Hierarchical Topic Integration Through Semi-supervised Hierarchical Topic Modeling

Xian-Ling Mao[♠], Jing He[♡], Hongfei Yan[♠†], Xiaoming Li[♠]
[♠]Department of Computer Science and Technology, Peking University, China
[♡]Université de Montréal, Cannada
{xianlingmao, yanhf, lxm}@pku.edu.cn, hejing@iro.umontreal.ca

ABSTRACT

Lots of document collections are well organized in hierarchical structure, and such structure can help users browse and understand these collections. Meanwhile, there are a large number of plain document collections loosely organized, and it is difficult for users to understand them effectively. In this paper we study how to automatically integrate latent topics in a plain collection with the topics in a hierarchical structured collection. We propose to use semi-supervised topic modeling to solve the problem in a principled way. The experiments show that the proposed method can generate both meaningful latent topics and expand high quality hierarchical topic structures.

Categories and Subject Descriptors

G.3 [**Probability and Statistics**]: Stochastic processes; I.2.7 [**Artificial Intelligence**]: [Natural Language Processing, Text analysis]

Keywords

Topical Integration, Hierarchical Topic Modeling

1. INTRODUCTION

In the real world, we usually have a number of document collections, some of which have been well organized as hierarchical structure, and some others don't. For example, to provide an online question and answering service, it is very useful to make use of both cQA and online frequently asked question (FAQ) information. Most of the questions in the cQA systems such as Yahoo! Answers are well organized. However, FAQ does not contain the topical summary information. As a naive solution, we can generate some topical or clustering summary from the FAQ data independently. But such summary may be inconsistent to the hierarchical structure representation of cQA data. Particularly, the topics from the FAQ data may be overlapped with or even redundant to the topics from the cQA data.

† Corresponding author.

CIKM'12, October 29–November 2, 2012, Maui, HI, USA.

It is useful to seamlessly integrate the observed topics in the hierarchical document collections with the latent topics in the plain document collections. First, the hierarchical structure information can definitely guide the topic discovery in the plain document collections. Since the hierarchical structure is carefully organized by the people, the observed nodes in the structure can naturally reflect the semantic topics in a collection. Such semantic topics can provide some supervision for topic modeling in the other document collections. Second, the plain document collections can complement the information in the hierarchical structure. For example, assume that we have a hierarchical collection about "Apple Inc." (as of Nov. 1, 2010), when "ipad" was released, new documents about "ipad" appeared. Thus we need to find and integrate the latent topic "ipad" in the new generated documents into the existing hierarchical topic summary about "Apple Inc.".

In this paper, we study the integration problem for the topics in a hierarchical structured collection and the latent topics in a plain document collection. To the best of our knowledge, such an integration problem has not been studied in the existing work. We propose a generative method to solve this integration problem.

2. PRELIMINARY

The Chinese Restaurant Process (CRP) [1] is a distribution over partitions, and it can be described by the following metaphor. Imagine a restaurant with an infinite number of tables, and imagine customers entering the restaurant in sequence. The d^{th} customer sits at a table according to the following distribution,

$$p(c_d = k|c_{1:(d-1)}) \propto \begin{cases} m_k & \text{if } k \text{ is previous occupied} \\ \gamma & \text{if } k \text{ is a new tabel,} \end{cases} \quad (1)$$

where m_k is the number of previous customers sitting at table k and γ is a positive scalar. After D customers have sat down, their seating plan describes a partition of D items.

3. HIERARCHICAL TOPICAL INTEGRATION

Our approach primarily consists of two stages: (1) **topic integration with horizontal expansion** and (2) **vertical expansion**, illustrated in Figure 1.

