Author Tree-structured Hierarchical Dirichlet Process

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

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

Abstract. Three key aspects of online discussion venues are the multitude of participants, the underlying trends of content, and the structure of the venue. However, most models are unable to take into account all three of these. In hierarchically organized message forums, authors may participate differently at multiple levels of sections, with different interests and contributions across the hierarchy. Well-designed probabilistic models of online discussion are applicable to many tasks such as prediction of future content or authorship attribution. However, traditional models such as Hierarchical Dirichlet Processes (HDPs) do not fully take into account authors, and are further unable to fully take into account deep hierarchical venues where documents can arise at all tree nodes. We introduce the Author Tree-structured Hierarchical Dirichlet Process (ATHDP), allowing Dirichlet process based topic modeling of both text content and authors over a given tree structure of arbitrary size and height. Experiments on six hierarchical discussion data sets demonstrate better performance of ATHDP compared to traditional HDP based alternatives in terms of perplexity and authorship attribution accuracy.

Keywords: Hierarchical Dirichlet Processes, Topic Modeling, Message Forum

1 Introduction

Online forums (message boards) are popular social media platforms for information exchange and knowledge sharing, where users ask questions or start discussions by creating a thread, and other users post answers or comments. While some forums are specialized, general-interest forums cover a broad range of interests such as politics, health, beauty, cooking, product reviews, and so on. To help users navigate and participate, forums such as "Suomi24" (www.suomi24.fi) are organized into hierarchical sections. Hierarchical organization also occurs in online reviews for instance in retailer websites such as Amazon.com, where reviews follow the hierarchy of the products; we use Amazon reviews as a case study and point out dedicated review sites such as Yelp also feature hierarchical organization.

Three crucial aspects of online discussion are the huge diversity of interests being discussed, the huge pool of participants that contribute to the discussions, and the huge but still often structured diversity of online discussion areas where the discussion

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

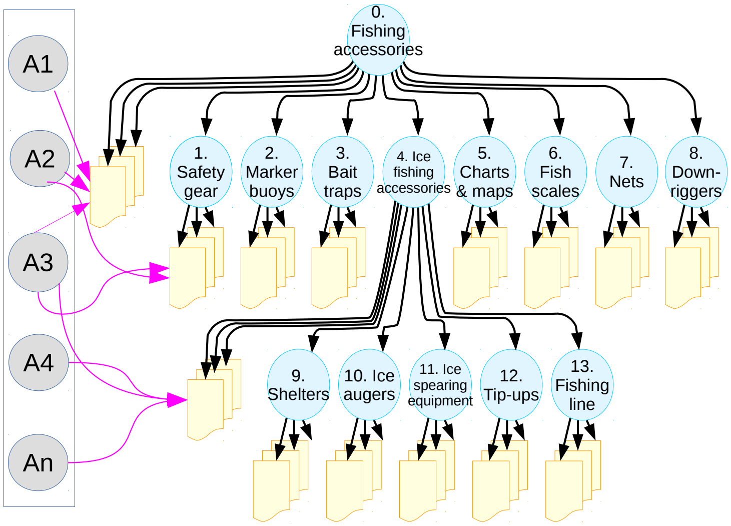

Fig. 1. Hierarchical document organization in a branch of the Amazon product hierarchy.

happens. The key question is how to take all three aspects into account in probabilistic modeling and machine learning of online discussion. In particular, the organization of online discussion areas and the identities of participants are at least partly observed data which can be taken into account for modeling the third aspect, diversity of the underlying topics of discussion.

The three aspects have different characteristics. The interests are expressed in a latent way through the observed text content, the authors are typically observed through author usernames but the pool of authors is unordered, whereas the venue is often both observed and structured: in particular, online discussion often occurs in venues having a prominent *hierarchical* organization for user-generated text content. Hierarchical structure of online forums is designed to cover a subset of prototypical user interests. However, user interests need not match the structure. For example, for an issue touching on multiple interests (say social security and mental health) there might be no dedicated section, and such issues might instead be discussed in multiple sections that each cover one of the interests. Discussion content is typically not regulated to strictly follow their section, hence users may start threads or write replies that deviate from the section theme, and threads commented by multiple users follow a mixture of their interests. Successful modeling of all three aspects of online discussions is important for studies of human online discussion behavior, for tracking trends of ideas and consumer interests, for recommender systems of discussion content or external content like targeted advertising, and for intelligent interfaces to browse and participate in discussions.

