Md Hijbul Alam, Jaakko Peltonen, Jyrki Nummenmaa, and Kalervo Järvelin. Tree-structured Hierarchical Dirichlet Process. In S. Rodriguez, J. Prieto, P. Faria, S. Klos, A. Fernandez, S. Mazuelas, M. D. Jimenez-Lopez, M. N. Moreno, and E. M. Navarro Martinez, editors, Distributed Computing and Artificial Intelligence, Special Sessions, 15th International Conference (Proceedings of DCAI 2018), pages 291-299, Springer, 2019.

The final publication is available at Springer via https://doi.org/10.1007/978-3-319-99608-0_33

Tree-structured Hierarchical Dirichlet Process

Md Hijbul Alam[1⋆], Jaakko Peltonen[1,2⋆]Jyrki Nummenmaa[1], and Kalervo Järvelin[1]

[1] University of Tampere, Finland
[2] Aalto University, Finland
{hijbul.alam, jaakko.peltonen, jyrki.nummenmaa, kalervo.jarvelin}@uta.fi

Abstract. In many domains, document sets are hierarchically organized such as message forums having multiple levels of sections. Analysis of latent topics within such content is crucial for tasks like trend and user interest analysis. Nonparametric topic models are a powerful approach, but traditional Hierarchical Dirichlet Processes (HDPs) are unable to fully take into account topic sharing across deep hierarchical structure. We propose the Tree-structured Hierarchical Dirichlet Process, allowing Dirichlet process based topic modeling over a given tree structure of arbitrary size and height, where documents can arise at all tree nodes. Experiments on a hierarchical social message forum and a product reviews forum demonstrate better generalization performance than traditional HDPs in terms of ability to model new data and classify documents to sections.

Keywords: Hierarchical Dirichlet Processes, Topic Modeling, Message Forum

1 Introduction

Modeling online discussions is important for studies of discussion behavior, for tracking trends of ideas and consumer interests, for recommendation of discussion content or targeted advertising, and for intelligent interfaces to browse discussions. Online discussion often occurs in venues having a prominent hierarchical organization such as hierarchical forums (message boards). General-interest forums cover a broad range of interests such as politics, health, product reviews, and so on. As a case study we use a popular Finnish forum Suomi24 (www.suomi24.fi) spanning 16 years and 6.5 million threads. Forums are organized into hierarchical sections created by administrators for prototypical interests. Hierarchical organization also occurs in online reviews at, e.g., websites such as Amazon.com, where reviews follow the hierarchy of the products.

Administrator-created sections are simplified divisions that do not suffice to describe the variety of semantic content in discussions; an important task in data analytics of online forums is to extract latent topics of discussion. Modeling text data is often done by generative topic models such as Latent Dirichlet Allocation [1] and Dirichlet Processes [2], which represent unstructured text as a bag of words arising out of a mixture of latent topics. In this paper we give a solution for the challenge of effectively taking hierarchical structure of data collections into account in such modeling. User in-

⋆ MHA and JP had equal contributions. The work was supported by Academy of Finland decisions 295694 and 313748.

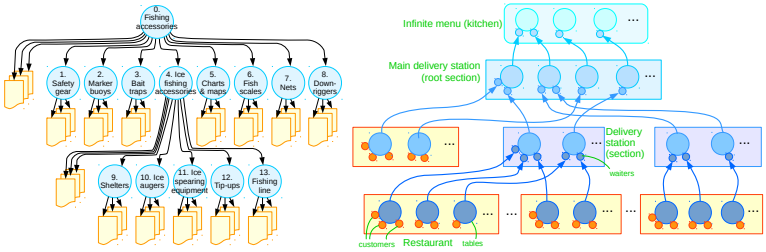

Fig. 1: (Left) Hierarchical document organization in part of the Amazon product hierarchy. Yellow icons denote documents which here are reviews (threads); they can appear under any section (blue circles) at any hierarchy level. (Right) An illustration of an Imperial Chinese Banquet.

terests need not match the administrator-created structure. Issues touching on multiple interests (say food and health) may have no dedicated section, and users may discuss them in multiple sections. Users may digress from the section theme; threads with many users follow a mixture of their interests. Thus, forum sections need not correspond to section themes.