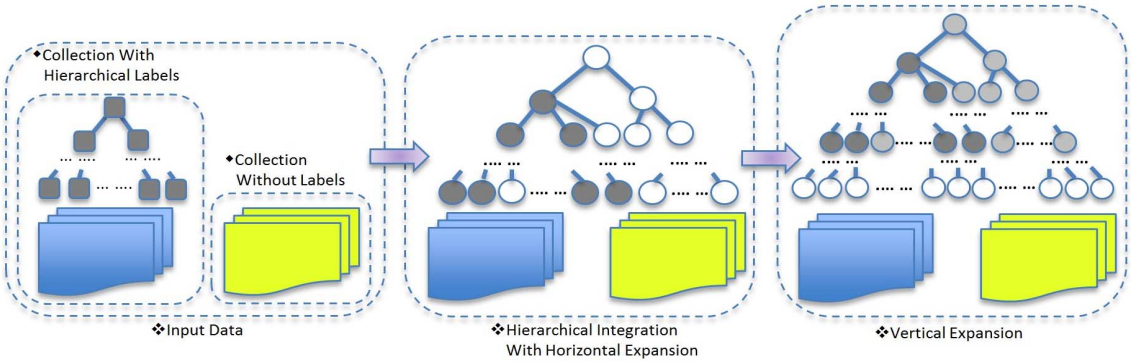

Figure 1: Processing Steps.

3.1 Topic Integration with Horizontal Expansion

3.1.1 Horrizontal expansion topic model

In this section, we will introduce a semi-supervised hierarchical integration topic model, i.e., the *Horizontal Expansion Hierarchical Latent Dirichlet Allocation* (HEHLDA).

We assume that the generative process for a document in a plain collection is as following: 1) For a document from the plain collection, we need to choose a path from a topical tree; 2) Then we need to generate the terms in this document by the topics corresponding to the nodes on its path. After generating M documents, we have chosen a topical tree for these M documents, we can take the M documents and its topical tree as our hierarchical labeled collection, and remaining documents as our plain collection. This metaphor gives us inspiration. To obtain topical integration, we can simulate the process to continue to generate the documents in the plain collection, and finally obtain a hierarchical topical integration automatically. For simplicity, we assume the document length and depth of each leaf node in the tree to be static values, denoted as N and L^l respectively. The derivation is still valid if the document lengths and depth of leaf nodes vary. Thus, the generative process can be formulated as follows:

(1) For each table $k \in T$ in the infinite tree, including labeled table,
   (a) Assume that there have been tables corresponding to labels, and each table has had custermers which are corresponding documents on this label, m_{c_i} is the number of documents assigned to the table c_i. Let c_1 be the root node.
   (b) Draw a topic $\beta_k \sim Dir(\eta)$.

(2) For each document, $m \in \{1, 2, ..., D\}$
   (a) If d is a labeled document, its topical path is the corresponding path in the hierarchy of labels;
   (b) Else if m is a plain document, for each level $l \in \{2, ..., L^l\}$:
      (i) Draw a table from restaurant c_{l-1} using Formula (1). Set c_l to be the restaurant referred to by that table.
   (c) Draw an L-dimensional topic proportion vector from $Dir(\alpha)$.
   (d) For each word $n \in \{1, ..., N\}$:
      (i) Draw $z \in \{1, ..., L^l\}$ from $Mult(\theta)$.
      (ii) Draw w_n from the topic associated with restaurant c_z.

3.1.2 Approximate inference by Gibbs sampling

According to the assumption and generative process above, in this section, we describe a Gibbs sampling algorithm for sampling from the posterior and corresponding latent topics in the HEHLDA model. The Gibbs sampler provides a method for simultaneously exploring the parameter space (the latent topics of the plain collection) and the model space (L-level trees).

In HEHLDA, we sample the per-document paths c_m in plain collection and the per-word level allocations to topics in all paths of the whole topical tree $z_{m,n}$. Thus, we approximate the posterior $p(\boldsymbol{c}_m^p, \boldsymbol{z}_m|\gamma, \eta, \boldsymbol{w}, \boldsymbol{c}^l)$. The hyperparameter γ reflects the tendency of the customers in each restaurant to share tables, η denotes the expected variance of the underlying topics (e.g, $\eta \ll 1$ will tend to choose topics with fewer high-probability words), and $w_{m,n}$ denotes the n^{th} word in the m^{th} document. $c_{m,l}$ represents the restaurant corresponding to the l^{th} topic in document m, and superscript "p" and "l" in a notation denote "from plain collection" and "from hierarchical labeled collection", such as $\boldsymbol{c}_m^p$ means the topical path for the m^{th} document in plain collection; if there is no superscipt in a notation, the document is from the whole collection; and $z_{m,n}$, the assignment of the n^{th} word in the m^{th} document to one of the L available topics. All other variables in the model – θ and β – are integrated out. The Gibbs sampler thus assesses the values of $z_{m,n}$ and $c_{m,l}$.

The Gibbs sampler can be divided into two main steps: the sampling of level allocations and the sampling of path assignments for plain documents.