The content of the discussions is text data, and probabilistic modeling of text data is often done by generative topic models such as Latent Dirichlet Allocation [5] and Dirichlet Processes [15]. Such models represent text content of documents in an unstructured way as a bag of words arising out of a mixture of latent topics; the latent topics are fitted to a collection of documents and represent themes of discussion occurring over the collection. Basic topic models represent text content alone, whereas recent work on text mining ([14, 18] and others) has attempted author modeling for text analysis, however, most such works are not applicable to documents with observed authors in a deep hierarchical tree such as Figure 1, where documents (with yellow icons) can appear under any section (with blue icon) at any hierarchy level. The column at left denotes the pool of authors $A_1, \ldots, A_n$, where multiple authors can contribute to each thread and each author can contribute to threads at different nodes across all hierarchy levels (illustrative examples shown as purple arrows). We review related work in Section 2.

We give a solution for the challenge of effectively taking hierarchical structure of data collections and author information into account in such modeling. We introduce the *Author Tree-structured Hierarchical Dirichlet Process* (ATHDP), a new model which identifies latent topics of each section in a hierarchy and their association with authors. ATHDP is a generative model for the documents and their authors, which can model documents with multiple authors occurring at all nodes of a multi-level hierarchy. Our contributions are as follows: **1.** We develop a new nonparametric hierarchical topic model to model forum threads where multiple authors can contribute to documents, and documents and their authors can occur at any position of the section hierarchy. **2.** We develop a Gibbs sampling algorithm that extracts topics and their usage across threads and hierarchical sections. **3.** In experiments, our model outperforms the nearest state-of-the-art baseline models in terms of perplexity of held-out documents and in terms of accuracy in an author prediction task.

We point out that the latter task we consider, author prediction based on text content and section of the venue, can have many uses in online discussion venues. Authors who usually post while logged in may sometimes post with a guest username for convenience; author prediction can help associate such posts to the correct author. Moreover, when authors use different accounts on different forums, author prediction can help associate posts from the other forum to authors in the forum of interest. Author prediction models could also be applied to author similarity modeling: if posts of a known author match well (having a high classification probability) to another author, such two authors are similar and could for example be recommended as followers of each other, or could be served similar ads or other content. Such tasks assume the correct author is available in the set of candidates; in principle documents that do not match any author well could be detected simply from poor perplexity scores for all author candidates, but in this paper we do not consider such outlier detection scenarios.

The rest of the paper is structured as follows. In Section 2 we discuss related previous work. In Section 3 we introduce our new model, and in Section 4 we derive Gibbs sampling based Bayesian inference equations for the model. In Section 5 we carry out experiments on six data sets arising from two kinds of data, online forum data and online reviews data. Lastly, in Section 6 we draw conclusions.

2 Related work

A topic model [5] is a parametric Bayesian model for count data such as bag-of-words representations of text documents. Several variations expand the basic topic model setting. One of the pioneering works is the Author Topic Model (ATM) [14], which explores relationships between authors, documents, topics and words. Jiang et al. [8] recommend points of interest using ATM. However, ATM models documents arising from a uniform mixture of a group of authors, and cannot model different proportions of authors, and cannot take into account hierarchical organization of documents in a venue. Yang et al. [17] proposed a model that explores asker-answerer networks between users topics for question answering applications, however no hierarchical organization of documents is considered. Another model variant considers modeling sentiment with topics jointly [3]. Author-aware Aspect Topic Sentiment Model (AATSM) [13] explores relationship between authors and sentiment to retrieve supporting opinions from reviews; again no hierarchical document organization is considered. In Link-LDA [6] occurrences of words and entities (such as authors) are not paired. It only models that the document contains a set words and a set of entities, but not which word associated with which entity. The Entity topic model (ETM) [9] models the influence of entities on word content of topics, but does not model the influence of entities on which topics are active in the first place. Thus, it cannot not model influence of sections on active topics. Moreover, all the above models are parametric models and require the number of topics to be predefined.

Teh et al. [15] proposed the Hierarchical Dirichlet Process, a nonparametric model where the number of topics does not need to be pre-specified. However, HDP does not consider author information in the model. There are several parametric/nonparametric models that consider author information. HDPauthor [18] generates documents by a group of authors, and Junyu et al. [16] proposed an infinite author topic model based on mixed Gamma-Negative Binomial Process. However, in these models it is not known which words come from which authors. Moreover, each author always has the same topic distribution regardless of where the topics occur. The only thing that then differentiates the topic proportions of different documents is the proportions of participating authors. Thus, such models cannot properly model the influence of discussion venue sections on document content, and furthermore, these models are not readily applicable to a scenario where documents could arise at any node in a deep hierarchy of sections (as shown in Figure 1), where not only modeling the influence of sections is important, but also modeling the relationships of content among sections.

