A LDA-Based Approach for Semi-Supervised Document Clustering

Ruizhang Huang, Ping Zhou, and Li Zhang

Abstract—**In this paper, we develop an approach for semi-supervised document clustering based on Latent Dirichlet Allocation (LDA), namely LLDA. A small amount of labeled documents are used to indicate user's document grouping preference. A generative model is investigated to jointly model documents and the small amount of document labels. A variational inference algorithm is developed to infer the document collection structure. We explore the performance of our proposed approach on both a synthetic dataset and realistic document datasets. Our experiments indicate that our proposed approach performs well on grouping documents based on different user grouping preferences. The comparison between our proposed approach and state-of-the-art semi-supervised clustering algorithms using labeled instance shows that our approach is effective.**

Index Terms—**Semi-supervised clustering, document clustering, latent dirichlet allocation, generative model.**

I. Introduction

Latent Dirichlet allocation (LDA) [1], one important algorithm for topic modeling which shows promising performance in representing text documents with its related topics, has receiving more and more interest in recent years. Besides topic modeling, LDA also shows effective document clustering performance [2]-[6] when regarding latent topics as document partition criteria. The LDA model has become one of the most heavily investigated document clustering approaches due to its ability on dimensionality reduction which is extremely useful for analyzing high-dimensional text documents. One problem for using the LDA approach for document clustering is that documents are grouped by only considering the characteristic of unlabeled documents. In reality, users usually have different document grouping preferences in mind. For example in the news document clustering task, a user can choose to group news documents according to general categories, such as "sports", "finance", etc. Alternatively, another user can also choose to group news documents according to location of news events, such as "China", "American", and "Canada". Therefore, it is useful to let user provide supervised information to guide document clustering. Semi-supervised document clustering, which dealing with the problem of grouping documents with the consideration of a small amount of user-provided information, is a problem derived from the traditional document clustering problem and has receiving considerable attention recently. However, there is no existing semi-supervised document clustering model designed with the LDA model.

Considering the effectiveness of the LDA model on the document clustering problem, in this paper, we investigate a LDA-based model for semi-supervised document clustering, namely LLDA. Labeled documents are used as the type of supervised-information and are used to indicate user's document grouping preferences. A generative model is investigated by using which documents are partitioned by maximizing the joint generative likelihood of text documents and the user-provided document labels. These labels were treated as variables which obey normal distribution and are regressed on the topic proportions. The computational cost of LLDA parameter estimation is also a problem for developing the LLDA model for the semi-supervised document clustering. Traditionally, there are two algorithms to infer LLDA parameters, in particular, the variational inference algorithm and the Gibbs sampling algorithm. Compared with the Gibbs sampling algorithm, the variational inference algorithm shows better computational performance due to the high dimensional representation of text documents. In this paper, we also derived a variational inference algorithm for the LLDA model.

We have conducted extensive experiments on our proposed LLDA model by using both synthetic and realistic datasets. We also compared our approach with state-of-the-art semi-supervised document clustering algorithms with labeled documents as supervised information. Experimental results show that the LLDA model is effective for semi-supervised document clustering.

II. Related Work

A. Semi-Supervised Clustering

Recently, semi-supervised clustering which makes use of a small amount of supervised information to improve clustering accuracy, has attracted much attention. Most semi-supervised clustering algorithms use supervision in the form of document supervision such as labeled instances or instance pairwise constraints for general clustering problems. In this paper, we consider labeled documents as the type of user-provided information. Regarding how supervised information is used, existing semi-supervised clustering methods fall into three categories, namely, constraint-based, distance-based and a combination of the previous two. Constraint-based methods [7]-[12] directly use the constraints to improve clustering algorithms. In [8], the objective function is modified to satisfy paired constraints. Ruiz *et al.* [10] made the clustering process follow the constraint conditions. Cluster seeds are derived from the constraints to initialize the cluster centroids [7], [9]. In [11], a comparative study of investigating annealing process for

Manuscript received March 20, 2014; revised May 22, 2014.

The authors are with the College of Computer Science and Technology, Guizhou University, Guiyang, CO 550025 China (corresponding author: Li Zhang; e-mail: cse.rzhuang@gzu.edu.cn, gs.pzhou11@mail.gzu.edu.cn, lizhang_2004@126.com).