First, given the values of the HEHLDA hidden variables, we sample the $c_{m,l}^p$ variables which are associated with the CRP prior. The conditional distribution for $\boldsymbol{c}_m^p$, the L_1 topics associated with document m, is:

$$p(\boldsymbol{c}_m^p|\boldsymbol{z}, \boldsymbol{w}, \boldsymbol{c}_{-m}^p, \boldsymbol{c}^l) \propto p(\boldsymbol{w}_m^p|\boldsymbol{z}, \boldsymbol{w}_{-m}, \boldsymbol{c}^p, \boldsymbol{c}^l) p(\boldsymbol{c}_m^p|\boldsymbol{c}_{-m}^p, \boldsymbol{c}^l) \tag{2}$$

where

$$p(\boldsymbol{w}_m^p|\boldsymbol{z}, \boldsymbol{w}_{-m}, \boldsymbol{c}^p, \boldsymbol{c}^l) = \prod_{l=1}^{|\boldsymbol{c}_m|} \left(\frac{\Gamma(n_{c_{m,l},-m}^{\cdot} + |V|\eta)}{\prod_w \Gamma(n_{c_{m,l},-m}^w + \eta)} \times \frac{\prod_w \Gamma(n_{c_{m,l},-m}^w + n_{c_{m,l},m}^w + \eta)}{\Gamma(n_{c_{m,l},-m}^{\cdot} + n_{c_{m,l},m}^{\cdot} + |V|\eta)} \right) \tag{3}$$

$|V|$ is the size of vocabulary, $\Gamma(.)$ denotes the standard gamma function, $n_{c_{m,l},-m}^w$ is the number of instances of word w that

Table 1: The statistics of the datasets.

Datasets	#labels	#paths	Max level	#docs
Y!A	46	35	4	6,345,786
ODP	9126	8869	10	79,193

have been assigned to the topic indexed by $c_{m,l}$, not including those in the document m. $p(\boldsymbol{c}_m^p|\boldsymbol{c}_{-m}^p, \boldsymbol{c}^l)$ as the prior on $\boldsymbol{c}_m^p$ implied by the nested CRP.

Second, given the current state of the HEHLDA, we sample the $z_{m,n}$ variables of the underlying HEHLDA model as follows:

$$p(z_{m,n} = j|\boldsymbol{z}_{-(m,n)}, \boldsymbol{w}, \boldsymbol{c}_m, \boldsymbol{\alpha}_l) \propto \frac{n_{-n,j}^m + \alpha_j}{n_{-n,\cdot}^m + |\boldsymbol{c}_m|} \cdot \frac{n_{-n,j}^{w_{m,n}} + \eta_{w_{m,n}}}{n_{-(m,n)}^{\cdot} + |V|} \quad (4)$$

3.2 Vertical Expansion

Note that HEHLDA builds a topical hierarchical structure, in which each node corresponds to a topic, and each document is assigned to one path in the structure. We illustrate it by the center part in Figure 1. Here, we take all these topics as observed topics, compared with the new topics we will detect. We want to expand our hierarchical topical summarization vertically, illustrated in the right part of Figure 1. We will use *Semi-Supervised Hierarchical Latent Dirichlet Allocation* (SSHLDA) [13] to solve the problem of vertical topical expansion. Like hierarchical Labeled LDA (hLLDA) [16], SSHLDA can incorporate observed topics into the generative process of documents. On the other hand, like hierarchical Latent Dirichlet Allocation (hLDA) [1], SSHLDA can automatically explore latent topic in data space, and extend the existing hierarchy of observed topics.

4. EXPERIMENTS

We demonstrate the effectiveness of the proposed method on large, real-world datasets on three tasks: topic modeling, document classification and clustering.

4.1 Datasets

We first crawled question-answer pairs (QA pairs) from two top categories in Yahoo! Answers: *Computers & Internet* and *Health*. This gives rise to an archive of 6,345,786 QA documents. We refer to the dataset as ***Y!A***. Our ***ODP*** dataset contains the Web pages and their hierarchical structure in two top category (*Home* and *Health*) from Open Directory Project Website *. We removed all categories whose number of Web sites is less than 3 for its sparseness. For each of Web sites in categories, we further extended its description information by submitting the URL to Google and used the words in the snippet and title of the first returned result. The statistics of all datasets are summarized in Table 1. From this table, we can see that these two datasets are very different: Y!A dataset has much fewer categories than ODP dataset, and the depth of the hierarchical structure is much shallow, but it contains much more documents for each category.