Ahmed et al. [2] create a time-dependent topic cluster model based on a recurrent Chinese restaurant process, so that content is grouped at three levels of organization such as high-level topics, individual stories, and entities over time. In PAM [11], a document is modeled as a distribution over the topics at the leaves of the topic hierarchy. In the nested Chinese restaurant process [4], a document is modeled as a distribution over a single path from the root to the leaf node. In TS-SB [1], a document is modeled by a single node of the tree. In the recursive Chinese restaurant process [10], a document has a distribution over all of the nodes of the hierarchy. In the above models, HDP is used to learn a tree structure; the difference is that in ATHDP we do not need to perform any learning on the structure of the data, our model is based on a known hierarchy which is fixed during inference. Instead we focus on modeling authors and the given hierarchy as the model structure, where documents can occur under any node in the hierarchy.

3 Author Tree-structured Hierarchical Dirichlet Process

ATHDP is a generative model for documents arising from multiple authors at different nodes of a multilevel hierarchy of sections. Each document is represented as a bag of $(word, author)$ tuples, arising out of a latent mixture of topics. Topic mixtures in the model are drawn from Dirichlet process priors: the Dirichlet process is a nonparametric prior over topic distributions that requires only a base distribution and a concentration parameter, and does not require pre-specifying the number of topics; the inference of the resulting ATHDP model will learn the number of topics from the data of documents and their authors over the hierarchy.

In the following, we describe ATHDP first as a top-down generative process from its associated graphical model shown in Figure 2. We then introduce a restaurant-related metaphor called Fine Chocolates Banquet (FCB) for the model which provides useful terminology and intuition; such food-related metaphors are commonly used in Dirichlet process based modeling, as the Dirichlet process itself is often also described as a Chinese restaurant process. The FCB metaphor will be used in the next section to describe inference for ATHDP.

Generative process. Consider a given tree hierarchy which, in a top-down fashion, can be described as a root node (root section) connected to a set of child nodes, those in turn connected to grandchild nodes, and so on. Documents can be observed under any node, not only under leaf nodes. ATHDP is a nonparametric topic model that generates a Dirichlet process prior into each node of the tree and into each document. From that prior the topic distribution of the document is drawn, and each topic generates $(word, author)$ tuples as content for the documents.

The generative process first draws a global distribution G^0_{root} from a Dirichlet process with base distribution H and concentration parameter α^0 for the root node of a given tree, denoted as $G^0_{root} \sim DP(H, \alpha^0)$. A node can contain child nodes and/or documents. We index a node with v and a document with j. Therefore, for each child section v of the root node in the tree, a discrete distribution G^1_v is generated from a Dirichlet process with base distribution G^0_{root} and concentration parameter α^1, denoted as $G^1_v \sim DP(G^0_{root}, \alpha^1)$. The process is repeated recursively for every child node to generate its grandchild sections, so that a node v at level l in the hierarchy (l steps down from the root) is generated a discrete distribution G^l_v by drawing it from a Dirichlet process with base distribution $G^{l-1}_{pa(v)}$ and concentration parameter α^l, where $pa(v)$ is the parent node of v, denoted as $G^l_v \sim DP(G^{l-1}_{pa(v)}, \alpha^l)$.The Dirichlet process priors describe which topics are active in each node; in order to generate topic content which are $(word, author)$ tuples, for each topic z two distributions are drawn, a distribution ϕ_z over the vocabulary V of possible words from a Dirichlet prior, denoted as $\phi_z \sim Dirichlet(\beta)$, and a distribution ϑ_z over the pool of possible authors A from another Dirichlet prior, denoted as $\vartheta_z \sim Dirichlet(\gamma)$, where β and γ are hyperparameters.

A document, or several documents, can arise under any node. Therefore, to generate a document j under a node v at level l, the model draws G_j from a Dirichlet process with base distribution G^l_v and concentration parameter α^{l+1}, denoted as $G_j \sim DP(G^l_v, \alpha^{l+1})$. From G_j, a topic index z_{ji} is drawn. Based on the topic index

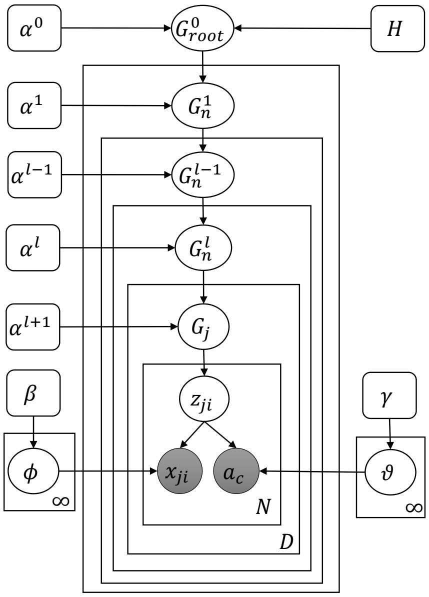

Fig. 2. Graphical model of the Author Tree-structured Hierarchical Dirichlet Process

the model samples a word x_{ji} and an author a_{ji} from the distribution of words in that topic and the distribution of authors in that topic, respectively. Figure 2 shows the plate representation graphical model of ATHDP. In summary, the full generative process of the ATHDP model is as follows:

- For each topic $z = 1, 2, \dots,$
  1. Sample a distribution over words, $\phi_z \sim Dirichlet(\beta)$.
  2. Sample a distribution over authors, $\vartheta_z \sim Dirichlet(\gamma)$.
- For the root, $G^0_{root} \sim DP(\alpha^0, H)$.
- For each section v at level l from the root, $G^l_v \sim DP(\alpha^l, G^{l-1}_{pa(v)})$.
- For each document j in section v, $G_j \sim DP(\alpha^{l+1}, G^l_v)$.
- For each word x_{ji} and author a_{ji} in a document j, $z_{ji} \sim G_j, \quad x_{ji} \sim \phi_z, \quad a_{ji} \sim \vartheta_z$

Food-based metaphor. The formal generative process above can also be described implicitly as an iterative process where documents are filled one observed $(word, author)$ tuple at a time. We describe the process by the FCB metaphor; the mathematical details are then provided in the next section as Gibbs sampling based inference equations.

In the FCB metaphor, a chocolate-tasting banquet, where dishes are assortment boxes of fine chocolates prepared by famous chocolatiers, is arranged in a multilevel palace: each level has several *food-delivery stations*, each of which serves several *restaurants* (dining rooms) at that level. Each topic in ATHDP is a dish, that is, a chocolate-assortment box containing a particular mixture of chocolate candies (words) created by a team of chocolatiers (authors). A customer chooses which assortment they want to

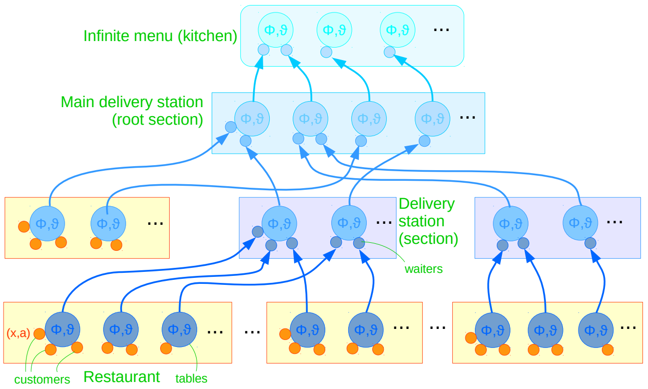

Fig. 3. An illustration of a Fine Chocolates Banquet.

eat from, and then takes a chocolate from the assortment box: each chocolate is provided in a wrapper signed by the chocolatier, thus when a customer takes a chocolate from the box they will observe a tuple of the candy itself (word) and the identity of the chocolatier (author). Attendees (i.e., customers) visit the chocolate restaurants to eat from popular dishes (popular chocolate assortments): each restaurant has tables for customers, and there is a responsible *waiter* at every table who brings a dish (chocolate-assortment box) to the table, fetching it from a table in a food-delivery station. At food-delivery stations, each table contains a pile of a particular dish (boxes of a particular chocolate-assortment), and each table also has a responsible waiter who brings the dish to the table from an upper-level delivery station, recursively. At the topmost level there is a kitchen where the chocolatiers work to create the different types of dishes (assortments). Each time a customer/waiter chooses a table, they prefer popular tables that other customers/waiters have also picked, but can also pick a new table; this property enables the FCB to make available as many dishes as are needed without specifying the number beforehand. In practice, although a potentially infinite number of dishes are available, inference yields a finite number of dishes suitable for modeling the data set.

We illustrate FCB in Figure 3. Yellow boxes are restaurants (documents), and orange circles denote customers that each pick a chocolate representing a $(word, author)$ tuple (x, a) from their table. Each table serves a dish (chocolate assortment box) which represents a topic, having a distribution ϕ over words and a distribution ϑ over authors. Each dish is brought to the table by a waiter from an upper-level delivery station (blue boxes), where each waiter chooses one table in the delivery station. Ultimately the dishes are created in the uppermost level (kitchen) which hosts an infinite menu of chocolate-assortment dishes. The content of the available dishes and their prevalences across restaurants and delivery stations are not observed, and will be inferred from the observed data as described in Section 4.

