss.

2.2 Model Order Reduction
Model Order Reduction (MOR) is a branch of systems and control theory, which studies properties of dynamical systems in application for reducing their complexity, while preserving (to the possible extent) their input-output behavior. It is a high-dimensional state vector is actually belongs to a low-dimensional subspace.

Goals

- Automatic

- Good approximation
- Parameterized reduced models.

2.3 EM Algorithm

An expectation-maximization (EM) algorithm is a method for finding maximum likelihood or maximum a posteriori (MAP) estimates of parameters in statistical models, where the model depends on unobserved latent variables. EM is an iterative method which alternates between performing an expectation (E) step, which computes the expectation of the log-likelihood, evaluated using the current estimate for the latent variables, and maximization (M) step, which computes parameters maximizing the expected log-likelihood found on the *E* step.

2.4 Backward elimination

Backward elimination is one of several computer-based iterative variable-selection procedures. It begins with a model containing all the independent variables of interest. Then, at each step the variable with smallest F-statistic is deleted. The method based on backward elimination approach which is capable of components more than he needed value and then prunes the mixtures to reach their optimum size during the elimination process, proper selection of some latent variables which must be deleted.

1. This procedure begins with a model that includes all the independent variables.
2. It then attempts to delete one variable at a time by determining whether the least significant variable currently in the model can be removed.
3. Once a variable has been removed from the model it cannot reenter at a subsequent step.

3. PREVIOUS WORK

The key idea of LSA [7] is to map documents and by symmetry terms to a vector space of reduced dimensionality, the latent semantic space, which in a typical application in document indexing is chosen to have of the order dimensions (Deerwester et al., 1990; Dumais, 1995). The mapping of the given document term vectors to its latent space representatives is restricted to be linear and is based on a decomposition of the co-occurrence matrix by SVD. Compared to standard Latent Semantic Analysis which stems from linear algebra and performs a Singular Value Decomposition of co-occurrence tables, the proposed method is based on a mixture decomposition derived from a latent class model [4].

Tahereh Emami Azadi, Farshad Almasganj [1] proposed that.Model selection algorithm begins by taking a larger latent dimension than needed, and then continues by pruning the unvaluable latent variables to finally arrive at a model which has an optimum latent dimension. Model selection in clustering requires (i) to specify a clustering principle and (ii) to decide an appropriate number of clusters depending on the noise level in the data.The optimized case must indeed perform a high quality clustering process. The steps for suggested approach are:

- Model initialization and learning.
- Model order reduction.
- Denoting the ‘‘validated” model

Model selection is the problem of picking among different mathematical models which all purport to describe

- Efficient.

the same data set. A good model selection technique will balance goodness of fit with simplicity.

Algorithm1: Model Selection (n (wi, dj))

Input: n (wj, dj) (occurrences of word wi and dj for all i and j);
Output: K* (Optimum number of latent dimension) and the ‘‘validated” model parameters;
(1) K Kmax;
(2) Randomly initialize p(wijzk) and p(zkjdj) for all i, j and k;
(3) Model learning by EM algorithm (terminate when stop condition is met);
(4) Repeat
(5) Delete a latent variable by calling model order reduction algorithm;
(6) K K _ 1;
(7) Until K > 1
(8) compute BICK using (7);
(9) K K; BICK_ BICK ;
(10) Repeat
(11) Compute BICK;
(12) If BICK > BICK_
(13) K* K; BICK_ BICK ;
(14) End if
(15) Until K < Kmax
(16) Return K* and pK_ ðwijzkÞ and pK_ ðzkjdjÞ for all i, j and k.

The proposed approach for realizing the model selection capability is based on the hypothesis that, if we search for solutions (i.e. correct document clusters) in an incorrect solution space (i.e. Using an incorrect number of clusters), result obtained from each run of the document clustering will be quite randomized because the solution does not exist. Otherwise, results obtained from multiple runs must be very similar assuming that there is only one genuine solution in the solution space. Translating this into the model selection problem, it can be said that, if our guess on the number of clusters is correct, each run of the document clustering will produce similar sets of document clusters; otherwise, clustering result obtained from each run must be unstable, showing a large disparity. Model Selection is a difficult and pervasive local optima problem and its quite computational cost. The problem of model selection complexity control arises when a set of possible models consists of parametric models of varying complexity.

Tahereh Emami Azadi, Farshad Almasganj [1] proposed that Model reduction or model order reduction is a mathematical theory to find a low-dimensional approximation for a system of ordinary differential equations (ODEs). The most important part of this approach is the ‘‘order reduction” stage. It is specialized for selection and removal of non-efficient latent variables.