In our experiment, we randomly partition a dataset with L-height hierarchical structure into two parts. The random partition processes as follows: (i) first choose top-l levels sub-tree ($l < L$) in the hierarchy, denoted as T^l; (ii) sample a subset of paths S_p from the set of paths in T^l according to a proportion ζ_1. (iii) and then sample some documents from the documents in these paths according to a proportion ζ_2. The sampled documents and their corresponding paths will be taken as our hierarchical labeled collection, denoted as C^l; and the remaining documents C^p, are taken as the plain collection.

*http://dmoz.org/

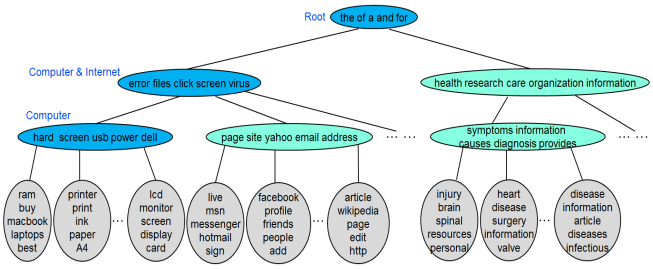

Figure 2: A sub network discovered on Yahoo! Answer dataset using proposed method, and the whole tree has 74 nodes. In the figure, the blue nodes are observed topics with observed labels; The green nodes are the latent topics without labels from plain collection, which is result of HEHLDA; And the shaded nodes are the latent topics without labels from all the documents, which is the result of SSHLDA. Each topic represented by top 5 terms.

In this paper, for both datasets *Y!A* and *ODP*, we choose $\zeta_1 \in \{1/3, 1/2, 2/3\}$, and $\zeta_2 = 3/4$; meanwhile, we choose $l = 3$ for dataset *Y!A* and $l = 4$ for dataset *ODP*. All experimental results are average values, i.e we sample S_p 5 times according ζ_1, then obtain a experimental value for each samping collections, and finally compute average value over these 5 results. In particular, we run HEHLDA and SSHLDA models both for each sampling collection, with a burn-in of 10000 Gibbs sampling iterations, symmetric priors $\alpha = 0.1$ and free parameter $\eta = 1.0$; Then we compute scores by the measures introduced in latter sections, and obtain average values. For $\boldsymbol{\mu}$, we can obtain the estimation of μ_{c_i} by fixed-point iteration [15].

4.2 Case Study

In our models, each latent topic is modeled as a word distribution. Naturally, the words with the highest probability in a word distribution about a topic can be used as a description for this topic [4, 3]. We show an example in Figure 2, which is the topical integration resultant over *Y!A* dataset. The Hierarchical structure is generated by our integration and expansion algorithm. In Figure 2, the blue nodes are those in the existing hierarchical structure; The green nodes are the latent topics generated in topic integration and horizontal expansion (HEHLDA model); And the shaded nodes are the latent topics generated in the vertical expansion (SSHLDA model).

We have three findings from the example: (i) During horizontal topic integration, the new document from the plain collection can be assigned to either a node in the existing hierarchical structure (green nodes in Figure 2), or a node in the expanded latent topics (blue nodes). On the other hand, in the vertical expansion, we can discover finer-grained structure and latent topics (shaded nodes); (ii) During the horizontal integration and vertical expansion, our models can make use of the information from existing hierarchical structure, and it can help generate a logical, structural hi-

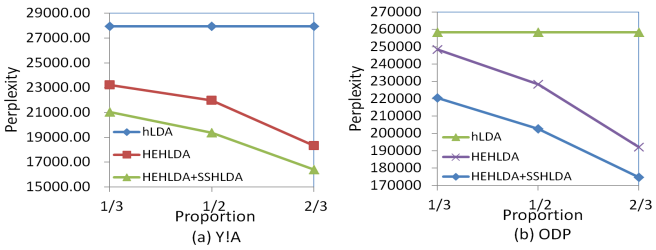

Figure 3: Perplexities of hLDA, HEHLDA and "HEHLDA + SSHLDA". The X-axis is the proportion of documents in BT tree, and Y-axis is the perplexity. (a) The results are run over the *Y!A* dataset, with observed height $l = 3$ and topical height $L = 4$; (b) The results are run over the *ODP* dataset, with observed height $l = 4$ and topical height $L = 8$.

