Collective Latent Dirichlet Allocation

Zhi-Yong Shen[1,2], Jun Sun[1,2], and Yi-Dong Shen[1]
[1]State Key Laboratory of Computer Science,
Institute of Software, Chinese Academy of Sciences, Beijing 100190, China
[2]Graduated University, Chinese Academy of Sciences, Beijing 100049, China
{zyshen,junsun,ydshen}@ios.ac.cn

Abstract

In this paper, we propose a new variant of Latent Dirichlet Allocation(LDA): Collective LDA (C-LDA), for multiple corpora modeling. C-LDA combines multiple corpora during learning such that it can transfer knowledge from one corpus to another; meanwhile it keeps a discriminative node which represents the corpus ID to constrain the learned topics in each corpus. Compared with LDA locally applied to the target corpus, C-LDA results in refined topic-word distribution, while compared with applying LDA globally and straightforwardly to the combined corpus, C-LDA keeps each topic only for one corpus. We demonstrate that C-LDA has improved performance with these advantages by experiments on several benchmark document data sets .

1. Introduction

Modeling the content of documents is a standard task of information retrieval, text ming and natural language processing. Latent Dirichlet Allocation (LDA) based topic models [2, 11] have attracted much attention recently due to their ability of discovering the low-dimensional semantic structures of a corpus. In LDA, the documents are assumed to be sampled from a mixture distributions over latent topics; meanwhile each topic is characterized by a distribution over words. By considering a prior probability on these distributions, LDA establishes a complete generative model for the corpus. There are dozens of LDA based models including: temporal text mining [14], author- topic analysis [12], supervised topic models [1], latent Dirichlet co-clustering [10] and LDA based bio-informatics [3]. Most of these models are designed for a single corpus while in practice, we always face numerous corpora such as newsgroups, web pages and scientific papers. In this paper, we consider how LDA can be used to model multiple corpora collectively.

Transfer learning is an hot area in machine learning and data mining domains recently, which emphasizes the transferring of knowledge across different domains or tasks. The performance of learning models can be improved by knowledge transferred from extra (even can be irrelevant) *auxiliary* data sets. For example, Wu and Dietterich [15] propose how to adjusting SVM classifiers with auxiliary data sources. Raina et al. [9] investigate learning logistic regression classifiers by incorporating labeled data from irrelevant categories through constructing informative prior from the irrelevant labeled data. Raina et al. [8] propose a new learning technique − self-taught learning which uses irrelevant unlabeled data to enhance the classification performance. As mentioned in [9], LDA may also model a corpus better with auxiliary corpora. This is realistic in the human reading behavior since readers always use knowledge across the reading domains. A straightforward application of LDA on multiple corpora is to combine the target corpus with auxiliary corpora and treat the combination as a single corpus. Although by this means, knowledge in each corpus can be transferred to the others, there are some shortcomings: First, the supervised information − from which corpus a document comes, is discarded. Second, the learned topics are across all corpora, which may not satisfy the learning objective that the learned topics should be specific to the target corpus.

In this paper, we propose a new variant of LDA: Collective LDA (C-LDA), for multiple corpora modeling. C-LDA combine multiple corpora when learning such that it can transfer knowledge from one corpus to another; meanwhile it keeps a discriminative node which represents the corpus ID to constrain the learned topics in target corpus. Compared with LDA locally applied on the target corpus, C-LDA results in refined topic-word distribution, while compared with applying LDA globally and straightforwardly on the combined corpus, C-LDA keeps learned topics only for the target corpus. By experiments on several benchmark document data sets, we demonstrate that C-LDA has significantly improved performance with these advantages. Teh et al. [13] propose Hierarchical Dirichlet Processes (HDP) which also can learn topics from multiple corpora. How-

ever, the objective of HDP is to share topics among different corpora while our setting is to learn topics specifically for the target corpus.