DOI: 10.7763/IJMLC.2014.V4.430

varies model-based semi-supervised document clustering approaches with labeled documents are presented. Recently, Yan *et al.* [12] investigated a semi-supervised fuzzy co-clustering approach. Pairwise constrains are used to improve the objective function. For comparative study, the effectiveness of labeled documents were also discussed. Distance-based methods [4], [13] improve the clustering quality by learning a more accurate distortion measure over the data space using constraints. The distortion measure is trained based on the constraints. In [13], Xing *et al.* presented an algorithm to learn a distance metric representing the examples of similar points. Their method is based on the idea of posing metric learning as a convex optimization problem. The original high-dimensional feature space can be projected into low-dimensional feature subspaces guided by constraints [4].

B. The LDA Model

The latent dirichlet allocation (LDA) model, one of the most important topic probabilistic models, has been proved as a promising approach for the topic modeling. In recent years, researches are conducted to explore the LDA model to other related problems such as the clustering problem [5], [6], [10], [14]-[17]. For the document clustering problem, Shafiei et al. presented a four-level hierarchical Bayesian model for simultaneously clustering documents and terms [15]. Yun et al. combined LDA with explicit human-defined concepts in Wikipedia [6]. Considering the spatial and temporal structure Wang et al. put forward spatial latent dirichlet allocation for computer vision field [5]. In [10], a generative algorithm jointly modeling text and tags is proposed. In addition to the document clustering problem, LDA model is also applied to images. Qi *et al.* [14] used nonparametric LDA to model the panchromatic image collection. In [16], a multiscale LDA approach is proposed to model satellite images. Wu *et al.* [17] mined the correlations between words and images to improve clustering results. There is no existing work on deriving the LDA model for the semi-supervised document clustering problem.

III. LLDA

Formally, we define the following terms:

- A word is an item from a vocabulary indexed by $\{1,2,\ldots,N\}$;
- A document d is represented as a N-dimensional vector $d=(w_1, w_2,\ldots, w_N)$ where w_j is the number of appearance of the word w_j of the document d;
- A document set D is a collection of M documents $\{d_1, d_2,\ldots, d_M\}$;

We aim to find a probability model that assigns high probability not only to reasonable document to cluster assignment but also the high satisfaction of the user-provided document labels.

We introduce a preference variable λ to indicate the user-provided document labels. Our model assumes the generative process for a document d in document set D is as follows:

1) Choose $N \sim Poisson(x)$
2) Choose $q \sim Dir(a)$
3) For each of the N words:
   a) Choose a topic $z_n \mid q \sim Multinomial(q)$
   b) Choose a word $w_n \mid z_n, b_{1,K}$: $Multinomial(b_{1,K})$
4) Choose the preference variable l for labeled document, $l \mid z_{1:N}, h, s^2 \sim N(h^T \bar{z}, s^2)$

where $\bar{z}$ is the average of the topic variable for each word calculated as $(1/N)\sum_{n=1}^{N} z_n$; K is the number of clusters. The graphical representation of the LLDA model is shown in Fig. 1. We partition the document set to two parts, in particular, labeled document set D_L and unlabeled document set D_U. For the unlabeled document set, we aim to find the document partition with the LDA model. For the labeled document set, we aim to find the document partition with both the consideration of document characteristics and the satisfaction of the preference parameter λ for each labeled document.

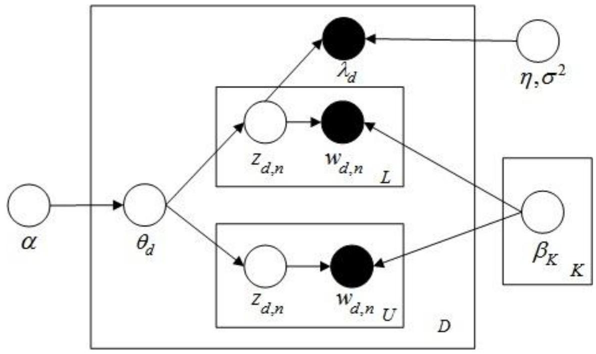


Fig. 1. The graphical representation of the LLDA model.

