Constrained-hLDA for Topic Discovery in Chinese Microblogs

Wei Wang, Hua Xu, Weiwei Yang, and Xiaoqiu Huang

State Key Laboratory of Intelligent Technology and Systems,
Tsinghua National Laboratory for Information Science and Technology,
Department of Computer Science and Technology, Tsinghua University,
Beijing 100084, China
{ww880412,alexalexhxqhxq}@gmail.com,
xuhua@tsinghua.edu.cn, ywwbill@163.com

Abstract. Since microblog service became information provider on web scale, research on microblog has begun to focus more on its content mining. Most research on microblog context is often based on topic models, such as: Latent Dirichlet Allocation(LDA) and its variations. However,there are some challenges in previous research. On one hand, the number of topics is fixed as a priori, but in real world, it is input by the users. On the other hand, it ignores the hierarchical information of topics and cannot grow structurally as more data are observed. In this paper, we propose a semi-supervised hierarchical topic model, which aims to explore more reasonable topics in the data space by incorporating some constraints into the modeling process that are extracted automatically. The new method is denoted as constrained hierarchical Latent Dirichlet Allocation (constrained-hLDA). We conduct experiments on Sina microblog, and evaluate the performance in terms of clustering and empirical likelihood. The experimental results show that constrained-hLDA has a significant improvement on the interpretability, and its predictive ability is also better than that of hLDA.

Keywords: Hierarchical Topic Model, Constrained-hLDA, Topic Discovery.

1 Introduction

In the information explosion era, social network not only contains relationships, but also much unstructured information such as context. Furthermore, how to effectively dig out latent topics and internal semantic structures from social network is an important research issue. Early work on microblogs mainly focused on user relationship and community structure. [1] studied the topological and geographical properties of Twitter. Others work such as [2] studied user behaviors and geographic growth patterns of Twitter. Only little research on content analysis of microblog was proposed recently. [3] was mainly based on traditional text mining algorithms. [4] proposed MB-LDA by overall considering contactor relevance relation and document relevance relation of microblogs. In this paper, we propose a novel probabilistic generative model based on hLDA, called constrained-hLDA, which focuses on both text content and topic hierarchy.

V.S. Tseng et al. (Eds.): PAKDD 2014, Part II, LNAI 8444, pp. 608–619, 2014.

Previous work on microblog text was mainly based on LDA. To our best knowledge, there was little research on the topic hierarchy on microblog text. However, hierarchical topic modeling is able to obtain the relations between topics. [5] proposed an unsupervised hierarchical topic model, called hierarchical Latent Dirichlet Allocation (hLDA), to detect automatically new topics in the data space after fixing the level. Based on the stick-breaking process, [6] proposed the fully nonparametric hLDA without fixing the level. After that, some modifications of hLDA were proposed [7–9]. Given a parameter L indicating the depth of the hierarchy, hLDA makes use of nested Chinese Restaurant Process(nCRP) to automatically find useful sets of topics and learn to organize the topics according to a hierarchy in which more abstract topics are near the root of the hierarchy and more concrete topics are near the leaves. However, the traditional hLDA is an unsupervised learning which does not incorporate any prior knowledge. In this paper, we attempt to extract some prior knowledge and incorporate them to the sampling process.

The rest of the paper is organized as follows. Section 2 introduces the previous work related to this paper. Section 3 describes the hLDA briefly. Section 4 introduces the novel model constrained-hLDA. The experiment is introduced in Section 5, which is followed by the conclusion in Section 6.

2 Related Work

There have been many variations of probabilistic topic models, which was first introduced by [10]. The probabilistic topic model is based on the idea that documents are generated by mixtures of topics which is a multinomial distribution over words. One limitation of Hofmann's model is that it is not clear how the mixing proportions for topics in a document are generated. To overcome this limitation, [11] propose Latent Dirichlet Allocation(LDA). In LDA, the topic proportion of every document is a K-dimensional hidden variable randomly drawn from the same Dirichlet distribution, where K is the number of topics. Thus, generative semantics of LDA are complete, and LDA is regarded as the most popular approach for building topic models in recent years[12–16].