erarchy with parent-child relations Particularly, the latent topic about a parent node is usually general than a topic about a child node. For example, in Figure 2, we can find that the "Computer" node (with topical description "hard, screen" etc.) at the 3rd level contains a lot of children, three of which are about the topic about laptop, printer and monitor respectively. (iii) In a hierarchy of topics, if a topical node is derived directly from the existing hierarchical structure, it usually contains other description information besides the high probability topical words. Many hierarchy organizers label these nodes manually, and these labels can help people understand the themes about the nodes and their descendant in the hierarchy. For example, when we know node "error files click screen virus" in Figure 2 has its label "Computers & Internet", we can understand the child topic "site yahoo email page address" is about "Internet".

These observations show that topical integration is very interesting and useful.

4.3 Generalization Power of Hierarchical Topic Models

Topic models are usually evaluated by its ability to generate the unseen data. *Perplexity* is widely used in the language modeling and topic modeling community [4]. Lower perplexity score of a topic model indicates stronger generalization ability to the new data. In our experiment, we keep 80% of the data collection as the training set and use the remaining collection as the test set.

Since the hierarchical topic integration is a novel problem, we don't have direct baseline models. Instead, we use hLDA model [1] as the baseline, since hLDA can also obtain a hierarchical topics for a plain collection. For comparison, we remove the structure information from the hierarchical structured document collection, and combine it with the plain document collection, and then model this combined plain collection by hLDA model. This method provide us a reasonable baseline.

We present the results on the *Y!A* and *ODP* datasets in Figure 3. As our proposed method has two stages, and it would generate a model as a result of each stage. In the figure, HEHLDA indicated the result from the HEHLDA model and HEHLDA+SSHLDA indicates the result from both HEHLDA and SSHLDA model. From the figure, we can see that the perplexities of proposed models, HEHLDA and the combination of HEHLDA and SSHLDA, are lower than that of hLDA at different proportion value of observed paths. It shows that the proposed models have stronger generalization power than the hLDA model. We can test on different tree height parameter values, and the results are robust to the height selection.

4.4 Evaluating the Hierarchical Structures

The evaluation by perplexity just tells one side of the story: the accuracy of the topics generated by the model. Another important motivation of the integration problem is to expand a semantically meaningful hierarchical structure that can help users browse and understand a document collection. In this section, we investigate the accuracy of the expanded structure from our proposed model.

In our experimental dataset described in section 4.1, we partition the original hierarchical structured collection C (with a gold standard tree T) into two parts: hierarchical structured sub-collection C_1 (with the hierarchical tree T_1) and plain collection C_2. Our models can integrate the documents from C_1 and C_2 into a unified hierarchical structure $T^{'}$. To evaluate the accuracy of the integration, we can compare the rebuilt tree $T^{'}$ with the gold standard tree T. In our models, the documents from the hierarchical structured collection (C_1) are assigned to their original paths, so we need to compare the paths of those documents from the plain collection (C_2).

We use classification and clustering metrics to evaluate two kinds of documents in C_2: (i) $C_{2,1}$: the documents that are selected from the paths in T_1, and (ii) $C_{2,2}$: those documents that are selected from the paths that do not exist in T_1.

4.4.1 Evaluating hierarchical structure as classification

For a document from $C_{2,1}$, we can check whether the its assigned path is exactly the same as its path in gold standard tree T. So we can cast it as a classification problem, and evaluate it with a widely used classification metric *micro-averaged* F_1 (*Mi-*F_1) [20]. We choose *RBF-SVM* as our baseline. The label of a node will be taken as the target value of all documents in this node, and words are used as features.

We show the result of Mi-F_i in Table 2. Whatever the sampling proportion is, from Table 2, proposed method performances significantly better than that of the baseline on both datasets. Why our method performances better? One possible explanation is that baseline method does not consider the relation among nodes, and our method uses indirectly the hierarchical relation among nodes. On the other hand, with the sampling proportion increases, the Mi-F_1 decreases. This is because the number of categories increases when sampling proportion increases, which will generally decrease the performance of classification.