The joint generative probability of document set and preference variables can be derived by jointly considering the unlabeled document set and the labeled document set. Let M_l denote the number of labeled documents and M_u denote the number of unlabeled documents. The probability of a document set D corpus can be obtained as follows:

$$p(D, l \mid a, b, h, s^2) = \prod_{i=1}^{M_l} p(d_i, l_i \mid a, b, h, s^2) \cdot \prod_{j=1}^{Mu} p(d_j \mid a, b) \tag{1}$$

Notice that the preference variable, λ, with labeled data comes from a normal linear model. By regressing the preference variable on the empirical topic frequencies, we treat the preference variable as nonexchangeable with words. Since dirichlet distribution is the conjugate prior for the parameter of multinomial distribution, the marginal distribution of a labeled document and its preference variable conditioned on the latent variables becomes:

$$p(d, l \mid a, b, h, s^2) = \int p(q \mid a) \prod_{n=1}^{N} \left(\sum_{z_{1:N}} p(z_n \mid q) p(w_{d_n} \mid z_n, b_{1:K}) \right) p(l \mid z_{1:N}, h, s^2) dq \tag{2}$$

The likelihood of unlabeled document is derived from the LDA model which is given as follows:

$$p\left(d \mid a, b\right) = \int p\left(q \mid a\right) \prod_{n=1}^{N} \left(\sum_{z_{1:N}} p\left(z_n \mid q\right) p\left(w_{d_n} \mid z_n, b_{1:K}\right) \right) dq \quad (3)$$

IV. Algorithm

In this section, we present a variational inference algorithm to infer the cluster structure for our proposed LLDA model.

For the unlabeled document set D_U, the marginal distribution of document is identical to the corresponding terms for LDA. Therefore, we only investigate the variational inference algorithm for the labeled document set D_L. Because the variables θ and β are coupled, the posterior distribution of hidden variable is intractable to compute. The fully factorized distribution $q(\theta, z \mid \gamma, \varphi)$ is used to approximate the posterior distribution $p(w_d, \lambda, \theta, z \mid \alpha, \beta)$. The difference between two probability distributions p and q is measured by the KL divergence as follows:

$$D_{KL}\left(p\left(d, l, q, z \mid a, b\right) \parallel q\left(q, z \mid g, f\right)\right) = \int \sum_{z} q\left(q, z \mid g, f\right) \log \frac{p\left(d, l, q, z \mid a, b\right)}{q\left(q, z \mid g, f\right)} dq \quad (4)$$

where $q\left(q, z \mid g, f\right) = q\left(q \mid g\right) \prod q\left(z_n \mid f_n\right)$.

The basic idea of variational inference algorithm is to make use of Jensen's inequality to obtain an adjustable lower bound on the log likelihood for a document.

$$\log p\left(d, l, z_{1:N} \mid a, b_{1:K}, h, s^2\right) = E_q\left[\log p\left(d, l, q, z \mid a, b\right)\right] - E_q\left[\log q\left(q, z \mid g, f\right)\right] \quad (5)$$

Therefore, the lower bound of the log marginal likelihood is as follows:

$$L = E_q\left[\log p\left(d, l, q, z \mid a, b\right)\right] - E_q\left[\log q\left(q, z \mid g, f\right)\right] \quad (6)$$

Given latent topic assignments, the expected log probability of the preference variable is obtained.

$$E_q\left[\log p\left(l \mid z_n, h, s^2\right)\right] = -\frac{1}{2}\log\left(2ps^2\right) - \frac{\left(l^2 - 2lh^T E\left[\bar{z}\right] + h^T E\left[\bar{z}\bar{z}^T\right] h\right)}{2s^2} \quad (7)$$

where

$$E\left[\bar{z}\right] = \bar{f} := \left(1/N\right) \sum_{n=1}^{N} f_n,$$

$$E\left[\bar{z}\bar{z}^T\right] = \left(1/N^2\right)\left(\sum_{n=1}^{N} \sum_{m \neq n} f_n f_m^T + \sum_{n=1}^{N} diag\left\{f_n\right\} \right).$$