The model with chosen to overestimate the true number of clusters. Then minimize the BIC cost for this component model. Next, we delete the component whose removal is estimated to give the greatest decrease least

increase in BIC. There are various methods for selecting this component. In [2], the authors (one-by-one) *trial* deletes each component and then rerun the learning to retune the remaining components. Thus they measure the (immediate) effect of removing each component and permanently delete the component whose removal yields the lowest cost. Since models may have the components, however, trial-deletions can be quite computationally expensive. This procedure, with component pruning followed by parameter retuning, is repeated, reducing the model down to a single component. The final selected model (with associated order) is the one with least BIC cost. This procedure is obviously heuristic. Its effectiveness can be partially understood if we view component pruning as a way of removing poorly initialized components [3].

3.1 Model order reduction by trial-deletion procedure

Trial-deletion (one-by-one removing), can be also applied to reduce the latent variables of a PLSA model. The deleted variable is the one whose removal (followed by rerunning the learning process to fine-tune the model) yields the smallest decrease or greatest increase in the performance function, here BIC.Trial-deletion can be applied in our task, it has two main disadvantages. First, selection of the best option for removal, and quite time consuming. Second defect is related to the performance function. For model order reduction, we can follow a similar approach applied to the conventional simple mixture models. The conventional mixture model based on a trial and error method, have deleted the components one by one and then rerun the learning step to retune the remaining components.

Algorithm2 :(Model order reduction pk (wijzk), pk (zkjdj), s).

Input: pk (wijzk) and pk (zkjdj) for all i, j and k (the parameters of the model with K number of latent variables) and s (threshold value)
Output: pk_1 (wijzk) and pk_1 (zkjdj) and for all i, j and k (the parameters of the reduced order model)
(1) For k = 1… K do
(2) compute WSk using (8–10);
(3) End for
(4) z0 i ranked latent variables by WS;
(5) For i = 1,. . . ,K do
(6) Remove the ith top ranked latent variable ðz0 iÞ;
(7) Learn the reduced order model by running EM algorithm;
(8) compute BICi k_1;
(9) If BICi k_1 P s then
(10) go to line 14;
(11) End if
(12) Restore the model to the initial value (The parameters of the model with K number of latent variables)
(13) End for
(14) Remove the latent variable which has the maximum value of BICi k_1 for all i;
(15) Return pk_1 (wijzk) and pk_1 (zkjdj).

3.2 Model order reduction by combined method

Combined method works better than the trial-deletion in selecting the irrelevant latent topics to prune. It is used as choosing the best option for deletion. It is obvious that using the combined method instead of the one-by-one approach is computationally less costly. The latent variables superiorities for deletion are determined with their weighted similarities. The removal procedure and checking the changes in the objective function for new choices must be done in iterative manner. Initially select and remove the latent topic with the highest priority. It reduces their complexity. It is specialized for selection and removal of non-efficient latent variables.

4. PROPOSED WORK

In the proposed system, the method based on backward elimination approach which is capable of components more than the needed value and then prunes the mixtures to reach their optimum size during the elimination process, proper selection of some latent variables which must be deleted. To treat this problem introduces a new combined pruning method selects the best options for removal, while keeping a low computational cost, at all. The proposed novel approach enhance, namely DPMFS to address this issue. The proposed approach is designed 1) to group documents into a set of clusters while the number of document clusters is determined by the Dirichlet process mixture model automatically 2) to identify the discriminative words and separate them from irrelevant noise words via stochastic search variable selection technique. The DPM model is a mixture model with an infinite number of mixture components [12]. The infinite mixture model firstly describing the simple finite mixture model. In the finite mixture model, each data point is drawn from one of *K* fixed unknown distributions. The multinomial mixture model for document clustering assumes that each document *xn* is drawn from one of *K* multinomial distributions parameterized by *K* different multinomial parameters, $\theta 1,\ldots,\theta K$. The data point *xn* follows a general mixture model in which the parameter θ is generated from a distribution *G*. The process based on the DPM model considers both the data likelihood and the property of the DP prior that data points are more likely to be related to popular and large clusters [12, 13]. This flexibility of the DPM model makes it particularly useful for document clustering. One reason is that the high-dimensional representation of text documents is composed of all distinct words including discriminative words and a large number of irrelevant noise words. The proposed approach is Robust and effective for document clustering. This stochastic search variable selection technique has been used successfully in various applications to identify informative variables [9, 10]. As [10], proposed system combines this technique with DPM model.