Recent work on text mining has attempted hierarchical text analysis: most works [3–5] build an unsupervised hierarchy of topics from a document set; they ignore predefined organization of documents in a section hierarchy and cannot extract the topic distributions of a section in the hierarchy. HDP and its variations aim to take data division into account [2] but cannot readily be extended where sets of documents can arise at any node in the tree, as in Fig. 1 (Left).

We introduce the Tree-structured Hierarchical Dirichlet Process (THDP), a new model which identifies latent topics of each section in a hierarchy. THDP is a generative model for the documents in any position of a hierarchy and can be applied to data in hierarchies of arbitrary size and height. Our contributions: **1.** We develop a new nonparametric hierarchical topic model to model forum texts which can come from any place of the section hierarchy. The key is a new nonparametric generative process, the Imperial Chinese Banquet, representing a top-down percolation of topics to documents at different hierarchy levels. **2.** We develop a Gibbs sampling algorithm that extracts topics and their usage across threads and hierarchical sections. **3.** In experiments, evaluated with various metrics and use cases, our model outperforms the state-of-art models.

2 Related Work

A topic model [1] is a parametric Bayesian model for count data such as bag-of-words representations of text documents. Teh et al. [2] propose HDP (Fig. 2, Left), a non parametric model where the number of topics does not need to be pre-specified. The crucial difference to our work is that HDP by Teh et al. is not designed for deep hierarchies; as presented in Teh et al., their model was mainly used for a "flat" division of documents into groups: Dirichlet processes (DPs) of each document were only connected by one DP for each group, under an overall DP. In such a flat model, documents always occur at the groups and the parent level is unobserved; in our model, documents can occur

under any node in the deep hierarchy. Alternative models exist e.g. placing additional sparsity priors for topic sharing [6] but again not for deep hierarchies. Some variants involve a hierarchy: in the nested Chinese restaurant process [3] and knowledge-based hierarchical topic model [7], a document is modeled as a distribution over a path from the root to the leaf node; in the recursive Chinese restaurant process [8], a document has a distribution over all of the nodes of the hierarchy; in the tree-structured stick-breaking process [5], a document is modeled by a node of the tree. In these models, a tree structure is learned to represents topics; whereas in THDP we do not need to learn the structure as our model is based on a known hierarchy; we focus on modeling using the given deep hierarchy as the model structure.

3 Tree-structured Hierarchical Dirichlet Process

We describe a generative process, THDP, given a tree-structured hierarchy of sections, where documents can arise at any section. A global distribution G^0_{root} over topics is first drawn from a Dirichlet process (DP) with base distribution H and concentration parameter α^0 for the root node of a given tree, denoted $G^0_{root} \sim DP(\alpha^0, H)$. The root node corresponds to the root section. We index nodes as v. For each child section v of the root, a discrete distribution G^1_v is drawn from a DP with base distribution G^0_{root} and concentration parameter α^1, denoted $G^1_v \sim DP(\alpha^1, G^0_{root})$. This is repeated recursively for every child node to generate its grandchild sections: a node v at level l in the hierarchy (l steps down from the root) has a discrete distribution G^l_v generated from a DP with base distribution $G^{l-1}_{p(v)}$ and concentration parameter α^l, where $p(v)$ is the parent node of v, denoted $G^l_v \sim DP(\alpha^l, G^{l-1}_{p(v)})$. Lastly, G_j for a document j under a node v at level l is drawn from a DP with base distribution G^l_v and concentration parameter α^{l+1}, denoted $G_j \sim DP(\alpha^{l+1}, G^l_v)$. Document content is then generated: for each word i in a document j, draw the topic $\theta_{ji} \sim G_j$ and draw the observed word from the topic's word distribution as $x_{ji} \sim F(\theta_{ji})$. Fig. 2 (Middle) shows the plate representation graphical model of THDP, with an instantiation for an example hierarchy in Fig. 2 (Right).