To maximize the lower bound L, the update equations for each parameter are as follows:

$$g_i = a_i + \sum_{n=1}^{N} f_{n,i}, \quad (8)$$

$$f_{n,i} = b_{i,w_n} \exp\{y\left(g_i\right) - y\left(\sum_{j=1}^{K} g_j\right) + \left(\frac{l}{Ns^2}\right) h - \frac{\left[z\left(h^T f_{-n,i}\right)h + \left(h \circ h\right)\right]}{2N^2 s^2}\} \quad (9)$$

$$h = E\left(\left[A^T A\right]\right)^{-1} E\left[A\right]^T \quad (10)$$

$$s^2 = \left(1/M\right)\left\{ l\,l - l\,E\left[A\right]\left(E\left[A^T A\right]\right)^{-1} E\left[A\right]^T l \right\} \quad (11)$$

where $i \in \{1,\ldots,K\}$; $n \in \{1,\ldots,N\}$; $f_{-n,i} := \sum_{m \neq n} f_m$; and A is a $M \times (K+1)$ matrix in which each row is the vector $\bar{z}^T$. The detailed algorithm of the LLDA model is shown in Fig. 2. When the improvement of L is less than a threshold, say 10^{-5}, we regarded the LLDA model converge and estimate the latent clustering structure by the variational parameter γ. The cluster to which the document d belongs is determined by the value of γ. In particular, let the γ_i, be the largest value acquired by the document d, d will then be assigned to the cluster labeled by i.

Input: D, α, K, λ

Output: document topic matrix γ

Algorithm:

1. Initialization: randomly initialize β, η, σ^2;
2. **Repeat** until L converge
3. **For** each document d in the dataset
5. Initialization $f_{ni} = \frac{1}{K}$, $g_i = a_i + \frac{N}{K}$
6. **If** d is labeled
7. Update $f_{n,i}$ with the Equation (9);
8. **Else** update $f_{n,i}$ with the ordinary LDA model

$$f_{n,i} = b_{i,w_n} \exp\{y\left(g_i\right) - y\left(\sum_{j=1}^{K} g_j\right) \quad (12)$$

8. Update γ with the Equation (8);
9. Update η with the Equation (10);
10. Update σ^2 with the Equation (11) ;
11. Calculate L with Equation (6)

Fig. 2. The LLDA Algorithm.

V. Experiment

Two sets of experiments were conducted to evaluate the performance of the LLDA model. For the first experiments, the clustering result of LLDA is evaluated using a synthetic dataset. For the second experiments, our proposed approach is evaluated via real document datasets.

A. Evaluation Metric

Normalized mutual information that refers to NMI [18] can be used as clustering evaluation metric. NMI is an external measure, mainly used to evaluate the effect of clustering on a data set and the degree of similarity of the real division of the data set. The *NMI* value is between 0 and 1. The higher the *NMI* value is, the more perfectly the clustering results match the user-labeled class assignments. This evaluation metric is used in our experiments. *NMI* is estimated as follows:

$$NMI = \frac{\sum_{h,l} d_{h,l} \log\left(\frac{d \cdot d_{h,l}}{d_h \cdot c_l}\right)}{\sqrt{\left(\sum_h d_h \log\left(\frac{d_h}{d}\right)\right)\left(\sum_l \log\left(\frac{c_l}{d}\right)\right)}} \qquad (13)$$

where d is the number of documents, d_h is the number of documents in class h, c_l is the number of documents in cluster l and $d_{h,l}$ is the number of documents in class h as well as in cluster l.

B. Synthetic Dataset

Dataset and Experimental Setup. We derived a synthetic dataset to evaluate the effectiveness of the LLDA model on partitioning data points based on different user grouping preferences. The synthetic dataset consists of 200 data points with 600 features. Data points are generated from 4 classes, in particular, T_{AC}, T_{AD}, T_{BC} and T_{BD}. Each class is derived from 2 subclasses. Specifically, the class T_{AC} contains data points from subclasses T_A and T_C. The class T_{AD} contains data points from subclasses T_A and T_D. The class T_{BC} contains data points from T_B and T_C. The class T_{BD} contains data points from T_B and T_D. Each subclass has 150 distinctive features generated from a general multinomial distribution. 50 data points were then generated from each class by randomly selecting features from the two related subclasses. Taking the number of clusters K as 2, the synthetic dataset can be organized in 2 different ways. Data points can be organized from the perspective of the subclass T_A and T_B. In particular, we regard data points from T_{AC} and T_{AD} as in one cluster, while data points from T_{BC} and T_{BD} as in the other cluster. On the other hand, data points can be organized from the perspective of the subclasses T_C and T_D. In particular, we regard data points from T_{AC} and T_{BC} as in one cluster, while data points from T_{AD} and T_{BD} as in the other cluster.