LDA is a useful algorithm for topic modeling, but it fails to draw the relationship between one topic and another and fails to indicate the level of abstract for a topic. To address this problem, many models have been proposed to build the relations, such as hierarchical LDA(hLDA) [5, 6], Hierarchical Dirichlet processes(HDP) [17], Pachinko Allocation Model(PAM) [18] and Hierarchical PAM(HPAM) [19] etc. These models extend the "flat" topic models into hierarchical versions for extracting hierarchies of topics from text collections. [6] proposed the most up-to-date hLDA model, which is a fully nonparametric model. It simultaneously learns the structure of a topic hierarchy and the topics that are contained within that hierarchy. Furthermore, it can also learn the most appropriate levels and hyper-parameters although it is time-consuming. In recent years, some modifications of hLDA has also been proposed. [7] proposed a supervised hierarchical topic model, called hierarchical Labeled Latent Dirichlet Allocation(hLLDA), which uses hierarchical labels to automatically build corresponding topic for each label. [8] propose an unsupervised hierarchical topic model, called

Semi-Supervised Hierarchical Latent Dirichlet Allocation (SSHLDA), which can not only make use of the information from the hierarchy of observed labels, but also can explore new latent topics in the data space. Although our work has some slight resemblance with their work, there still exist several important differences:

1. Our constrained-hLDA mainly focuses on the text of microblogs or reviews without observed labels.
2. The prior knowledge is extracted automatically from the corpus instead of first-hand observation.
3. The constraints are alterable by different parameters.

3 Preliminaries

The nested Chinese restaurant process (nCRP) is a distribution over hierarchical partitions[5, 6]. It generalizes the Chinese restaurant process (CRP), which is a single parameter distribution over partitions of integers. It has been used to represent the uncertainty over the number of components in a mixture model. The generative process is as follow:

1. There are N customers entering the restaurant in sequence, which is labeled with the integers $\{1, ..., N\}$.
2. First customer sits at the first table.
3. The nth customer sit at:
   (a) Table i with probability $\frac{n_i}{\gamma+n-1}$, where n_i is the number of customers currently sitting at table i, which has been occupied.
   (b) A new table with probability $\frac{\gamma}{\gamma+n-1}$.
4. After N customers have sat down, their seating plan describes a partition of N items.

In the nested CRP, suppose there are an infinite number of infinite-table Chinese restaurants in a city. One restaurant is identified as the root restaurant and its every table has a card with the name that refers to another restaurant. This structure repeats infinitely many times, thus, the restaurants in the city are organized into an infinitely branched, infinitely-deep tree. When a tourist arrives at the city, he selects a table, which is associated with a restaurant at next level, using the CRP distribution at each level. After M tourists have visited in this city, the path collection, which they selected, describes a random subtree of the infinite tree.

Based on identifying documents with the paths generated by the nCRP, the hierarchical topic model, which consists of an infinite tree, is defined. Each node in the tree is associated with a topic, which is a probability distribution across words. Each document is assumed to be generated by a mixture of topics on a path from the root to a leaf. For each token in the document, one picks a topic randomly according to the distribution, and draws a word from the multinomial distribution of that topic. To infer the topic hierarchy, the per-document paths $\mathbf{c}_d$ and the per-word level allocation to topics in those paths $z_{d,n}$ must be sampled. Then we will introduce the process briefly.

For the path sampling, the path associated with each document conditioned on all other paths and the observed words need to be sampled. Assume the depth is finite and let T denotes it, the posterior distribution of path $\mathbf{c}_d$ is as denote:

$$p(\mathbf{c}_d|\mathbf{w}, \mathbf{c}_{-d}, \mathbf{z}, \eta, \gamma) \propto p(\mathbf{c}_d|\mathbf{c}_{-d}, \gamma)p(\mathbf{w}_d|\mathbf{c}, \mathbf{w}_{-d}, \mathbf{z}, \eta) \tag{1}$$