4 Inference

We introduce a Gibbs sampling scheme for ATHDP, derived based on the FCB representation. We sample tables, pointers to ancestor tables, and dishes for tables. Let $f_k^{-x_{ji},a_{ji}}(x_{ji}, a_{ji})$ denote the conditional density or likelihood of $(x_{ji},\, a_{ji})$ given all data items except $(x_{ji},\, a_{ji})$, where k is the dish at the table of $(x_{ji},\, a_{ji})$. We have for a pre-existing dish and for a brand-new dish

$$f_k^{-x_{ji},a_{ji}}(x_{ji}, a_{ji}) \propto \frac{n_{kw}^{-ji}}{n_{k.}^{-ji}} \times \frac{n_{ka}^{-ji}}{n_{k.}^{-ji}} \quad \text{and} \quad f_{k_{new}}^{-x_{ji},a_{ji}}(x_{ji}, a_{ji}) \propto \frac{1}{V \times A}$$

respectively, where is the word index of x_{ji}, a is the author index of a_{ji}, n_{kw}^{-ji} is the number of occurrences of w from dish k (other than x_{ji}), n_{ka}^{-ji} is the number of occurrences of a from dish k (other than a_{ji}), and $n_{k.}^{-ji}$ is the sum over different word indices; note that since words and authors occur in tuples, the sum over word indices is the same as the sum of n_{ka}^{-ji} over author indices. We denote

$$f_k^{-x_{jt},a_{jt}}(x_{jt}, a_{jt}) = \frac{\prod_w (\beta + n_{kw} - 1)...(\beta + n_{kw}^{-jt})}{(V\beta + n_{kw} - 1)...(V\beta + n_{k.}^{-jt})} \frac{\prod_a (\gamma + n_{ka} - 1)...(\gamma + n_{ka}^{-jt})}{(A\gamma + n_{ka} - 1)...(A\gamma + n_{k.}^{-jt})}$$

as the conditional density of $(x_{jt},\, a_{jt})$ given all data items associated with mixture component k leaving out $(x_{jt},\, a_{jt})$, where β and γ are hyperparameters.

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 $k_{jt^{new}}$:

$$p(x_{ji}, a_{ji} | \boldsymbol{t}_{-ji}, t_{ji} = t^{new}; \boldsymbol{k}) = \sum_{k=1}^{K} Q_k f_{k_{jt}}^{-x_{ji},a_{ji}}(x_{ji}, a_{ji}) + Q_{k^{new}} f_{k_{jt}^{new}}^{-x_{ji},a_{ji}}(x_{ji}, a_{ji})$$

where $\boldsymbol{t}_{-ji}$ denotes table choices of all words other than t_{ji} and $\boldsymbol{k}$ denotes dish choices of all tables, and Q_k or $Q_{k^{new}}$ denote dish probabilities that are computed recursively, traveling from a leaf node to all the way up to the root node by summing the number of tables in each node that are assigned to a topic.

$$Q_k(v) = \frac{m_{.k}^v}{m^v.. + \alpha^l} + \frac{\alpha^l}{m^v.. + \alpha^l} Q_k(pa(v)) \ ,$$

where $m_{.k}^v$ is the number of tables assigned to topic k in node v, and $m_{..}^v$ is the number of tables in node v, and l is the level of the node. Therefore, at a restaurant the conditional distribution of t_{ji} is:

$$\begin{aligned} p(t_{ji} = t) &\propto n_{jt.}^{-ji} \times f_k^{-x_{ji},a_{ji}}(x_{ji}, a_{ji}) \\ p(t_{ji} = t^{new}) &\propto \alpha^{l+1} p(x_{ji}, a_{ji} | \boldsymbol{t}_{-ji}, t_{ji} = t^{new}; \boldsymbol{k}) \end{aligned} \qquad (1)$$

Part 2. Sampling a table t from delivery-station v for a new waiter with first customer x_{ji}: The likelihood for $t_{jt} = t^{new}$ can be calculated as follows:

$$\begin{aligned} p(t_{jt}|\boldsymbol{t}_{-jt}, t_{jt} = t^{new}; \boldsymbol{k}) = &\sum_{k=1}^{K} \frac{c_{vt.}}{c_{v..} + \alpha^l} f_{k_{jt}}^{-x_{ji},a_{ji}}(x_{ji}, a_{ji}) \\ &+ \frac{\alpha^l}{c_{v..} + \alpha^l} f_{k_{jt}^{new}}^{-x_{ji},a_{ji}}(x_{ji}, a_{ji}) \end{aligned}$$

where $c_{v.k}$ is the number of tables assigned to k in node v, $c_{vt.}$ 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

$$\begin{aligned} p(t_{jt} = t) &\propto \frac{c_{vt.}^{-jt}}{c_{v..} + \alpha_j} f_{k_{tj}}^{-x_{ji},a_{ji}}(x_{ji}, a_{ji}) \\ p(t_{jt} = t^{new}) &\propto \frac{\alpha_j}{c_{v..} + \alpha_j} p(t_{jt}|\boldsymbol{t}_{-jt}, t_{jt} = t^{new}; \boldsymbol{k}) \end{aligned} \tag{2}$$