In our proposed algorithm for this synthetic dataset, we set α=1. For each experiment setting, we ran our proposed approach 10 times. The performance is computed by taking the average of these 10 experiments.

Experimental Performance. We conducted experiments for the LLDA model on labeled as user preferences on the plots. We investigated the clustering performance by varying the percentage of the labeled data points from 0 to 50%. When the percentage of labeled data points is set to 0, the LLDA model is reduced to the ordinary LDA model. The experimental results are depicted in Fig. 3 and Fig. 4.

Noticed that the LLDA model tends to group the data points to the T_C and T_D when no labeled data points are provided. The reason is because features of each data point are not evenly but randomly selected from the two underlying subclasses. In our generated synthetic dataset, T_C and T_D contribute more to the generation of data points and provide more discriminative features. However, with a small number of labeled documents, the LLDA model is able to organize the data points in the right the direction indicated by the labeled points.

Guided by labeled data points, the LLDA model is able to organize data points differently for the same set of data points. Perfect clustering results are achieved when the percentage of labeled data points are small. Therefore, the LLDA model is effective on discovering different data grouping preferences.

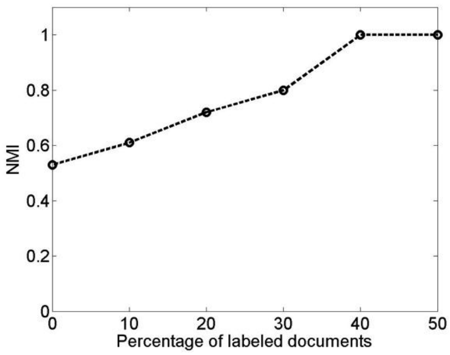


Fig. 3. Clustering performance of the LLDA model on the synthetic dataset. Data points are organized from the perspective of the subclass T_A and T_B indicated by the labeled data points.

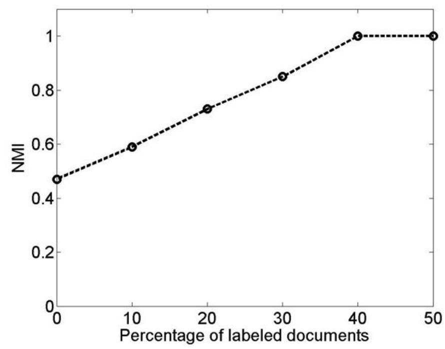


Fig. 4. Clustering performance of the LLDA model on the synthetic dataset. Data points are organized from the perspective of the subclass T_C and T_D indicated by the labeled data points.

C. Real Document Datasets

Datasets and Experimental Setup. Two real-world document datasets are used to evaluate our proposed LLDA model, in particular, the *re0* dataset[1] and the *Yahoo_k1* dataset[2]. The *re0* dataset is derived from the *Reuters-21578* collection. This collection contains messages collected from 13 different categories. The *Yahoo_k1* dataset is from the WebACE project. Each document corresponds to a web page listed in the subject hierarchy of Yahoo. We pre-processed the two datasets by removing headers and stop-words. Low-frequency words that occur less than 0.5% are also removed. The purpose of such processing is to eliminate those words that obviously not define the latent cluster structure. A summary of the datasets used in this paper is shown in Table I.

TABLE I: SUMMARY DESCRIPTION OF DATASETS (M: NUMBER OF DOCUMENTS, K: NUMBER OF CLUSTERS, N: NUMBER OF WORDS)

Datasets	M	K	N
re0	1504	13	2837
Yahoo_k1	2340	6	3671

Parameters Discussion. We investigated the sensitivity of the choices of the LLDA model parameters.

[1] http://www.daviddlewis.com/resources/testcollections/reuters21578
[2] ftp://ftp.cs.umn.edu/dept/users/boley/pddpdata/doc-K

Choice of α. We investigated the sensitivity of the choice of parameters α that influenced the distribution of topics. We simulated with different values of α where α was set to be 0.1, 0.5, 1.0 and 5.0 under the LLDA model. For four different values of α, K was fixed as 13 on *re0* dataset and 6 on *Yahoo_k1* dataset. The percentage of labeled documents is 5%. Our proposed approach achieved stable clustering results in all these experiments as shown in Fig. 5. This indicates that our model is robust to the choice of α.