DPM Model with Feature Selection

The following generative process for the *D* documents in a dataset:
1. Choose $\gamma \mid \omega \sim p(\gamma)$.
2. Choose $Nij \sim Poisson(\xi j)$, $i = 1, 2\ldots D$, $j = 1, 2$.
3. Choose G $\mid \gamma, \lambda \sim$ DP (α, G0), where $\lambda = (\lambda 1,\ldots, \lambda W)$ and G0 is a Dirichlet distribution with parameter $\lambda 1\gamma 1,\ldots, \lambda W\gamma W$.
4. Choose $\eta i \mid$ G ~ G, $i = 1, 2\ldots D$.
5. Choose $\eta 0 \mid \gamma, \lambda \sim Dirichlet(\lambda 1(1-\gamma 1)\ldots \lambda W(1-\gamma W))$.
6. Choose $xi\ \gamma \mid \eta i \sim Multinomial(\eta i; Ni1)$, $i = 1\ldots D$.
7. Choose $xi(1-\gamma) \mid \eta 0 \sim Multinomial(\eta 0; Ni2)$, $i = 1\ldots D$.

Ni1 is the total appearances of the discriminative words in document *xi* and *Ni2* is the total appearance of the irrelevant noise words in *xi*. *Ni1* and *Ni2* are both unobservable and considered as latent variable *ηi* denotes the multinomial parameter for the discriminative words in *xi* and *η0*, as the multinomial parameter for the irrelevant noise words, is shared by all the documents in the dataset.

5. EXPERIMENTAL RESULTS

The performance of the models is evaluated by two well-known quality measures (Steinbach, Karypis, & Kumar, [8] 2000). The first is the F-measure, which combines the Precision and Recall ideas from the Information Retrieval literature. F-measure is a measure of a test's accuracy. F-measure is the weighted harmonic mean of precision and recall. The precision *p* and the recall *r* of the test to compute the score. The F-measure is often used in the field of information retrieval for measuring search, document classification, and query classification performance Precision and recall are two widely used metrics for evaluating the correctness of a pattern recognition algorithm. Precision is the probability that a (randomly selected) retrieved document is relevant. Recall is the probability that a (randomly selected) relevant document is retrieved. They can be seen as extended versions of accuracy, a simple metric that computes the fraction of instances Precision can be seen as a measure of exactness or fidelity, whereas recall is a measure of completeness.The Recall, Precision and F-measure of this retrieval are formulated as

$$\text{Recall(l,k)}=\frac{D_{lk}}{D_l},\ \text{Precision(c,k)}=\frac{D_{lk}}{D_k} \quad (1)$$

F (1, k) = 2 * Recall (l, k) * Precision (l, k) / Recall (l, k) (2)

The overall weighted F-measure can be given as:

$$\text{F-measure}=\sum_{l=1}^{L}\frac{D_l}{D}\max\left(\text{F(l,k)}\right) \quad (3)$$

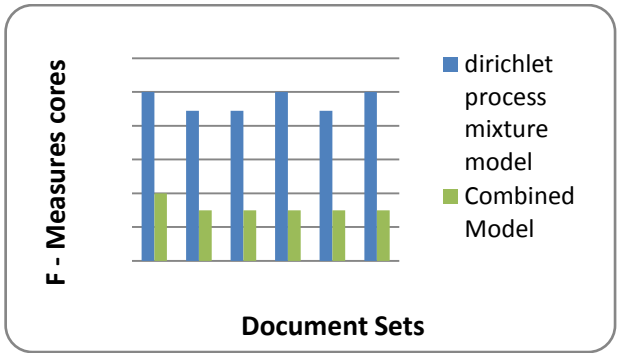

Fig1: F-Measure Scores

The second measure is the Entropy. Entropy indicates how homogeneous a cluster is. If a cluster homogeneity is high then the entropy criterion for that cluster will be low, and vice versa.