Part 3. Sampling a delivery-station table t for a waiter with several existing customers: The likelihood for $t_{jt} = t^{new}$ for many customers in a table can be calculated as follows:

$$\begin{aligned} p(t_{jt}|\boldsymbol{t}_{-jt}, t_{jt} = t^{new}; \boldsymbol{k}) = &\sum_{k=1}^{K} \frac{c_{vt.}}{c_{v..} + \alpha^l} f_{k}^{-\boldsymbol{x}_{jt},\boldsymbol{a}_{jt}}(\boldsymbol{x}_{jt}, \boldsymbol{a}_{jt}) \\ &+ \frac{\alpha^l}{c_{v..} + \alpha^l} f_{k_{new}}^{-\boldsymbol{x}_{jt},\boldsymbol{a}_{jt}}(\boldsymbol{x}_{jt}, \boldsymbol{a}_{jt}) \end{aligned}$$

Therefore, the conditional distribution of t_{jt}, given all customers in the table, is

$$\begin{aligned} p(t_{jt} = t) &\propto \frac{c_{vt.}^{-jt}}{c_{v..} + \alpha^l} f_{k}^{-\boldsymbol{x}_{jt},\boldsymbol{a}_{jt}}(\boldsymbol{x}_{jt}, \boldsymbol{a}_{jt}) \ , \\ p(t_{jt} = t^{new}) &\propto \frac{\alpha^l}{c_{v..} + \alpha^l} p(t_{jt}|\boldsymbol{t}_{-jt}, t_{jt} = t^{new}; \boldsymbol{k}) \end{aligned} \tag{3}$$

If the upper level is the root level, a dish or topic is sampled from the kitchen instead of a table pointer, and the dish is propagated to all descendants of the waiter.

We summarize the Gibbs sampling algorithm for ATHDP inference in Algorithm 1. We sample a table assignment for each $(word, author)$ tuple in a document with a recursive procedure in line 3. For a $(word, author)$ tuple, we sample a table using Eq. (1), and we sample a parent table using Eq. (2) from delivery stations. If it's a new table, then we move to the parent node to sample a table from the parent node in line 15. The process is repeated until a parent table is selected or the root node is reached. If the root node is reached a topic selected using Eq. (3). After that, we update the topic of all tables in the descendant's nodes of the table in the root. We maintain a data structure to keep track of topics of all tables. For simplicity, we do not include them in the algorithm. 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 Eq. (3). We repeat the process until the root is reached and eventually sample a topic for the root table using Eq. (3).

Algorithm 1 Gibbs Sampling for ATHDP

```
Input: words w in documents d, # topics K, # iterations I
Output: Topic assignments z
for $i$ in $I$ do
  for $w$ in d do
    SampleTable(node)
  for $t$ in d do
    SampleParentTable(node)
Procedure SampleTable(node)
  if node == document then
    $table$ ← Sample a table by Eq.(1)
  else
    $table$ ← Sample a table using Eq. (2)
  if $table == t_{new}$ then
    if node.parent == root.node then
      $topic$ ← Sample a topic by Eq. (3)
    else
      SampleTable(node.parent)
Procedure SampleParentTable(node)
  if node.parent == root.node then
    $topic$ ← Sample a topic by Eq. (3)
  else
    node ← node.parent
    $table$ ← Sample a table by Eq. (3)
    if $table == t_{new}$ then
      node← node.parent
      SampleParentTable(node)
```

Table 1. Data sets. Total document counts at different levels from the root given in the 4th column.

	Threshold	# Authors	# nodes	#docs at different level of the tree	# Train docs	# Test docs
Amazon Sports	50	97	599	100, 57, 434, 1655, 3775, 645, 8	5965	709
Amazon Food	100	26	40	3166, 2, 56, 42, 22	2948	340
Amazon Home	100	37	567	25, 85, 227, 2015, 2179, 394, 80	4489	516
Amazon Health	100	67	484	107, 137, 550, 5501, 2912, 209	8444	972
S24 Relationship	100	50	25	0, 12148	13425	1514
S24 Health	20	71	64	0, 464, 1882	2509	661

5 Experimental Results

We evaluate ATHDP's performance by two performance measures: (1) held-out perplexity, representing ability to model unseen documents and (2) author prediction of unseen documents. Since the methods described in the related work are all not directly applicable to our case, we take the Hierarchical Dirichlet Process [15] as a baseline that would be readily available to the practitioner, and we use it in two ways to take author information into account, as described in the *Quantitative comparison* paragraph below. We used Gibbs sampling to train the models and took a sample at 100th iterations. We first describe the data sets, summarized in Table 1. We begin by a qualitative analysis of ATHDP results, and then present quantitative comparisons between ATHDP and comparison methods.