We describe a metaphor for THDP which we call the *Imperial Chinese Banquet* (ICB); we will use it for inference. A banquet is arranged in a multilevel palace: each level has several *food-delivery stations*, each serving several *restaurants* (dining rooms) at that level. Attendees (i.e., customers) visit dining rooms (i.e., restaurants) to eat popular dishes: each restaurant has tables for customers, and there is a responsible *waiter* at every table who brings a dish to the table, fetching it from a table in a food-delivery station. At food-delivery stations, the tables also have responsible waiters who bring the dishes from an upper-level delivery station, recursively. Each time a customer/waiter chooses a table, they prefer popular tables that other customers/waiters have also picked.

4 Inference

We introduce a Gibbs sampling scheme for THDP, based on the ICB representation. We sample tables, pointers to ancestor tables, and dishes for tables. Let $f_k^{-x_{ji}}(x_{ji})$ denote the conditional density or likelihood of x_{ji} given all data items except x_{ji}, where k is the

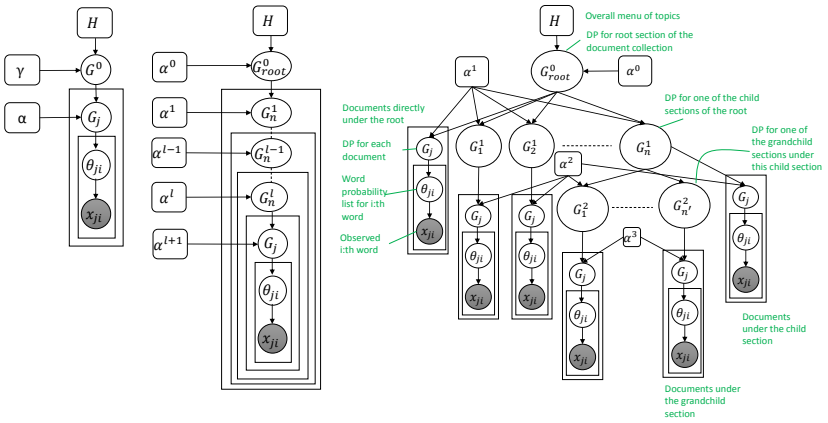

Fig. 2: (Left) Hierarchical Dirichlet Process. (Middle) Tree-structured Hierarchical Dirichlet Process (THDP). (Right) Detailed plate model instantiation example of the THDP, for an example hierarchy having three levels and documents arising at each level.

dish at the table of x_{ji}, and j is a document index. We have $f_k^{-x_{ji}}(x_{ji}) \propto n_{kw}^{-ji}/n_{k.}^{-ji}$ for an existing dish and $f_{k_{new}}^{-x_{ji}}(x_{ji}) \propto 1/V$ for a new dish [2], where w is the word index of x_{ji}, V is the vocabulary size, n_{kw}^{-ji} is the number of occurrences of w from dish k (other than x_{ji}), and $n_{k.}^{-ji}$ is the sum over different word indices. We denote $f_k^{-\boldsymbol{x}_{jt}}(\boldsymbol{x}_{jt}) = (\prod_w(\beta + n_{kw}-1)...(\beta+n_{kw}^{-jt}))/((V\beta+n_{kw}-1)...(V\beta+n_{k.}^{-jt}))$ as the conditional density of $\boldsymbol{x}_{jt}$ given all data items associated with mixture component k leaving out $\boldsymbol{x}_{jt}$, where β is a hyperparameter.