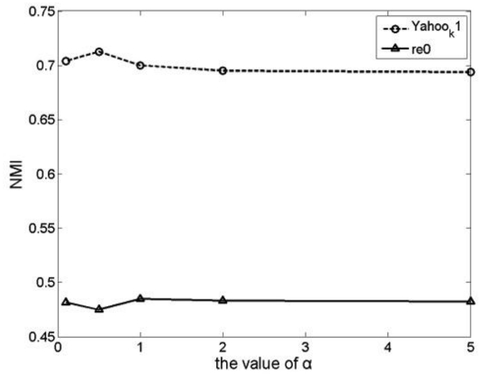
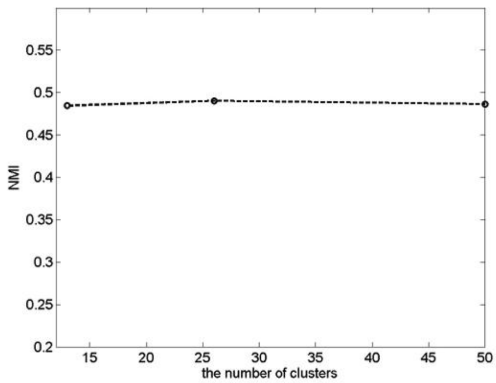
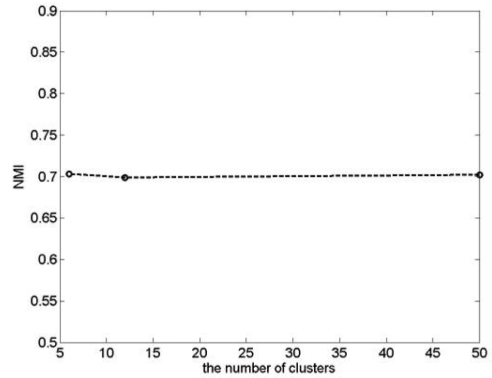


Fig. 5. Document clustering performance for the LLDA model on the *re0* and *Yahoo_k1* datasets when α is chosen to be different values



(a) *re0*

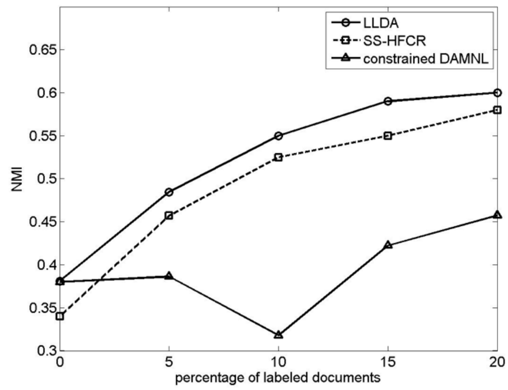


(b) *Yahoo_k1*

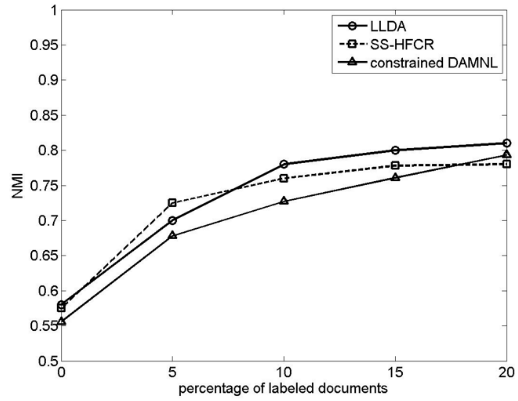
Fig. 6. Document clustering performance for the LLDA model on the *re0* and the *Yahoo_k1* datasets when K is set with different values.

Choice of K. The parameter K affects the number of clusters to which documents belong. Some care is needed to choose this parameter in a reasonable range since a much larger value for it will result in a model with more computing time. On *re0* dataset, we experimented with different values of K where K was set to 13, 26 and 50. On *Yahoo_k1* dataset, the parameter K was set to 6, 12 and 50. α was fixed as 0.01 and the percentage of labeled documents was set to 5%. Experimental results as shown in Fig. 6 indicate that K does not affect much to the document partition performances when K is set to a larger value. Most documents are partitioned to a reasonable number of clusters and leave the rest of clusters assigned with a small amount of outlier documents that contribute less to the document clustering performance.