4.4.2 Evaluating hierarchical structure as clustering

Besides the documents that are assigned to $T^{'}$ in the gold standard, there are still some documents ($C_{2,2}$) are not assigned to any paths in $T^{'}$, since $T^{'}$ is only a sub-tree of T. Since the expanded nodes in tree $T^{'}$ do not have identities, we cannot compare these new paths to those in T. Alternatively, we use the clustering evaluation technique to measure the quality of these expanded structures.

In this paper, we use the FScore [12] to measure the qual-

Table 2: The Mi-F_1 score over two datasets.

Datasets	Methods	Proportion (ζ_1)		
		1/3	1/2	2/3
Y!A ($l = 3$)	proposed	0.8598	0.8236	0.7543
	SVM	0.8193	0.7862	0.7273
ODP ($l = 4$)	proposed	0.6725	0.5383	0.4466
	SVM	0.6132	0.4915	0.4034

ity of the expanded structure. We take those nodes in T but not in T_1 as gold standard expanded paths, and each node is corresponding to a class. Similarly, each node in $T^{'}$ but not in T_1 can be considered as a cluster generated by our models. In general, the higher the FScore values, the better the clustering solution is.

In this aspect, we demonstrate the strength of the hierarchical integration models by comparing it with a single-linkage clustering algorithm CLUTO [9]. It can generate the clusters for documents in $C_{2,2}$. We denote the method as *h-cluster*.

For *Y!A*, $l = 3$ and $L \in \{3, 4\}$; for *ODP*, $l = 4$ and $L \in \{4, 5, 6\}$. The clustering results are presented in Table 4. From the table, we can see that our proposed model can achieve consistent improvement over the baseline at smaller depth for *Y!A* and *ODP*. However, in higher height, *h-cluster* maybe performance better than our proposed method. This is reasonable because higher height means more clusters to form for proposed method, which will decrease the performance of clustering. For HEHLDA, it always performances better than h-cluster over *Y!A* and *ODP*. For example, for *Y!A*, the performance of HEHLDA can reach 0.5183, 0.5939 and 0.6987 with proportion $\zeta_1 \in \{1/3, 1/2, 2/3\}$ while the h-cluster only achieve 0.40845, 0.5362 and 0.5826. The result shows that our model can achieve about 26.9%, 10.8% and 19.9% improvements over h-clustering with proportion $\zeta_1 \in \{1/3, 1/2, 2/3\}$. The improvements are significant by t-test according to significance 95%. We can also obtain similar conclusion over *ODP*.

5. RELATED WORKS

To the best of our knowledge, no previous study has addressed the problem of integrating a hierarchy of label topics with latent topics in text documents. Topic model has been widely and successfully applied to mine topic patterns [3, 4]. Our work adds to this line yet another novel use of such models for topical integration.

Unsupervised non-hierarchical topic models are widely studied, such as pLSA [8], LDA [4] and Concept TM [5, 7, 6] etc. However, the above models cannot capture the relation between super and sub topics. To address this problem, many models have been proposed to model the relations, such as Hierarchical LDA (HLDA) [1], Hierarchical Dirichlet processes (HDP) [21], Pachinko Allocation Model (PAM) [11] and Hierarchical PAM (HPAM) [14] etc.

Although unsupervised topic models are sufficiently expressive to model multiple topics per document, they are inappropriate for labeled corpora because they are unable to incorporate the supervised label set into their learning procedure. Several modifications [20, 2, 3, 10, 17, 19, 20, 20, 18] of LDA to incorporate supervision have been proposed in the literature.

None of these models, however, leverage dependency structure, such as parent-child relation, in the label space. As far as we know, only hLLDA [16] is proposed to capture the structral relation.

Table 3: The FScore over two datasets.

Datasets	Methods	Height (L)	Proportion (ζ_1)		
			1/3	1/2	2/3
Y!A ($l = 3$)	Proposed	3(HEHLDA)	0.5183	0.5939	0.6987
		4	0.4221	0.4976	0.6195
	h-cluster	-	0.4084	0.5362	0.5826
ODP ($l = 4$)	Proposed	4(HEHLDA)	0.3684	0.3847	0.4566
		5	0.3284	0.3447	0.3966
		6	0.2927	0.3392	0.3663
		7	0.2652	0.3113	0.3439
		8	0.2214	0.2852	0.3072
	h-cluster	-	0.2688	0.3037	0.3594

6. CONCLUSION

In this paper, we try to use semi-supervised probabilistic topic modeling to solve a novel problem of hierarchical topical integration which aims at integrating topics expressed in a well-written labels with latent topics hidden in plain collections to generate a unified hierarchical topical summary.