Part 1. Sampling table t for a customer x_{ji} at a restaurant: For an individual customer the likelihood for a new table $t_{ji} = t^{new}$ can be calculated by integrating out the possible values of the new dish $k_{jt^{new}}$: $p(x_{ji}|\boldsymbol{t}_{-ji}, t_{ji} = t^{new}; \boldsymbol{k}) = \sum_{k=1}^{K} \frac{m_{.k}}{m_{..}+\alpha^0} f_{k_{jt}}^{-x_{ji}}(x_{ji}) + \frac{\alpha^0}{m_{..}+\alpha^0} f_{k_{jt}^{new}}^{-x_{ji}}(x_{ji})$. Here the m values are total counts of tables from restaurants at all leaf nodes (observed documents), and α is a hyperparameter. We make a computationally efficient approximation in the right-hand term (corresponding to a new table at the parent node) by evaluating its word probabilities directly from the root instead of recursively traveling up. Therefore, at a restaurant the conditional distribution of t_{ji} is: $p(t_{ji} = t) \propto (n_{jt.}^{-ji}/(n_{.j.} + \alpha^{l+1}))f_{k_{jt}}^{-x_{ji}}(x_{ji})$, and $p(t_{ji} = t^{new}) \propto (\alpha^{l+1}/(n_{j..} + \alpha^{l+1}))p(x_{ji}|\boldsymbol{t}_{-ji}, t_{ji} = t^{new}; \boldsymbol{k})$, where $n_{jt.}$ is the number of customers in restaurant j at table t.

Part 2. Sampling a table t from delivery-station v for a new waiter with first customer x_{ji}: When a customer x_{ji} sits at a new restaurant table, it has no dish yet: the waiter at that table must fetch a dish for this first customer from the delivery station for the restaurant, and must then choose some table t_{jt} from delivery-station v. The delivery-station table can be either a table that other waiters have also picked, or a new delivery-station table; in the latter case a new dish must then be brought from the upper-level delivery station. The likelihood for $t_{jt} = t^{new}$ can be calculated as follows:

Table 1: Data set properties. Section counts at level 2-4 below the root given in parentheses.

	#sections	#Train docs	#Test docs	#terms	Avg. doc len
Suomi24 Politics	49 (16+16+17)	980	245	50217	323.5
Suomi24 Health	15 (5+9+1)	300	75	14700	209.9
Suomi24 Relationship	18 (14+4+0)	360	90	17804	264.1
Amazon Fishing Acc.	13 (0+8+5)	260	65	3206	256.6

$p(t_{jt}|\boldsymbol{t}_{-jt}, t_{jt} = t^{new}; \boldsymbol{k}) = \sum_{k=1}^{K}(c_{vt.}/(c_{v..}+\alpha^l))f_{k_{jt}}^{-x_{ji}}(x_{ji}) + (\alpha^l/(c_{v..}+\alpha^l))f_{k_{jt}^{new}}^{-x_{ji}}(x_{ji})$, where $c_{vt.}$ is the number of tables point to table t in node v and $c_{v..}$ is the number of tables point to tables in node v. Therefore, the conditional distribution of t_{jt} (with a customer at a restaurant) is $p(t_{jt} = t) \propto (c_{vt.}^{-jt}/(c_{v..}+\alpha_j))f_{k_{jt}}^{-x_{ji}}(x_{ji})$ and $p(t_{jt} = t^{new}) \propto (\alpha_j/(c_{v..}+\alpha_j))p(t_{jt}|\boldsymbol{t}_{-jt}, t_{jt} = t^{new}; \boldsymbol{k})$.

Part 3. Sampling a delivery-station table t for a waiter with several existing customers: The likelihood for $t_j = t^{new}$ for many customers in a table can be calculated as: $p(t_{jt}|\boldsymbol{t}_{-jt}, t_{jt} = t^{new}; \boldsymbol{k}) = \sum_{k=1}^{K}(c_{vt.}/(c_{v..}+\alpha^l))f_k^{-\boldsymbol{x}_{jt}}(\boldsymbol{x}_{jt}) + (\alpha^l/(c_{v..}+\alpha^l))f_{k_{new}}^{-\boldsymbol{x}_{jt}}(\boldsymbol{x}_{jt})$. Therefore, the conditional distribution of t_j, given all customers in the table, is $p(t_{jt} = t) \propto (c_{vt.}^{-jt}/(c_{v..}+\alpha^l))f_k^{-\boldsymbol{x}_{jt}}(\boldsymbol{x}_{jt})$ and $p(t_{jt} = t^{new}) \propto (\alpha^l/(c_{v..}+\alpha^l))p(t_{jt}|\boldsymbol{t}_{-jt}, t_{jt} = t^{new}; \boldsymbol{k})$. If the sampled value of t_{ji} is t^{new}, we create a new table at the upper level, and recursively sample its dish. If the upper level is the root level, a topic is sampled $k_{j^{root}t}$ with respect to k_{jt} and propagated to all its descendants.