Moreover, since a categorized corpus can be divided into sub-corpora according to the categories, C-LDA can also employ such supervised information to improve the performance of LDA. In this case C-LDA can be viewed as a implementation of supervised LDA. There are some supervised LDA models in literatures. Blei and McAuliffe introduce supervised LDA (sLDA) in [1]. Flaherty et al. [3] propose *labeled* LDA (lLDA) for biological data. Both of these works try to refine the latent semantic structure learned by LDA using some supervised information. However, sLDA is designed for continuous response values and lLDA uses the labels associated with each gene, i.e. with each word as in the text mining domain. Li et al. [5] present a LDA based model integrating category information to learn natural scene categories of images whose goal is for the categories not the latent semantic structures. Therefore, C-LDA's extended application for documents with categorical labels is also novel.

2. Collective Latent Dirichlet Allocation

We firstly introduce the original LDA model. The graphical model representations of LDA and C-LDA are shown in Figure 1.

Original LDA models describe a process to generate a corpus of documents represented by *bags of words*. Suppose there are T topics and a vocabulary with totally V unique words. First, for each document d, LDA samples a multinomial distribution θ_d over topics from a Dirichlet distribution with parameter α; second, LDA samples a topic z_{di} for the i-th word in document d from the multinomial θ_d; third, emit the word w_{di} from the multinomial distribution associated to z_{di} with parameter ϕ_z which can be smoothed by another Dirichlet distribution with parameter β. The graphical representation of this process is shown in Figure 1 (a). The goal of learning LDA is to find the configuration of latent variables that explain the observed corpus best.

It is natural to apply LDA directly to the combination of multiple corpora such that each document is represented as a random mixture over uniform latent topics and each topic is specified by a distribution over a uniform vocabulary. We'll call this global strategy as LDA-G in the rest of this paper and the strategy that apply original LDA on the target corpus locally will be denoted as LDA-L. The intuition behind our work is to adjust the original LDA such that the topics are specified on the uniform vocabulary while the topic of a document are constrained by the document's corpus ID. The constraint is modeled by assuming that a documents corpus ID is generated from a multinomial distribution (ψ) according to its topic distribution. The graphical model of the proposed generative probabilistic model is shown in Figure 1 (b). We also give the Gibbs sampling process for parameter estimation presented in the next subsection:

Table 1. Notations

SYMBOL	DESCRIPTION
T	Number of topics
D	Number of documents
C	Number of corpora
V	Number of words
v	A word in vocabulary
d	A document in a corpus
c	the corpus ID of a document or a token
N_d	Number of tokens in document d
θ	the multinomial distribution over topics
ϕ	the multinomial over words
ψ	the multinomial distribution over corpora
α (β, γ)	the Dirichlet prior for θ (ϕ, ψ)
$\mathbf{w}_d$	the word set of document d
w_{di}	the i-th word token in document d
z_{di}	the topic associated with w_{di}

1. Draw T multinomials from a Dirichlet prior γ over corpus IDs; and draw T multinomials from a Dirichlet prior β over words; then for each document:
2. Draw $N_d \sim$ Poisson(ξ); draw $\theta_d \sim$ Dirchilet(α); then for each of the N_d token w_{di}:
   (a) Draw a topic $z_{di} \sim$ Multinomial(θ_d)
   (b) Draw a token $w_{di} \sim$ Multinomial($\phi_{z_{di}}$).
   (c) Draw a corpus ID $c_{di} \sim$ Multinomial($\psi_{z_{di}}$).

The graphical representation of the sampling process is shown in Figure 1(c). Since all the corpus IDs of words in a document are observed as the same as the document's corpus ID. Therefore, the constraint on the topics according to the corpus ID still remains. Moreover, like that mentioned in [14], the corpus ID generated for each document (Figure 1 (b)) can be sampled by rejection or importance sampling from a mixture of per-topic multimomial distributions over corpus IDs.

2.1. Parameter Estimation

LDA models are too complex for exact learning, thus there are some approximate learning means available in the literature: variational methods, expectation propagation and Gibbs sampling. We choose Gibbs sampling and this part of work follows [11, 14, 4]. According to Figure 1 (c), the complete probability model is:

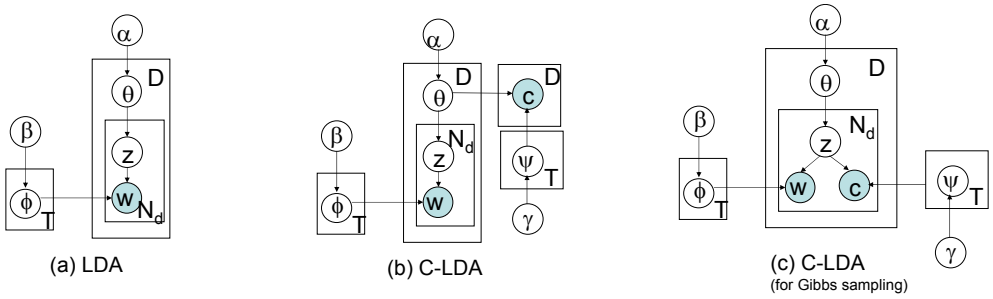

Figure 1. Graphical representations of LDA and the proposed C-LDA. Shaded nodes are observed.

$$
\begin{aligned}
\theta_d|\alpha &\sim \text{Dirichlet}(\alpha) \\
\phi_z|\beta &\sim \text{Dirichlet}(\beta) \\
\psi_z|\gamma &\sim \text{Dirichlet}(\gamma) \\
z_{di}|\theta_d &\sim \text{Multinomial}(\theta_d) \\
w_{di}|\phi_{z_{di}} &\sim \text{Multinomial}(\phi_{z_{di}}) \\
c_{di}|\psi_{z_{di}} &\sim \text{Multinomial}(\psi_{z_{di}})
\end{aligned}
$$

So the conditional posterior distribution for z_{di} is given by:

$$
\begin{aligned}
P(z_{di} = j|\mathbf{z}_{-di}, \mathbf{w}, \mathbf{c}) \propto \quad & P(c_{di}|z_{di} = j, \mathbf{z}_{-di}, \mathbf{c}_{-di}) \\
& \times P(w_{di}|z_{di} = j, \mathbf{z}_{-di}, \mathbf{w}_{-di}) \\
& \times P(z_{di} = j|\mathbf{z}_{-di})
\end{aligned} \tag{1}
$$

In the right side of (1), $\mathbf{w}_{-di}$ is discarded in the first term and $\mathbf{c}_{-di}$ also disappears in the condition of w. This is because w and c are independent to each other when z is given. The parameters θ, ϕ and ψ are omitted in the above expression because we'll integrate each of the terms on the right hand over the values of these parameters. For the first term, we have:

$$
\begin{aligned}
& P(c_{di}|z_{di} = j, \mathbf{z}_{-di}, \mathbf{c}_{-di}) \\
= \quad & \int P(c_{di}|z_{di} = j, \psi_j) P(\psi_j|\mathbf{z}_{-di}, \mathbf{c}_{-di}) d\psi_j
\end{aligned} \tag{2}
$$

where (using the Bayes' rule),

$$
P(\psi_j|\mathbf{z}_{-di}, \mathbf{c}_{-di}) \propto P(\mathbf{c}_{-di}|\psi_j, \mathbf{z}_{-di}) P(\psi_j) \tag{3}
$$

Since $P(\psi_j) \sim \text{Dirichlet}(\gamma)$ and conjugate to the multinomial $P(\mathbf{c}_{-di}|\psi_j, \mathbf{z}_{-di})$, the posterior distribution of $P(\psi_j|\mathbf{z}_{-di}, \mathbf{c}_{-di})$ will be $\text{Dirichlet}(\gamma + n_{z_{di}}^{(c)-1})$, where $n_z^{(c)}$ is the number of instances of corpus ID c assigned to topic z. The term -1 is for excluding the current one. Moreover, only the corpus ID assigned to topic j can influence the posterior distribution of ψ_j and $P(c_{di}|z_{di} = j, \psi_j) = \psi_{c_{di},j}$, we have:

$$
P(c_{di}|z_{di} = j, \mathbf{z}_{-di}, \mathbf{c}_{-di}) = \frac{n_{z_{di}}^{c_{di}} + \gamma - 1}{n_{z_{di}}^{(.)} + C\gamma - 1} \tag{4}
$$

where $n_{z_{di}}^{(.)}$ is the total number of tokens assigned to topic z_{di}, almost in the same way:

$$
\begin{aligned}
& P(w_{di}|z_{di} = j, \mathbf{z}_{-di}, \mathbf{w}_{-di}) \\
= \quad & \int P(w_{di}|z_{di} = j, \phi_j) P(\phi_j|\mathbf{z}_{-di}, \mathbf{w}_{-di}) d\phi_j
\end{aligned} \tag{5}
$$