Acknowledgments

This work was supported by NSFC with Grant No. 61073082, 60933004, 70903008.

7. REFERENCES

[1] D. Blei, T. Griffiths, M. Jordan, and J. Tenenbaum. Hierarchical topic models and the nested chinese restaurant process. *Advances in neural information processing systems*, 16:106, 2004.

[2] D. Blei and J. McAuliffe. Supervised topic models. In *Proceeding of the Neural Information Processing Systems(nips)*, 2007.

[3] D. Blei and J. McAuliffe. Supervised topic models. *Arxiv preprint arXiv:1003.0783*, 2010.

[4] D. Blei, A. Ng, and M. Jordan. Latent dirichlet allocation. *The Journal of Machine Learning Research*, 3:993–1022, 2003.

[5] C. Chemudugunta, A. Holloway, P. Smyth, and M. Steyvers. Modeling documents by combining semantic concepts with unsupervised statistical learning. *The Semantic Web-ISWC 2008*, pages 229–244, 2008.

[6] C. Chemudugunta, P. Smyth, and M. Steyvers. Combining concept hierarchies and statistical topic models. In *Proceeding of the 17th ACM conference on Information and knowledge management*, pages 1469–1470. ACM, 2008.

[7] C. Chemudugunta, P. Smyth, and M. Steyvers. Text modeling using unsupervised topic models and concept hierarchies. *Arxiv preprint arXiv:0808.0973*, 2008.

[8] T. Hofmann. Probabilistic latent semantic analysis. In *Proc. of Uncertainty in Artificial Intelligence, UAI'99*, page 21. Citeseer, 1999.

[9] G. Karypis. Cluto: Software for clustering high dimensional datasets. *Internet Website (last accessed, June 2008), http://glaros.dtc.umn.edu/gkhome/cluto/cluto/overview*, 2005.

[10] S. Lacoste-Julien, F. Sha, and M. Jordan. ndisclda: Discriminative learning for dimensionality reduction and classification. *Advances in Neural Information Processing Systems*, 21, 2008.

[11] W. Li and A. McCallum. Pachinko allocation: Dag-structured mixture models of topic correlations. In *Proceedings of the 23rd international conference on Machine learning*, pages 577–584. ACM, 2006.

[12] C. Manning, P. Raghavan, and H. Schutze. *Introduction to information retrieval*, volume 1. Cambridge University Press Cambridge, 2008.

[13] X. Mao, Z. Ming, T. Chua, S. Li, H. Yan, and X. Li. Sshlda: A semi-supervised hierarchical topic model. *Conference on Empirical Methods on Natural Language Processing*, 2012.

[14] D. Mimno, W. Li, and A. McCallum. Mixtures of hierarchical topics with pachinko allocation. In *Proceedings of the 24th international conference on Machine learning*, pages 633–640. ACM, 2007.

[15] T. Minka. Estimating a dirichlet distribution. *Annals of Physics*, 2000(8):1–13, 2003.

[16] Y. Petinot, K. McKeown, and K. Thadani. A hierarchical model of web summaries. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: short papers-Volume 2*, pages 670–675. Association for Computational Linguistics, 2011.

[17] D. Ramage, P. Heymann, C. Manning, and H. Garcia-Molina. Clustering the tagged web. In *Proceedings of the Second ACM International Conference on Web Search and Data Mining*, pages 54–63. ACM, 2009.

[18] D. Ramage, C. Manning, and S. Dumais. Partially labeled topic models for interpretable text mining. In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 457–465. ACM, 2011.

[19] M. Rosen-Zvi, T. Griffiths, M. Steyvers, and P. Smyth. The author-topic model for authors and documents. In *Proceedings of the 20th conference on Uncertainty in artificial intelligence*, pages 487–494. AUAI Press, 2004.

[20] T. Rubin, A. Chambers, P. Smyth, and M. Steyvers. Statistical topic models for multi-label document classification. *Arxiv preprint arXiv:1107.2462*, 2011.

[21] Y. Teh, M. Jordan, M. Beal, and D. Blei. Hierarchical dirichlet processes. *Journal of the American Statistical Association*, 101(476):1566–1581, 2006.