Part 4. Sampling k: The conditional probability of a dish at the root level $k_{j^{root}t}$ i.e., k_{jt} is: $p(k_{jt} = k) \propto (m_{.k}^{-jt}/(m..+\alpha^0))f_k^{-\boldsymbol{x}_{jt}}(\boldsymbol{x}_{jt})$ and $p(k_{jt} = k_{new}) \propto (\alpha^0/(m..+\alpha^0))f_{k_{new}}^{-\boldsymbol{x}_{jt}}(\boldsymbol{x}_{jt})$.

Part 5. Sampling k for a new table: If a customer is given a new table ($t_{ji} = t^{new}$) we sample $k_{jt^{new}}$ as follows: $p(k_{jt^{new}} = k) \propto (m_{.k}/(m..+\alpha^0))f_k^{-x_{ji}}(x_{ji})$ and $p(k_{jt^{new}} = k_{new}) \propto (\alpha^0/(m..+\alpha^0))f_{k_{new}}^{-x_{ji}}(x_{ji})$.

We summarize the Gibbs sampling algorithm for THDP inference: sample a table assignment for each word in a document with a recursive procedure as follows. For a word, sample a table as in Part 1; if it's a new table, move to the parent node to sample a table from the parent node as in Part 2; repeat until the root node is reached; then select a topic for the table in the root as in Part 5, and update the topic of all tables in the descendant's nodes of the table in the root. Similarly, for each table (i.e., a group of words associated with a table) in a document, we sample a parent table i.e. a table from the parent using as in Part 3. We repeat the process until the root is reached and eventually sample a topic for the root table using as in Part 4.

5 Experimental Results

We first describe the data sets, summarized in Table 1. We begin by qualitative comparisons, and then present quantitative comparisons.

We evaluate the THDP model against the baseline HDP [2] on difficult modeling tasks where relatively little observation data is available, and a well-chosen model struc-

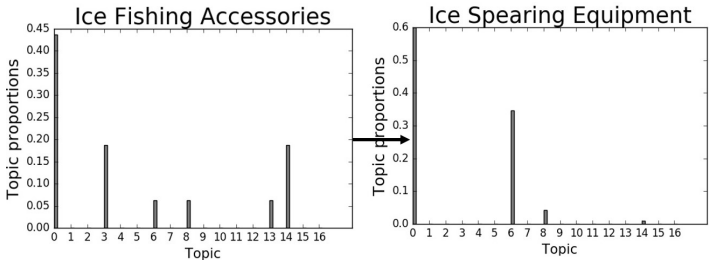

Fig. 3: THDP topic proportions in two example sections of the fishing accessories data set.

Table 2: THDP topics for Amazon data set, sections where they are active, and top words