then

$$
P(w_{di}|z_{di} = j, \mathbf{z}_{-di}, \mathbf{w}_{-di}) = \frac{n_{z_{di}}^{w_{di}} + \beta - 1}{n_{z_{di}}^{(.)} + W\beta - 1} \tag{6}
$$

for the last term of (1), we have:

$$
P(z_{di} = j|\mathbf{z}_{-di}) = \int P(z_{di} = j|\theta_d) P(\theta_d|\mathbf{z}_{-i}) d\theta_d \tag{7}
$$

and analogously we get:

$$
P(z_{di} = j|\mathbf{z}_{-di}) = \frac{n_{z_{di}}^{d} + \alpha - 1}{n_{(.)}^{d} + T\alpha - 1} \tag{8}
$$

where $n_{z_{di}}^{d}$ is the number of tokens in d assigned to topic z_{di} and $n_{(.)}^{d}$ is the total number of tokens in d. Finally, from (4), (6), (8) and (1), and considering the denominator in (8): $n_{(.)}^{d} + T\alpha - 1$ is a constant value for document d we have:

$$
\begin{aligned}
P(z_{di} = j|\mathbf{z}_{-di}, \mathbf{w}, \mathbf{c}) \quad \propto \\
(n_{z_{di}}^{(d)} + \alpha - 1) \quad \times \frac{n_{z_{di}}^{c_{di}} + \gamma - 1}{n_{z_{di}}^{(.)} + C\gamma - 1} \quad \times \frac{n_{z_{di}}^{w_{di}} + \beta - 1}{n_{z_{di}}^{(.)} + W\beta - 1}
\end{aligned} \tag{9}
$$

For LDA-L or LDA-G the above conditional probabilities become:

$$
P(z_{di} = j|\mathbf{z}_{-di}, \mathbf{w}) \propto (n_{z_{di}}^{(d)} + \alpha - 1) \times \frac{n_{z_{di}}^{w_{di}} + \beta - 1}{n_{z_{di}}^{(.)} + W\beta - 1} \tag{10}
$$

Table 2. Corpora used in this paper.

	newsgroup	sector	sraa	Total
Multi-1	Newsgroup-1 rec.autos rec.motorcycles rec.sport.baseball rec.sport.hockey	Sector-1 air.courier.industry oil.and.gas.operations.industry airline.industry oil.well.services.and.equipment coal.industry railroad.industry misc.transportation.industry trucking.industry oil.and.gas.integrated.industry water.transportation.industry	Sraa-1 realauto	*Doc: 9641* *tokens:906114* *words: 23000*
Multi-2	Newsgroup-2 talk.politics.guns talk.politics.misc talk.politics.mideast talk.religion.misc	Sector-2 banking.sector investment.services.industry consumer.financial.services. misc.financial.services.industry insurance.sector	Sraa-2 realauto	*Doc: 9728* *tokens:1201059* *words: 26973*
Total	*Doc.: 7987* *tokens: 1048519* *words: 27252*	*Doc.: 1819* *tokens: 346723* *words: 9992*	*Doc.: 4781*	

2.2. Parameter Initialization

In some literature [11, 12, 14], the hyper-parameters α and β are set symmetrically as $\alpha = 50/T$ and $\beta = 0.1$ for text mining. We follow these settings and set the new hyper-parameter $\gamma = \alpha$. These hyper-parameters determine how heavily the empirical distributions are smoothed and can be chosen to give the desired balance or skewness in the resulting multinomial distribution. We can estimate the values of all these hyper-parameters of the proposed model from data using EM like algorithms with the moment match technique as proposed in [7] and used in [10].

In original LDA, the setting of the topic number T is flexible and can be automatically learned by non-parametric Bayes methods. However, it is not so straightforward to set the topic numbers for each sub-corpus when we model the combined corpus. Both C-LDA and LDA-G will face this issue and there are various strategies to handle it. The following strategy is used in our experiments: we first apply a LDA-L process for each sub-corpus with specific topic numbers to *burn in* a initialization of topic assignments. Then C-LDA or LDA-G is applied to the combined corpus following this initialization. For LDA-L, we just need to continue the burning in process.