In Equation 1, two factors influence the probability that a document belongs to a path. The first factor is the prior on paths implied by the nested CRP. The second factor is the probability of observing the words in the document given a particular choice of path with equation organized as follows:

$$p(\mathbf{w}_d|\mathbf{c}, \mathbf{w}_{-d}, \mathbf{z}, \eta) = \prod_{t=1}^{T} \frac{\Gamma(n^{(\cdot)}_{c_d,t,-d} + V\eta)}{\prod_w \Gamma(n^{(w)}_{c_d,t,-d} + \eta))} \frac{\prod_w \Gamma(n^{(w)}_{c_d,t,-d} + n^{(w)}_{c_d,t,d} + \eta))}{\Gamma(n^{(\cdot)}_{c_d,t,-d} + n^{(\cdot)}_{c_d,t,d} + V\eta)} \tag{2}$$

where $n^{(w)}_{c_d,t,-d}$ is the number of word w that have been allocated to the topic indexed by $c_{d,t}$, not including those in the current document, V denotes the total vocabulary size, and $\Gamma(\cdot)$ is the standard gamma function. When $\mathbf{c}$ contains a previously unvisited restaurant, $n^{(w)}_{c_d,t,-d}$ is zero.

After selecting the current path assignments, the level allocation variable $z_{d,n}$ for word n in document d conditioned on the current values of all other variables need to be sampled as:

$$p(z_{d,n}|\mathbf{z}_{-(d,n)}, \mathbf{c}, \mathbf{w}, m, \pi, \eta) \propto p(w_{d,n}|\mathbf{z}, \mathbf{c}, \mathbf{w}_{-(d,n)}, \eta)p(z_{d,n}|\mathbf{z}_{d,-n}, m, \pi) \tag{3}$$

where $\mathbf{z}_{-(d,n)}$ and $\mathbf{w}_{-(d,n)}$ are the vectors of level allocations and observed words leaving out $z_{d,n}$ and $w_{d,n}$, $\mathbf{z}_{d,-n}$ denotes the level allocations in document d , leaving out $z_{d,n}$.

4 Constrained-hLDA

In this section, we will introduce a constrained hierarchical topic model, i.e., the constrained herarchical Latent Dirichlet Allocation(constrained-hLDA). As we have known, similar to LDA, the original hLDA is a purely unsupervised model without considering any pre-existing knowledge. However, in semi-supervised clustering framework, the prior knowledge can help clustering algorithm produce more meaningful clusters. In our algorithm, the extracted prior knowledge can help to pre-establish a part of the infinite tree structure. In this section, we will give an introduction to the constraint extraction and the proposed constrained-hLDA which can use pre-existing knowledge expressed as constraints.

4.1 Path Constraints Extraction

To construct constrained hierarchical topic model, we adopt hLDA and incorporate the constraints from the pre-existing knowledge. Compared with hLDA, constrained-hLDA has one more input for improving path sampling. The input is a set of constrained indicators, which is in the form of $\{\{w_{1,1}, w_{1,2}, \ldots\}, \ldots, \{w_{N,1}, w_{N,2}, \ldots\}\}$. Each subset

$\{w_{i,1}, w_{i,2}, \ldots\}$, which corresponds to a node in constrained-hLDA, consists of several high correlation words. In our work, these words, which can indicate the correlation of a path and a document, are called **constrained indicators**. These corresponding nodes, which are pre-allocated several constrained indicators, are called **constrained nodes**.

The intuition of above idea is very simple and easy to follow. In this paper, we just attempt to solve it based on a correlation approach, more novel and efficient method will be further explored in the future. Algorithm 1 summarizes the main steps of constraints extraction. First, the FP-tree algorithm is adopted to extract the one-dimension frequent items according to the minimum support and maximum support(Line 1). The maximum support is used to filter some common words in order to make sure that the occurrences of each candidate are close, therefore, there will not be hierarchical relationship of these frequent items. Next, for each fis_i, it is added to an empty collection CS_i first, and then the correlation of fis_i with other items is computed. If the correlation of fis_i and fis_j is greater than the given threshold, it is assumed that fis_i and fis_j should constitute a must-link and fis_j is appended to CS_i (Line 2 - Line 9). In this work, the correlation is calculated by overlap as follows:

$$overlap(A, B) = \frac{P_{A\&B}}{min(P_A, P_B)} \tag{4}$$

where $P_{A\&B}$ is the co-occurrence of word A and word B, P_A is the occurrence of word A, and P_B is the occurrence of word B. The range of equation 4 is between $[0, 1.0]$, so the threshold can be easily given for different corpora. In the end, we delete the same set only retaining one from CS (Line 10 - Line 14). Based on Algorithm 1, the prior set CS, each of which contains several high correlation indicators, can be acquired. In this paper, the threshold of overlap is set as 0.4, the maximum support is set as five times as minimum support, all these parameters are estimated number. Additionally, we attempt to utilize different minimum supports to obtain different set so that different experiment results can be made for sure.