We used two different data sources, *Suomi24* (s24) and Amazon for our experiments. S24 has in total 2434 sections in the hierarchy. The data source [3] is publicly available in original and lemmatized forms. From this source, we created several datasets

[3] https://www.kielipankki.fi/corpora/

Table 2. Sample ATHDP topic proportion for three sections in the food data set

Section id	Section name	Top 3 topic proportions for each section
44	Peanut Butter	66:0.46, 86:0.31, 37:0.23
3	Coffee Substitutes	9:0.247, 28:0.24, 37:0.24
292	Jams & Preserves Gifts	22:0.62, 9:0.36, 48:0.002

Table 3. ATHDP topics for Amazon food data set, sections where they are active, and top words

Topic	Top 3 author ids	Stemmed top words of the topic
9	10, 19, 7	cup recommend good coffe flavor tast drink pack highli brew keurig energi tea brewer make bold free star larg
22	18, 20, 23	make enjoy tast nice flavor bit good eat work meal ad cup star love give lot packag mix protein morn
28	23, 0, 1	coffe tast flavor great recommend highli make cup love chocol tea good stuff bit year product buy thing amaz awesom
37	14, 1, 0	good tast coffe flavor great love make product sweet tea time chocol stuff perfect snack work cup soup free chicken
48	7, 5, 24	clean top cook grape flavor work nice red kit simpl week conveni good great allergi bag expect sodium basic water
66	17, 7, 19	sugar calori product protein flavor ingredi bar tast fiber fat high oil organ food wheat natur time sweet make raisin
86	24, 17, 10	bar protein tast flavor calori bit eat snack fiber good meal cinnamon sugar textur fill chocol nice ingredi diet raisin

by taking thematic branches of the hierarchy, such as s24 relationship, s24 health for our experiments. The second data source is *reviews on Amazon.com*, one of the top shopping sites in the world with hundreds of shopping sections. We select thematic branches corresponding to several top categories (or department) such as Sports and Outdoors (Sports), Home and Kitchen (Home), Health and Personal Care (Health), Clothing Shoes and Jewelry (Clothes), Grocery and Gourmet Food (Food). Under each top category the site contains many sections. For example, there are 1933 sections under sports [7]. We select reviewers that have more than 50 or 100 reviews in each category. For each reviewer we randomly select 90% reviews for training and 10% reviews for testing. The numbers of train and test reviews for each category along with the number of reviews in different levels of the hierarchy are given Table 1. In Amazon data sets each product is considered as a document, and in s24 data sets each thread is considered as a document. A document consists of many reviews or comments from many authors.

Qualitative analysis of ATHDP results. We examine how ATHDP topics covered themes within and across sections. For brevity we present the analysis of the food dataset only. We present latent topics and their proportions for three sample section, as described in Table 2. We see that sections are mixed of latent topics with different proportions. For example, the top 3 topics of the Peanut Butter section are 66, 86, and 37. The top words of each latent topic of the sample sections are presented in Table 3. We observe that extracted latent topics covered many themes including section themes. For example, top words of topics 66, 86, and 37 include many words related to peanut butter

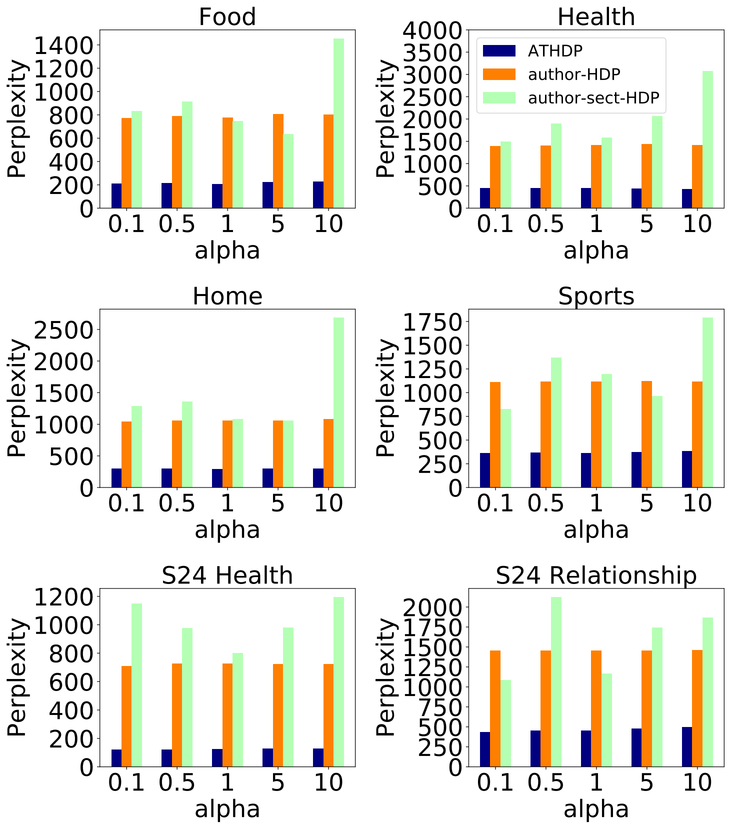

Fig. 4. Perplexity on held-out data sets with different alpha values. Smaller values are better; ATHDP outperforms author-HDP and author-sect-HDP.