3. Experimental Results

In this section, we present experimental results of comparison between C-LDA and LDA. We demonstrate the ability of C-LDA by lower perplexity values over various numbers of topics. We also investigate the benefit from various number of auxiliary corpora comparing with LDA-L. Finally, we demonstrate the extended application of C-LDA for a single categorized corpus.

3.1. Data Sets and Evaluation Criterions

We fit the models on numerous text corpora to compare their abilities with various evaluation criterions.

3.1.1 Data Sets

The multiple corpora data sets are generated from three text corpora sources including *20-Newsgroups* (shortening: *newsgroup*), *Industry Sector* (shortening:*sector*) and *SRAA*. All these data are downloaded from a researcher's home page, to which we refer for the detailed data information[1]. The Bow toolkit [6] is used to retrieve the documents and generate the bag of words matrices.

All the data sets have two levels of categories. We divide the three corpora by the top categories and choose totally six parts (two from each) as given in Table2. The second level categories and some brief information are listed. We generate two multiple text corpora: *Multi-1* and *Multi-2*, by combining the rows of this table. They are used for experiments on collective learning, i.e. each sub-corpus is the target corpus and meanwhile the others are auxiliary. In this case, the models are set blind to the low level categories and we call this part of experiments as *Collective Modeling* (Section 3.2). We also combine the parts from the same data source by column and treat them as single corpora. The second level categories are now supervised information to be employed by C-LDA. This part of experiment is indicated as *Supervised Modeling* (Section 3.3).

To evaluate the predictive ability of the models, we compute *perplexity* which is a standard measure for estimating the performance of a probabilistic model in language modeling. The formal definition of perplexity for a corpus $\mathcal{D}$ with D documents is:

$$perplexity(\mathcal{D}) = exp\{-\frac{\sum_{d=1}^{D} \log p(\mathbf{w}_d)}{\sum_{d=1}^{D} N_d}\} \qquad (11)$$

The perplexity is monotonically decreasing in the likelihood of the data and a lower perplexity means better modeling performance[2].

3.2. Collective Modeling

In this part, target corpora are sub-corpora in Multi-1,2. We firstly compare perplexity values of LDA-L, LDA-G and C-LDA over each sub-corpus with increasing topic numbers from 30 to 80 by step 10. The topic assignments to word tokens are initialized with 50 iterations of LDA-L for each sub-corpus as described in Section 2.2. Then we run the models with the same initializations for another 50 rounds. The average results from five rounds are presented in the first three columns of Figure 2. LDA-G performs slightly better than LDA-L over the sub-corpora generated

[1] http://www.cs.umass.edu/ mccallum/code-data.html

[2] Perplexity is conventionally used on a held out test set with the purpose of examining the generalization of the model. The objective of our work is to integrating extra data sets including test sets into training process, thus we make no division of train and test sets in our experiments.

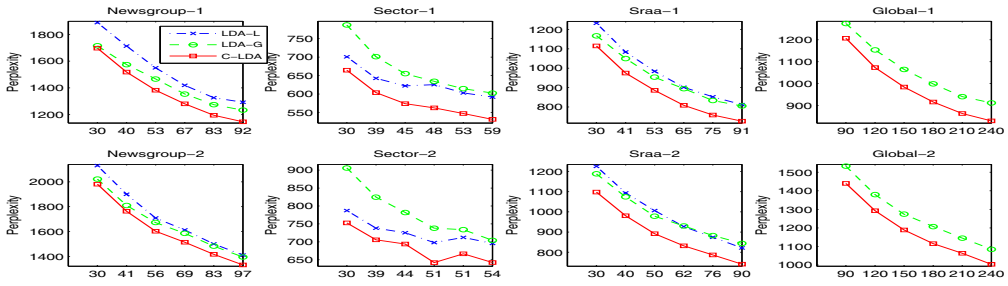

Figure 2. The perplexity comparison on the target corpora. X-axis: number of topics in C-LDA.

from *Newsgroup* and *Sraa* which demonstrates the benefits from transferring; meanwhile it performs worse over sub-corpora generated from *Sector*. This deems that contents of sub-corpora generated from *Sector* prefer the topics learned locally. C-LDA works better than LDA-G and LDA-L on all target sub-corpora by constraining the learned topics locally while transferring the knowledge from other sub-corpora. We also evaluate the global perplexity of LDA-G and C-LDA and results are shown in the last column of Figure 2. C-LDA naturally outperforms LDA-G over entire corpus since it works better on each sub-corpus.