Algorithm 1. Constraints extraction

1. Frequent Item Set $FIS \leftarrow$ FP-tree (D, min_sup, max_sup)
2. **for** each fis_i in FIS **do**
3. $\quad CS_i \leftarrow fis_i$
4. $\quad$ **for** each $fis_{j|j!=i}$ in FIS **do**
5. $\quad\quad$ **if** $correlation(fis_i, fis_j) > threshold$ **then**
6. $\quad\quad\quad CS_i \leftarrow fis_j$
7. $\quad\quad$ **end if**
8. $\quad$ **end for**
9. **end for**
10. **for** each CS_i in Constraint Set CS **do**
11. $\quad$ **if** there exists $CS_j == CS_i$ **then**
12. $\quad\quad$ delete CS_i from CS
13. $\quad$ **end if**
14. **end for**

4.2 Path Constraints Incorporation

To integrate constrained indicators into hLDA, we extend the nCRP to a more realistic situation. Suppose the root restaurant has infinite tables, some tables have a menu containing some special dishes. Suppose N tourists arrive at the city, some of them have a list of special dishes that they want to taste. When a tourist enters into the root restaurant, if he has a list, he will select a table whose menu contains the special dishes of his list. Otherwise, according to his willingness to taste the special dishes, he will use CRP equation to select a table among those tables without menus. To keep it simple in this paper, we assume only the root restaurant has menus.

In constrained-hLDA model, each constrained set CS_i corresponds to a menu and each constrained indicator corresponds to a special dish. Then the documents in a corpus are assumed drawn from the following generative process:

1. For each table $k \in T$ in the infinite tree
   (a) Draw a topic $\beta_k \sim Dir(\eta)$
2. For each document,$d \in \{1, 2, \ldots, D\}$
   (a) Let c_1 be the root node.
   (b) For level $l = 2$:
      i. If d contains the constrained indicators $\{i_{d,1}, i_{d,2}, \ldots\}$, select a table c_2 with probability $\frac{n_{i_{d,i}}+\gamma}{\sum(n_{i_{d,i}}+\gamma)}$, where $n_{i_{d,i}}$ is the number of table which contains $i_{d,i}$.
      ii. Otherwise, draw a table c_2 among $C_{nm,2}$ from restaurant using CRP, where $C_{nm,2}$ is the set of tables which have no menus on root restaurant.
   (c) For each level $l \in \{3, \ldots, L\}$
      i. Draw a table c_l from restaurant using CRP.
   (d) Draw a distribution over levels in the tree, $\theta_d|\{m, \pi\} \sim GEM(m, \pi)$
   (e) For each word $n \in \{1, 2, \ldots, N\}$
      i. Choose level $z|\theta_d \sim Discrete(\theta_d)$.
      ii. Choose word $w|\{z, \mathbf{c}, \beta\} \sim Discrete(\beta_{c_z})$, which is parameterized by the topic associated with restaurant c_z.

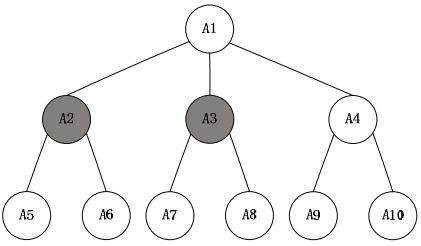

Fig. 1. One illustration of constrained-hLDA. The tree has 3 levels. The shaded nodes are constrained topics, which is pre-defined. The circled nodes are latent topics. After learning, each node in this tree is a topic, which is a corresponding probability distribution over words.

As the example shown in Figure 1, we assume that the height of the desired tree is $L = 3$, and the constrained-topics extracted are $\{A2, A3\}$. The constrained topics amount to the tables containing menu, each of which is pre-defined as the constrained indicators coming from a CS_i, and the constrained indicators amount to special dishes. In our work, because microblogs are mainly short texts, the maximum level is truncated to 3. Furthermore, it is notable that the constraints can be extended to the deeper level. For example, the constrained set can be extracted again from the documents which pass by the node $A2$, and then the constrained indicators set corresponding to $A2$ can be drawn from these documents.