In the following experiments, we set α to 1 and set K to the correct number. The parameter β was initialized randomly. We investigated the clustering performance by varying the percentage of the labeled data points from 0 to 20%. For each experiment setting, we conducted experiments 10 times and chose the result that acquired the largest value of Equation (5). The time complexity of the LLDA model is $O(MKN)$ where M is the number of documents, K is the number of clusters, N is the number of words and τ denotes the number of iterations.



(a) *re0*



(b) *Yahoo_k1*

Fig. 7. Document clustering performance for the LLDA model, the constrained-DAMNL, and the SS-HFCR model on the *re0* and the *Yahoo_k1* datasets.

Experimental Performance. For comparative investigation, two state-of-the-art semi-supervised document clustering approaches [11], [12] that use labeled documents as supervised information were investigated, labeled as constrained-DAMNL and SS-HFCR respectively, Fig. 7 shows the experimental performances of our proposed LLDA model, the constrained-DAMNL, and the SS-HFCR model on the *re0* and the *Yahoo_k1* datasets. Noticed that when the percentage of labeled data points is set to 0, the LLDA model is reduced to the ordinary LDA model. The experimental results show that our proposed LLDA model performs better than the LDA model with a small amount of labeled documents. When the number of labeled documents increases, the LLDA model performs better. Therefore, it is useful to incorporate a small amount of labeled documents to guide document clustering. Moreover, the LLDA model generally performs the best comparing with the constrained-DAMNL model and the SS-HFCR model for all experiments. When the percentage of labeled documents is 5% on the *Yahoo_k1* dataset, our proposed LLDA model performs slightly worse than the SS-HFCR model. One possible reason is due to the randomly generation of supervised information. The quality of the document labels cannot be controlled. When the number of supervised information is small, there may not be sufficient informative hints for directing document clustering provided which results in slightly worse document clustering performance. However, when the number of labeled documents increases, the LLDA model achieves better performance than both of the SS-HFCR model and constrained-DAMNL model. Therefore, our proposed LLDA model is effective on determining document partition based on different user grouping preferences.

VI. Conclusions and Future Work

In this paper, we proposed LLDA model handles document clustering with labeled instance. In our model, the document labels could be obtained by user's judgment or authentic resource. We treat document labels as preference variable follows normal distribution. The variational inference technique is used to estimate parameters. Our experiment shows that LLDA model groups document dataset into meaningful clusters with document labels provided by users. The comparison of our algorithm with some existing state-of-the-art algorithms indicates that our approach is more robust and effective for semi-supervised document clustering when the user's willing are satisfied. Our analysis of the experiment result also shows that supervised information inserted in the LDA model could reinforce the positive impact of labels and therefore improve the clustering quality.

An interesting direction for future research is to study how applying active learning approach to our proposed semi-supervised document clustering approach. Most semi-supervised clustering algorithms use supervision in the form of document supervision such as labeled instances or instance pairwise constraints for general clustering problems. The active learning approach can be incorporated to select document pairs while LLDA model is used with pairwise constraints.

References

[1] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent Dirichlet allocation," *Journal of Machine Learning Research*, vol. 3, pp. 993-1022, 2003.

[2] J. R. Millar, G. L. Peterson, and M. J. Mendenhall, "Document clustering and visualization with latent dirichlet allocation and self-organizing maps," in *Proc. 22nd International Florida Artificial Intelligence Research Society Conference*, pp. 69-74, Florida: AAAI Press, 2009.

[3] D. Ramage, P. Heymann, and D. Christopher, "Manning and hector Garcia-Molina, clustering the tagged web," in *Proc. Second ACM International Conference on Web Search and Data Mining*, pp. 54-63, Barcelona: ACM Press, 2009.

[4] W. Tang, H. Xiong, S. Zhong, and J. Wu, "Enhancing semi-supervised clustering: a feature projection perspective," in *Proc. 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 707-716, San Jose: ACM Press, 2007.

[5] X. Wang and E. Grimson, "Spatial latent Dirichlet allocation," in *Proc. 21st Annual Conference on Neural Information Processing Systems*, Vancouver: Curran Associates Inc, 2007.