To examine the ability of constraining the learned topics in target corpora, we compute co-occurrence matrices of low level categories and topics where the counts are for word tokens and the matrices are normalized by columns. We initialize 20 topics for each sub-corpus and then apply LDA-G or C-LDA with 50 iterations. We can see the matrices generated by C-LDA are much *clearer* than those by LDA-G. In matrices from LDA-G, the topics are all across entire combined corpora especially those for Sector-1,2. This also explains why in the Second column of Figure 2, the performance of LDA-G is the worst. Furthermore, if we directly take the topic assignments as the cluster assignments, then the C-LDA can be viewed a strategy for semi-supervised co-clustering. This is because the known top level category information can be treated as constraints for semi-supervised clustering. The co-occurrence matrices are computed on low level categories thus clearer blocks means better clustering performance.

To analyze the impact from various numbers of auxiliary corpora, we discard Sraa-1,2 from Multi-1,2 and then applying the C-LDA. For the sub-corpora Newsgroup-1,2 and Sector-1,2, the numbers of auxiliary corpora decrease from two to one. Note that LDA-L can be viewed as C-LDA with zero number of auxiliary corpus. We give the perplexity results with various numbers of auxiliary corpora for Newsgroup-1,2 and Sector-1,2 in Figure 4. The results from using auxiliary corpora are significantly better than those without them. However, the comparison between using two or one auxiliary corpora implies that some auxiliary corpora might benefit the modeling less significantly. We can see sub-corpora $Sraa$-1,2 help to decrease the perplexity of Newsgroup-1,2 only a little and they do no help for the modeling of Sector-1,2. Otherwise, they do not make the results significantly worse either.

3.3. Supervised Modeling

We now evaluate the model's extended ability to employ the supervised category information. Actually, if we treat the corpus IDs of the multiple corpora as the category labels, this has been proved. For more convincible demonstration, we vertically combine the *Newsgroup* and *Sector* corpora in Table 2 and generate two single corpus with known low level categories. We use no initialization strategy in this part of experiments since there's no need to set a desired number of topics for each category. The average results of five rounds are given in Figure 5. It seems in the *Newsgroup* corpus, the category label information significantly helps to decrease the model perplexity while in the *Sector* corpus, this information benefits less.

4. Conclusions

In this paper, we discuss a novel derivation for LDA named C-LDA with capability for modeling multiple corpora. The key idea of the proposed model is to employ auxiliary corpora for better topic representations while constraining each topic to only one corpus. Since the model transfers knowledge from other data sets, it can be viewed as a transfer learning version of LDA. With various experiments on some benchmark corpora, we demonstrate C-LDA outperforms the original LDA models that ignore auxiliary corpora (LDA-L) or simply combine auxiliary corpora into the target corpus (LDA-G). Moreover, C-LDA can be directly extended for a single corpus with category information. In this case, C-LDA becomes a kind of supervised LDA and its advantage is also empirically demonstrated.

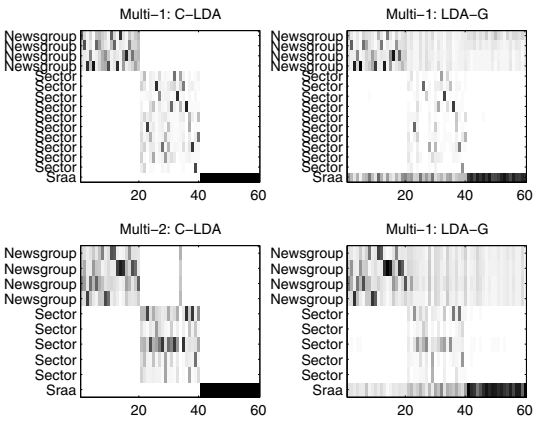

Figure 3. The co-occurrence matrix for low level categories and topics. We only label the corpus IDs of these categories for simplicity.