The entropy of cluster k is given by:

$$-\sum_{l=1}^{l} p(l,k)\log_2 p(l,k),\ p(l,k)=\frac{D_{lk}}{D_k} \quad (4)$$

The overall entropy is the sum of the entropies of all of the clusters, weighted by their document sizes computed as:

$$\text{Entropy}=\sum_{k=1}^{k}\frac{D_k}{D}E(k) \quad (5)$$

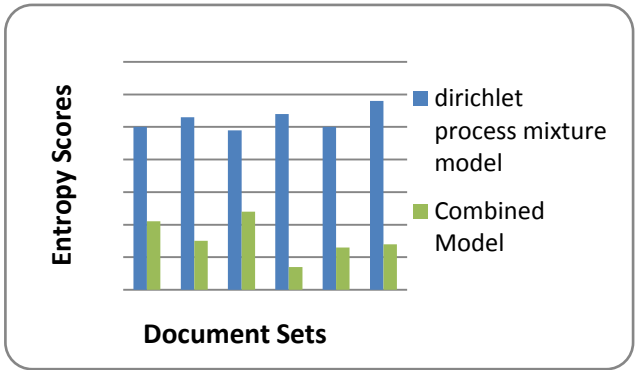

Fig2: Entropy Scores

6. CONCLUSION

A method based on backward elimination approach which is capable of components more than he needed value and then prunes the mixtures to reach their optimum size during the elimination process, proper selection of some latent variables which must be deleted. And its relation to the final performance of the pruned model is straight forward. To treat this problem we introduce a new combined pruning method selects the best options for removal, while keeping a low computational cost, at all. We proposed a novel approach, namely DPMFS [10] method which avoids the drawbacks of model selection and model order reduction method.

7. FUTURE ENHANCEMENT

The scope is to extend the approach to enable the user to have a good overall view of the information contained in the documents. Most classical clustering algorithms assign each data to exactly one cluster, thus forming a crisp partition of the given data, but fuzzy clustering allows for degrees of membership, to which a data belongs to different clusters. In this system, documents are clustered by using fuzzy c-means (FCM) clustering algorithm. FCM clustering is one of well-know unsupervised clustering techniques. It provides the best noise-feature separation and least prediction error.

8. REFERENCES

[1] Tahereh Emami Azadi, FarshadAlmasganj (2009) "Using backward elimination with a new model order reduction algorithm to select best double mixture model for document clustering", Expert Systems with Applications 36 (2009) 10485–10493

[2] M. A. T. Figueiredo and A. K. Jain, "Unsupervised learning of finite mixture models," IEEE Trans. Pattern Anal. Mach. Intell., vol. 24, no.3, pp. 381–396, Mar. 2002.

[3] M. W. Graham and D. J. Miller, "Unsupervised learning of parsimoniousmixtures on large feature spaces," Electrical Engineering Dept., Pennsylvania State, University Park, PA, Tech. Rep., 2004.

[4] Hofmann, T. (1999). Probabilistic latent semantic analysis. In Proceedings of the 22th annual international ACM/SIGIR conference on research and development in information retrieval (pp. 50–57).

[5] D. J. Miller and J. Browning, "A mixture model and EM-based algorithm for class discovery, robust classification, and outlier rejection in mixed labeled/unlabeled data sets," IEEE Trans. Pattern Anal. Mach. Intell.,vol. 25, no. 11, pp. 1468–1483, Nov. 2003.

[6] S.Vaithyanathan and B. Dom, "Generalized model selection for unsupervised learning in high dimensions," in Adv. Neural Inf. Process. Syst., vol. 11, 1999, pp. 970–976.

[7] S. C. Deerwester, S. T. Dumais, T. KLandauer, G.W. Furnas, and R. A. Harshman. Indexing by latent semantic analysis. Journal of the American Society of Information Science, 41 (6):391–407, 1990

[8] Steinbach, M., Karypis, G., & Kumar, V. (2000). A comparison of document clustering techniques. In Proceeding knowledge discovery and data mining (KDD) and workshop text mining. Boston.

[9] E. I. George and R. E. McCulloch. (1992). Variable selection via Gibbs sampling. *Journal of the American Statistical Association*, 88:881-889.

[10] S. Kim. (2006). Variable selection in clustering via Dirichlet process mixture models. Biometrika, 93(4):877-893.

[11] Document Clustering via Dirichlet Process Mixture Model with Feature Selection.GuanYu,Ruizhang Huang,Zhaojun WangKDD'10, July 25-28, 2010, Washington, DC, USA.

[12] Y. W. Teh, M. I. Jordan, M.J. Beal, and D.M. Blei. (2007).Hierarchical Dirichlet Processes. *Journal of the American Statistical Association*, 101(476):1566-1581.

[13] A. Vlachos, Z. Ghahramani, and A. Korhonen. (2008).Dirichlet process mixture models for verb clustering. *ICML Workshop on Prior Knowledge for Text and Language Processing*, Helsinki, Finland.