In constrained-hLDA, the idea of incorporating prior knowledge derives from [20], and the most important process is incorporating the constraints to the path sampling process according to the probabilities calculated using Equation 5:

$$p(\mathbf{c}_d|\mathbf{w}, \mathbf{c}_{-d}, \mathbf{z}, \eta, \gamma) \propto (\eta'\delta(\mathbf{w}_d, \mathbf{c}_d) + 1 - \eta')p(\mathbf{c}_d|\mathbf{c}_d, \gamma)p(\mathbf{w}_d|\mathbf{c}, \mathbf{w}_{-d}, \mathbf{z}, \eta) \quad (5)$$

where $\delta(\mathbf{w}_d, \mathbf{c}_d)$ is an indicator function, which indicates whether the nodes from $\mathbf{c}_d$ contain the same constrained indicator with that of $\mathbf{w}_d$: If $\mathbf{w}_d$ contains such node, $\delta(\mathbf{w}_d, \mathbf{c}_d) = 1$, otherwise, $\delta(\mathbf{w}_d, \mathbf{c}_d) = 0$. The hard constraint indicator can be relaxed by η', Let $0 \leq \eta' \leq 1$ be the strength of our constraint, where $\eta' = 1$ recovers a hard constraint, $\eta' = 0$ recovers unconstrained sampling and $0 < \eta' < 1$ recovers a soft constraint sampling.

4.3 Level Constraints Extraction and Incorporation

After revising the path sampling process and selecting a particular path, some prior knowledge can also integrate into level sampling process. As we have known, hLDA can discover the function words in root topic, furthermore, these words have no effect on the document interpretability and often appear in many documents. Therefore, we hope to improve level sampling process by pre-discriminating some function words and non-function words. In our work, the function words are discriminated according to the Part-Of-Speech(POS) and the term frequency in each document. Algorithm 2 describes our purpose, where RD_w denotes the ratio of the documents containing the word w in the current corpus. For each word, if its RD_w is greater than the given threshold

Algorithm 2 Constraints extraction

1. **for** each w_i in current document **do**
2. **if** $RD_w > threshold_{upper}$ AND $POS_w \notin S_{POS}$ **then**
3. $samleLevel_w \leftarrow 0$
4. **else if** $RD_w < threshold_{below}$ AND $POS_w \in S_{POS}$ **then**
5. $samleLevel_w \leftarrow 1, \ldots, K$
6. **else**
7. $samleLevel_w \leftarrow 0, \ldots, K$
8. **end if**
9. sample the level according $samleLevel_w$
10. **end for**

$threshold_{upper}$ and it does not belong to the pre-defined POS set S_{POS}, it would be likely to be a function word that is allocated to root node directly(Line 2 - Line 3). If its RD_w is less than the given threshold $threshold_{below}$ and it belongs to the pre-defined POS set S_{POS} at the same time, it would be likely to be a non-function word without being allocated to root node(Line 4 - Line 5). Finally, we sample the level according to these prior knowledge (Line 9). In this paper, $threshold_{upper}$ and $threshold_{below}$ are set to 0.02 and 0.005, and the pre-defined POS set S_{POS} is set as noun, adjective and verb.

5 Experiment

5.1 Data Sets

Due to the lack of standard data set for this kind of research yet, we collected the experiment data from sina microblog[1] by ourselves. It is generally known that Ya'an Earthquake[2] on 20th, April, 2013 was a catastrophe shocking everyone, which is exactly an ideal hot issue for research. We crawled 19811 microblog users all coming from Ya'an, and also crawled their posted microblogs from 8am 20th April 2013 to 8am 25th April 2013. There are 58476 original microblogs released by these users, each of which contains several sentences. As time passed by, people's concern level on this issue would decline gradually, therefore, we use the data on a daily level for further analysis. Table 1 depicts the data sets for evaluation. The designed experiments and sampling results can also be referred in [6]. For hLDA and constrained-hLDA, there is a restriction that documents can only follow a single path in the tree. In order to make each sentence of a document can follow different paths, we split texts into sentences, such a change can get a remarkable improvement for hLDA and constrained-hLDA in the corpus of microblogs. In our experiment, hLDA algorithm is completed with Java codes by ourselves according to [6]. In our constrained-hLDA, the stick-breaking procedures are truncated at three levels to facilitate visualization of results. The topic Dirichlet hyper-parameters are fixed at $\eta = \{1.0, 1.0, 1.0\}$, The nested CRP parameter γ is fixed at 0.5, the GEM parameters are fixed at $\pi = 100$ and $m = 0.25$.