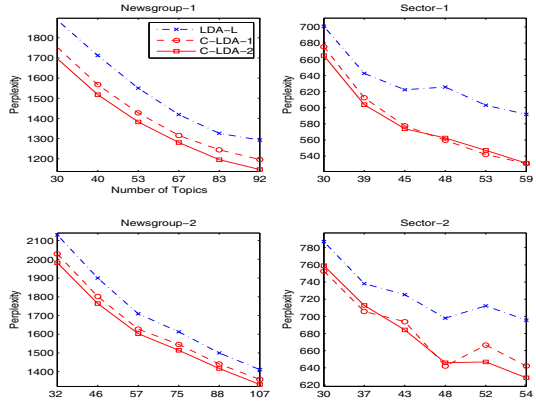

Figure 4. The perplexity results with various number of auxiliary corpora. C-LDA-1 denotes collective LDA with one auxiliary corpus and C-LDA-2 denotes two of them. Note that LDA-L can be viewed as C-LDA-0.

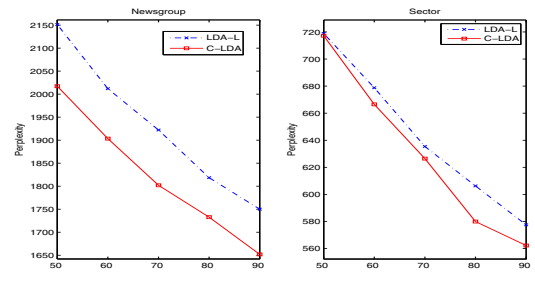

Figure 5. The perplexity comparison on single corpora. C-LDA employs the low level category labels while LDA-L ingnores.

5 Acknowledgments

This work is supported in part by NSFC grants 60673103 and 60721061.

References

[1] D. M. Blei and J. D. McAuliffe. Supervised topic models. In *Proceedings of the 21th In Advances in Neural Information Processing Systems*, 2007.

[2] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent dirichlet allocation. *Journal of machine Learning Research*, 3:993–1022, 2003.

[3] P. Flahertya, G. Giaever, J. Kumm, M. I. Jordan, and A. P. Arkin. A latent variable model for chemogenomic profiling. *bioinformatics*, pages 3286–3293, 2005.

[4] T. Griffiths. Gibbs sampling in the generative model of latent dirichlet allocation. www-psych.stanford.edu/ gruffydd/cogsci02/lda.ps, 2002.

[5] F. Li and P. Perona. A bayesian hierarchical model for learning natural scene categories. In *In Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 524–531, 2005.

[6] A. K. McCallum. Bow: A toolkit for statistical language modeling, text retrieval, classification and clustering. http://www.cs.cmu.edu/ mccallum/bow, 1996.

[7] T. P. Minka. Estimating a dirichlet distribution. Technical report, MIT, 2000.

[8] R. Raina, A. Battle, H. Lee, B. Packer, and A. Y. Ng. Self-taught learning: transfer learning from unlabeled data. In *Proceedings of the Twenty-fourth International Conference on Machine Learning*, page 759C766, 2007.

[9] R. Raina, A. Y. Ng, and D. Koller. Constructing informative priors using transfer learning. In *Proceedings of the Twenty-third International Conference on Machine Learning*, page 713C720, 2006.

[10] M. M. Shafiei and E. E. Milios. Latent dirichlet co-clustering. In *Proceedings of the Sixth International Conference on Data Mining*, pages 542 – 551, 2006.

[11] M. Steyvers. Probabilistic topic models. *Latent Semantic Analysis: A Road to Meaning*, pages 427–448, 2007.

[12] M. Steyvers, P. Smyth, and T. Griffiths. Probabilistic author-topic models for information discovery. In *Proceedings of the 10th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 306–315, 2004.

[13] Y. W. Teh, M. I. Jordan, and M. J. B. D. M. Blei. Hierarchical dirichlet processes. *Journal of The American Statistical Association*, 101:1566–1581, 2006.

[14] X. Wang and A. McCallum. Topics over time: a non-markov continuous-time model of topical trends. In *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 424–433, 2006.

[15] P. Wu and T. G. Dietterich. Improving svm accuracy by training on auxiliary data sources. In *Proceedings of the 21-st International Conference on Machine Learning*, pages 110–117, 2004.