Topic	Sections	Stemmed top words of the topic
0	all	fish work great good ice easi product bought make time made line order
1	0, 1, 5	oar clam gun collar book chart lock detail oarlock razor map guid
2	0, 5	fli cast video joan dvd learn watch wulff fish great teach instruct
3	0, 4, 9, 12	sled shelter shack chair wind warm set frame plenti heater pak front
4	0, 3, 8	planer board rapala belt fight releas descript pro tension troll brown
5	0, 2	buoy clamp float holder rod anchor outrigg kayak crab sand umbrella
6	0, 4, 11	gun spear band speargun load shoot dive shaft shot jbl cressi spearfish
7	0, 1	glass cabl wear sunglass cablz neck snow pair retain face read goggl
8	0, 4, 11	glove batteri heat provid hand pair warm finger cell pack wear chemic
9	0	spear frog sharpen tine gig hook sharp barb point head bend weld file
10	0, 6	scale weigh accur measur batteri scaler lip gripper digit pound tape
11	0, 3	trap crab bait pot door wire tank caught tie danielson fold blue
12	0, 7	net minnow cast throw return sink hook sein foot tradit styrofoam bow
13	0, 4	tini yard firelin leader bead invis knot sufix reel strong cast crystal
14	0, 4, 8, 10-13	grabber tag grab equip flop chip slipperi slimi soft slip northern
15	0, 1, 6	helmet complet stingray pad roman buyer paint gaiter foam hurt insid
16	0, 1	alarm loud sound night brother wrap backward speaker china lit led

ture can thus help. We used two different data sources, *Suomi24* and Amazon. Suomi24 has in total 2434 sections in its hierarchy. The data set (https://www.kielipankki.fi/corpora/) is publicly available in original and lemmatized forms. From this source, we created several data sets for our experiments. The second data source is *reviews on Amazon.com*, a major shopping site with numerous shopping sections, for example, 1933 sections under Sports and Outdoors department [9]. We select the Fishing Accessories data set which is under Sports and Outdoors $\rightarrow$ Sports $\rightarrow$ Hunting and Fishing $\rightarrow$ Fishing category. The data set contains products at different levels of hierarchy. For each section containing products, at whatever level of hierarchy, we select 20 threads for training and 5 threads for testing. Therefore, the Amazon Fishing Accessories data set contains in total 260 reviews for training and 65 reviews for testing. We lemmatized the words in all reviews. Table 1 shows the numbers of sections and total training and test set sizes.

Qualitative Analysis. We verify that extracted THDP topics in a section are related to the topic of the section. The analysis could be carried out for different alpha values;

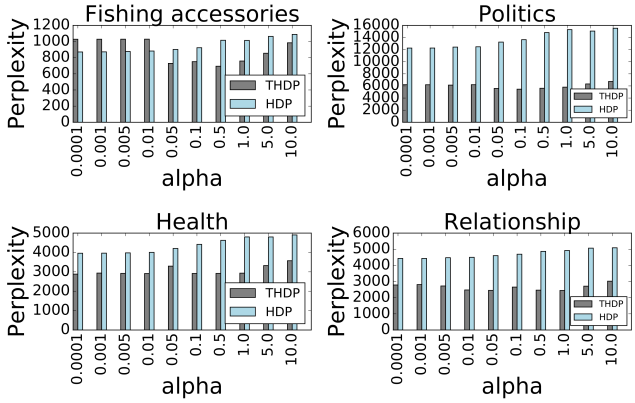

Fig. 4: Perplexity on different test data sets with different alpha values

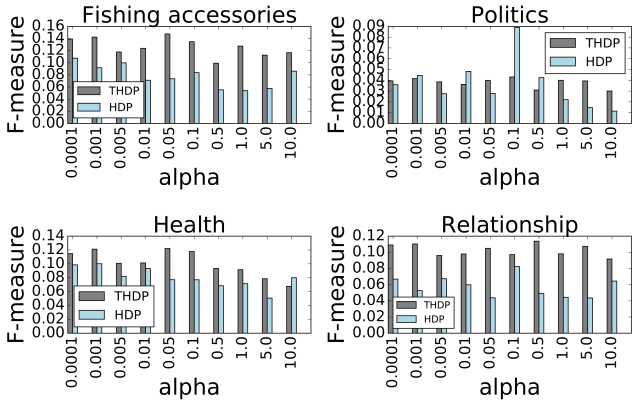

Fig. 5: Section prediction performance in terms of the F-measure of retrieving the correct section in different data sets with different alpha values.