[6] J. Yun, L. P. Jing, H. K. Huang, and J. Yu, "Multi-view LDA for semantics-based document representation," *Journal of Computational Information Systems*, vol. 7, no. 14, pp. 4999-5006, 2011.

[7] S. Basu, A. Banerjee, and R. J. Mooney, "Semi-supervised clustering by seeding," in *Proc. 9th International Conference on Machine Learning*, pp. 19-26, Sydney: Morgan Kaufmann Press, 2002.

[8] S. Basu, M. Bilenko, and R. J. Mooney, "A probabilistic framework for semi-supervised clustering," in *Proc. 10th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 59-68, New York: ACM Press, 2004.

[9] M. Bilenko,. S. Basu, and R. J. Mooney, "Integrating constraints and metric learning in semi-supervised clustering," in *Proc. 21st International Conference on Machine Learning*, pp.81-88, Banff: ACM Press, 2004.

[10] C. Ruiz, M. Spiliopoulou, and E. Menasalvas, "C-DBSCAN: Density-Based clustering with constraints," in *Proc. 11th International Conference on Rough Sets, Fuzzy Sets, Data Mining and Granular Computing*, pp. 216-223, Toronto: Springer Press, 2007.

[11] S. Zhong, "Semi-supervised model-based document clustering: A comparative study," *Machine Learning*, vol. 65, no. 1, pp. 3-29, 2006.

[12] Y. Yan, L. Chen, and W.-C. Tjhi, "Fuzzy semi-supervised co-clustering for text documents," *Fuzzy Sets and Systems*, vol. 215, pp. 74-89, 2013.

[13] E. P. Xing, A. Y. Ng, M. I. Jordan, and S. Russell, "Distance metric learning with application to clustering with side-information," *Advances in Neural Information Processing Systems*, vol. 15, pp. 505-512, MIT Press, 2002.

[14] Y. F. Qi, H. Tang, Y. Shu, L. Shen, J. W. Yue, and W. G. Jiang, "An object-oriented clustering algorithm for VHR panchromatic images using nonparametric latent Dirichlet allocation," in *Proc. 32nd IEEE International Geoscience and Remote Sensing Symposium*, pp. 2328-2331, Munich: Institute of Electrical and Electronics Engineers Inc Press, 2012.

[15] M. M. Shafiei and E. E Milios, "Latent Dirichlet co-clustering," in *Proc. Sixth International Conference on Data Mining*, pp. 542-551, Hong Kong : IEEE Computer Society Press, 2006.

[16] H. Tang, L. Shen, and Y. F. Qi, "A multiscale latent Dirichlet allocation model for object-oriented clustering of VHR panchromatic satellite images," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 53, no. 3, pp. 1680-1692, 2013.

[17] F. Wu, Y. H. Han, Y. T. Zhuang, and J. Shao, "Clustering web images by correlation mining of image-text," *Journal of Software*, vol. 21, no. 7, pp. 1561-1575, 2010.

[18] A. Strehl, J. Ghosh, and R. Mooney, "Impact of similarity measures on Web-page clustering," in *Proc. the Workshop on Artificial Intelligence for Web Search*, pp. 58-64, Austin: AAAI Press, 2000.

Ruizhang Huang received her B.S. degree in computer science from the Nankai University, China, in 2001 and the Mphil. and PhD. degrees in the systems engineering & engineering management from the Chinese University of Hong Kong, Hong Kong, in 2003 and 2008. In 2007, she joined the Hong Kong Polytechnic University, Hong Kong, as a lecturer. Since the year 2011, She has been with the Guizhou University as an associate professor. She is an active researcher on the area of data mining, text mining, machine learning, and information retrieval. She has published a number of papers including prestigious journals and conferences.

Ping Zhou was born in 1988, who received her B.S. degree in computer science from the Changzhou University, China, in 2009. Since the year 2011, She has been with the Guizhou University as a master degree candidate.

Li Zhang received her B.S. degree and the Mphil. degrees in the Department of Computer Science and Technology from the Guizhou University, China, in 1993 and 2007. Since the year 1993, She has been with the Guizhou University as a university lecturer. She is an active researcher on the area of algorithm optimization. She has published a number of papers including prestigious journals and conferences.