Table 1. Experiment data

Time	Token	Number of microblogs	Number of sentences
20-21	T1	19709	50631
21-22	T2	11678	32311
22-23	T3	9779	28215
23-24	T4	9308	27213
24-25	T5	8002	23159

[1] `http://weibo.com/`

[2] `http://en.wikipedia.org/wiki/2013_Lushan_earthquake`

5.2 Hierarchy Topic Discovery

Figure 2 depicts the hierarchical structure of cluster results. It is natural to conclude that the constrained-hLDA can well discover the underlying hierarchical structure of the content of micorblogs, and each topic and its child node mainly relate to pre-allocated the constrained indicator, which is the underlined word. For example, there are three sub-topic of *Ya'an*, the first sub-topic relates to *blessing*, the second talks about the *situations* of *Ya'an*, the third talks about *relief* of *Ya'an*. Furthermore, as we can find, the latent topic of second level is a meaningless topic, which is hard to summarize the interpretability of these topics. This phenomenon illustrates that the irrelevant information in microblog context that can be filtered well by our algorithm.

5.3 Comparison with hLDA

In this section, we compare the experimental results with hLDA, and the per-document distribution over levels is truncated at three levels. In order to evaluate our model, we use predictive held-out likelihood as a measure of performance to compare the two approaches quantitatively. The procedure is to divide the corpus into D_1 observed documents and D_2 held-out documents, and approximate the conditional probability of the held-out set given the training set:

$$p(\mathbf{w}_1^{held-out}, \ldots, \mathbf{w}_{D_2}^{held-out} | \mathbf{w}_1^{obs}, \ldots, \mathbf{w}_{D_1}^{obs}) \tag{6}$$

For this evaluation method, more details can be found in [6].

Figure 3 depicts the performance of constrained-HLDA on several data sets by different minimum support. Table 2 depicts the best performance of different constraints on several data sets. According to these experimental results, we can conclude that: (1) Both path sampling constraints and level sampling constraints can improve hLDA. (2) The smaller minimum support can obtain more constrained indicators so that it can achieve better log likelihood. (3) The likelihood of constrained-hLDA is better than the likelihood of hLDA, but for different corpus, the degree of improvement is different. When the topic of corpus is more concentrated, the improvement seems to be better.

Table 2. The Best Results of Different Prior Constraints(800 samplers)

Data set token	hLDA	hLDA + level constrains	hLDA + path constraints	constrained-hLDA
T1	-233776.355	-226566.45	-225814.798	-218583.987
T2	-169646.036	-164147.048	-162871.358	-158950.632
T3	-137735.633	-133976.967	-134931.846	-130895.533
T4	-130254.675	-127269.889	-128493.374	-124552.455
T5	-106223.172	-104269.033	-104670.728	-100796.114

In order to avoid interference from the values of hyperparameters, as with [6]'work, we also interleave Metropolis-Hastings (MH) steps between iterations of the Gibbs sampler to obtain new values of m,π,γ and η. Table 3 present the results by sampling the hyperparameters in the same case, from which we can see that constrained-hLDA still performs better than hLDA.