we present results for an example alpha value 1. The top words of THDP topics for the Fishing Accessories data set are shown in Table 2. For many sections, we observe that extracted latent topics correspond to the section themes. For example, top words of topics 6 and 10 are names of Ice Spearing Equipment and verbs for using them (e.g., spear, gun, load, shoot etc.) and details of Fishing Scales equipment (e.g., scale, battery, weight etc.). Similarly, Topics 3, 5, 12, and 13 are about Shelters, Marker Buoys, Nets, and Fishing Line, respectively. THDP topics are also shared across different sections. For example, the THDP Topic 1 discusses both Safety Gear (e.g., oar, oarlock, lock etc.) and Charts & Maps (e.g., chart, maps, book etc.). Topic 4 is another example where the discussion is about both Downriggers and Bait Traps section with fish catching related equipment keywords such as rapala, belt and troll. However, Bait Traps is also specifically discussed in Topic 11 (e.g., trap, bait, wire, tank, crab etc.).

We also analyze topic proportions at sections in the hierarchy. Fig. 3 shows THDP topic proportions for the Ice Fishing Accessories section and one of its child sections Ice Spearing Equipment. In both charts Topic 0 is about fishing or price in general as shown in Table 2; the topic has a large portion in all sections. Ice Spearing Equipment section activates Topics 0, 6, and 8, which are also present in the parent section.

Quantitative Analysis. First, we evaluate the ability of THDP to represent new incoming documents with perplexity of held-out test documents [1]. Fig. 4 shows the results, lower perplexity is better. THDP outperforms HDP in perplexity for most of the data sets for most of the tried alpha values (except for some alpha values in Fishing accessories). Next, for section prediction, we train a HDP model for each section in the dataset. For THDP, we train a single model for each dataset. To predict the section for each test document, we compute perplexity for the test document under the model for each section, and assign the document to the section or sections that yield the lowest perplexity. Fig. 5 shows the resulting F-measures for different α values for different datasets, averaged over 5 runs. THDP outperforms HDP with higher F-measure for most data sets for most alpha values (except for some alpha values in Politics) since it incorporates section hierarchy information in the model.

6 Conclusions

We introduced the Tree-structured Hierarchical Dirichlet Process (THDP), a generative model for documents in deep tree-structured hierarchies such as online discussion forums. THDP extracts latent topics (discussion themes) shared across discussion sections, and outperforms the state-of-the-art model HDP in modeling new documents (measured by perplexity) and section prediction (measured by F-measure). Unlike previous work, THDP can incorporate a truly multilevel hierarchy. It can be adapted to many topic modeling applications to take into account their hierarchical data structure.

References

1. Blei, D., Ng, A., Jordan, M.: Latent Dirichlet allocation. JMLR **3** (2003) 993–1022
2. Teh, Y., Jordan, M., Beal, M., Blei, D.: Hierarchical Dirichlet processes. J Am Stat Assoc **101** (2006) 1566–1581
3. Blei, D., Griffiths, T., Jordan, M.: The nested chinese restaurant process and Bayesian nonparametric inference of topic hierarchies. J ACM **57** (2010) 7:1–7:30
4. Li, W., McCallum, A.: Pachinko allocation: DAG-structured mixture models of topic correlations. In: Proc. ICML, ACM (2006) 577–584
5. Adams, R., Ghahramani, Z., Jordan, M.: Tree-structured stick breaking for hierarchical data. In: Proc. NIPS, Curran Associates Inc. (2010) 19–27
6. Faisal, A., Gillberg, J., Leen, G., Peltonen, J.: Transfer learning using a nonparametric sparse topic model. Neurocomputing **112** (2013) 124–137
7. Xu, Y., Yin, J., Huang, J., Yin, Y.: Hierarchical topic modeling with automatic knowledge mining. Expert Systems with Applications **103** (2018) 106 – 117
8. Kim, J., Kim, D., Kim, S., Oh, A.: Modeling topic hierarchies with the recursive chinese restaurant process. In: Proc. CIKM, ACM (2012) 783–792
9. He, R., McAuley, J.: Ups and downs: Modeling the visual evolution of fashion trends with one-class collaborative filtering. In: Proc. WWW. (2016) 507–517