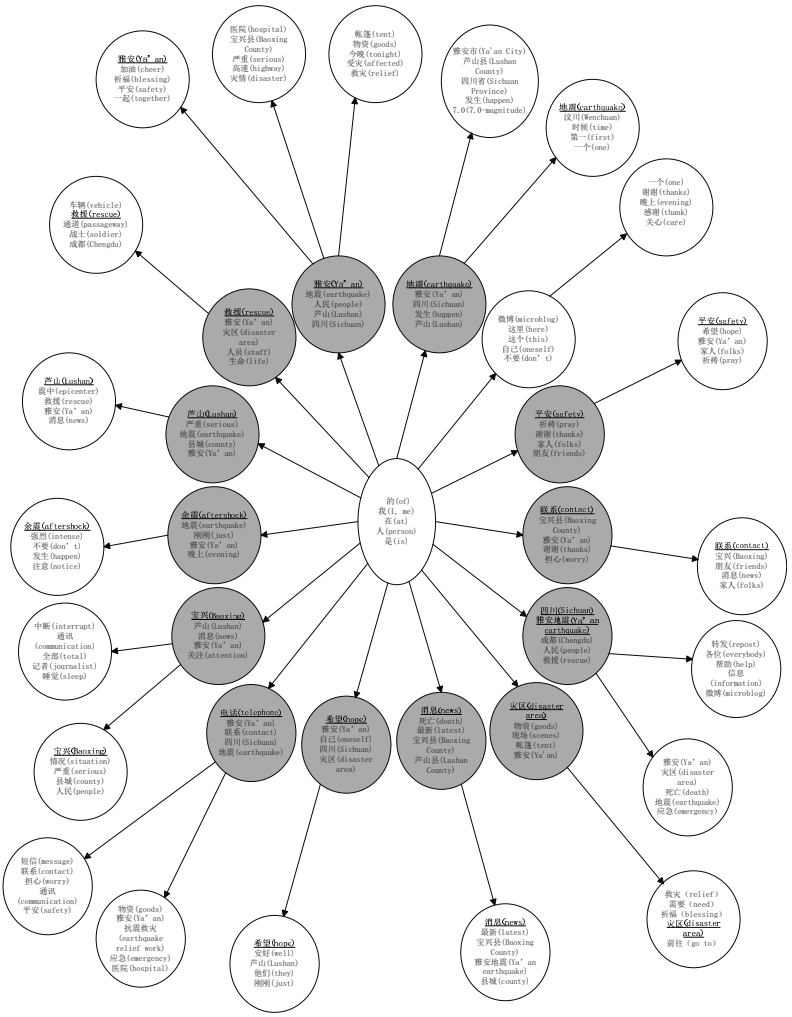

Fig. 2. A portion of the hierarchy learned from T1 Data. The shaded nodes are constrained topics, the bold and underlining words are the constrained indicators extracted by Algorithm 1.

Table 3. The Results by sampling the hyperparameters (800 samplers)

Data set token	hLDA	constrained-hLDA
T1	-210794.652	-195034.999
T2	-151989.286	-140812.547
T3	-123816.852	-115215.844
T4	-117760.271	-110516.351
T5	-94291.395	-90791.294

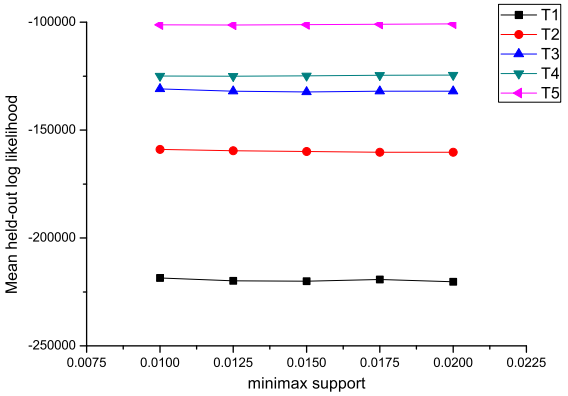

Fig. 3. The Results of Constrained-hLDA

6 Conclusions

This paper improves the popular topic modeling method hLDA by considering existing knowledge in the form of path sampling constraints and level sampling constraints. In the experiment, the proposed constrained-hLDA outperforms hLDA by a large margin, showing that constraints as prior knowledge can help unsupervised topic modeling. Moreover, this paper also proposes the extraction method for two types of constraints automatically. Experimental results show that their qualities are relatively higher than that of unsupervised one.

Acknowledgments. This work is supported by National Natural Science Foundation of China (Grant No: 61175110) and National Basic Research Program of China (973 Program, Grant No: 2012CB316305).

References

1. Java, A., Song, X., Finin, T., Tseng, B.: Why we twitter: understanding microblogging usage and communities. In: Proceedings of the 9th WebKDD and 1st SNA-KDD 2007 Workshop on Web Mining and Social Network Analysis, pp. 56–65. ACM (2007)
2. Krishnamurthy, B., Gill, P., Arlitt, M.: A few chirps about twitter. In: Proceedings of the First Workshop on Online Social Networks, pp. 19–24. ACM (2008)
3. Ramage, D., Dumais, S., Liebling, D.: Characterizing microblogs with topic models. In: International AAAI Conference on Weblogs and Social Media, vol. 5, pp. 130–137 (2010)
4. Zhang, C., Sun, J.: Large scale microblog mining using distributed mb-lda. In: Proceedings of the 21st International Conference Companion on World Wide Web, pp. 1035–1042. ACM (2012)
5. Blei, D.M., Griffiths, T.L., Jordan, M.I., Tenenbaum, J.B.: Hierarchical topic models and the nested chinese restaurant process. In: NIPS (2003)
6. Blei, D.M., Griffiths, T.L., Jordan, M.I.: The nested chinese restaurant process and bayesian nonparametric inference of topic hierarchies. Journal of the ACM 57(2), 7 (2010)

7. Petinot, Y., McKeown, K., Thadani, K.: A hierarchical model of web summaries. In: Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, pp. 670–675 (2011)
8. Mao, X.L., Ming, Z.Y., Chua, T.S., Li, S., Yan, H., Li, X.: Sshlda: a semi-supervised hierarchical topic model. In: Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, pp. 800–809. Association for Computational Linguistics (2012)
9. Mao, X.L., He, J., Yan, H., Li, X.: Hierarchical topic integration through semi-supervised hierarchical topic modeling. In: Proceedings of the 21st ACM International Conference on Information and Knowledge Management, pp. 1612–1616. ACM (2012)
10. Hofmann, T.: Probabilistic latent semantic analysis. In: Proceedings of the Fifteenth Conference on Uncertainty in Artificial Intelligence, pp. 289–296. Morgan Kaufmann Publishers Inc. (1999)
11. Blei, D.M., Ng, A.Y., Jordan, M.I.: Latent dirichlet allocation. The Journal of Machine Learning Research 3, 993–1022 (2003)
12. Chemudugunta, C., Steyvers, P.S.M.: Modeling general and specific aspects of documents with a probabilistic topic model. In: Advances in Neural Information Processing Systems 19: Proceedings of the 2006 Conference, vol. 19, p. 241. The MIT Press (2007)
13. Griffiths, T.L., Steyvers, M.: Finding scientific topics. Proceedings of the National Academy of Sciences of the United States of America 101(suppl. 1), 5228–5235 (2004)
14. Boyd-Graber, J., Blei, D., Zhu, X.: A topic model for word sense disambiguation. In: Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL), pp. 1024–1033 (2007)
15. Rosen-Zvi, M., Griffiths, T., Steyvers, M., Smyth, P.: The author-topic model for authors and documents. In: Proceedings of the 20th Conference on Uncertainty in Artificial Intelligence, pp. 487–494. AUAI Press (2004)
16. Mei, Q., Ling, X., Wondra, M., Su, H., Zhai, C.: Topic sentiment mixture: modeling facets and opinions in weblogs. In: Proceedings of the 16th International Conference on World Wide Web, pp. 171–180. ACM (2007)
17. Teh, Y.W., Jordan, M.I., Beal, M.J., Blei, D.M.: Hierarchical dirichlet processes. Journal of the American Statistical Association 101(476) (2006)
18. Li, W., McCallum, A.: Pachinko allocation: Dag-structured mixture models of topic correlations. In: Proceedings of the 23rd International Conference on Machine Learning, pp. 577–584. ACM (2006)
19. Mimno, D., Li, W., McCallum, A.: Mixtures of hierarchical topics with pachinko allocation. In: Proceedings of the 24th International Conference on Machine Learning, pp. 633–640. ACM (2007)
20. Andrzejewski, D., Zhu, X.: Latent dirichlet allocation with topic-in-set knowledge. In: Proceedings of the NAACL HLT 2009 Workshop on Semi-Supervised Learning for Natural Language Processing, pp. 43–48. Association for Computational Linguistics (2009)