including fat, oil, protein, sugar, calori and so on. We observe that topic 66 is directly related to peanut butters and can be regarded as discussion of *bread spreads*. There is some overlap between top words of topics 66 and 86. By looking at distinct words we observed that people are discussing diet, snack, meal, cinnamon, chocolate etc. in the peanut butter section, which refers to *how good peanut butter is as a diet*. Topic 37 is about having coffee or tea, which is an occasion where peanut butter based breads might also be enjoyed, hence it is a discussion of a *use scenario of peanut butter*. Table 3 also shows top 3 author ids for each topic. Overall, ATHDP extracts reasonable meaningful word-topic, section-topic and topic-author distributions. We also verified that the results regarding datasets other than food were similar.

Quantitative comparison. We compare ATHDP to two baseline models in two tasks, modeling of previously unseen documents and in author prediction. We use HDP as a baseline, which takes the hierarchy into account in two simple ways – model all documents belonging to the same author (author-HDP), and all documents belonging to the same author-section pair (author-sect-HDP). In author-HDP, for example, the sports

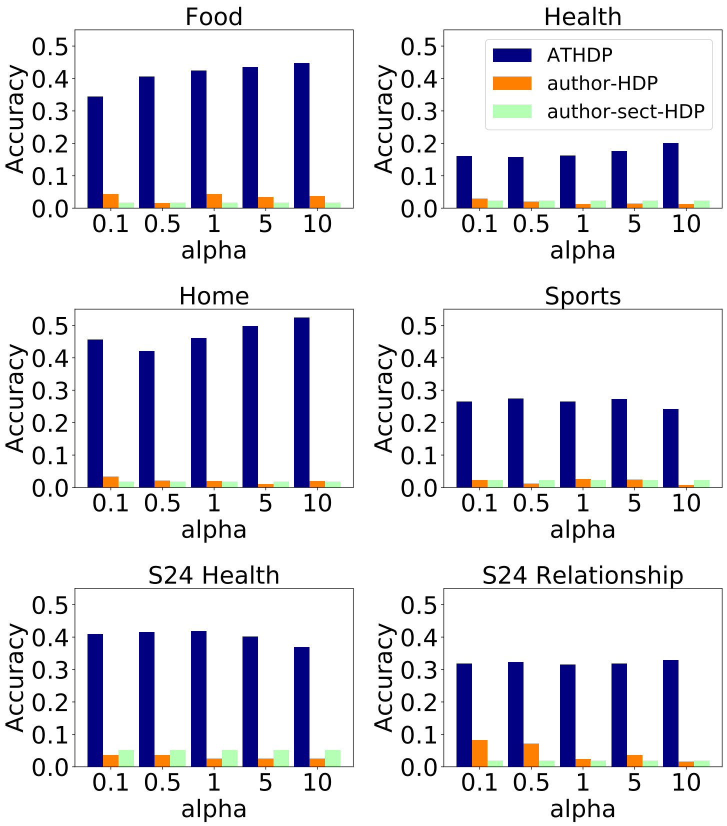

Fig. 5. Author prediction accuracy in different data sets with different alpha values. Larger values are better; ATHDP outperforms author-HDP and author-sect-HDP.

dataset consists of 97 authors, therefore we train 97 HDP models. In author-sect-HDP we train as many HDP models as there are author-section pairs. For ATHDP, we train a single model for each dataset. We run ATHDP and the author-HDP and author-sect-HDP baselines for all data sets with different values of the concentration hyperparameter alpha. We use the two HDP-based baselines as we found no related work that could fully take into account author information and the hierarchical structure of our data where documents arise in multiple places in the hierarchy, and the aim is to take the known document hierarchy into account. The baselines thus represent a natural way to run the existing HDP model with known divisions of data based on authors or based on authors and sections.

Modeling of previously unseen documents. We evaluate the ability of the proposed model to represent new incoming documents, by computing perplexity of held-out documents, a standard metric in information retrieval literature. We compute perplexity on held-out test documents as described in Table 1 as follows: $perplexity(D_{test}) = \frac{1}{M}\sum_{d=1}^{M} \exp\left(-\frac{logP(w_d)}{N_d}\right)$. We compute perplexity for different α values. We use the same alpha values for all levels in ATHDP. The results are shown in Figure 4. Lower perplexity indicates the better model. We observe that ATHDP outperforms author-HDP and author-sect-HDP in perplexity for all the data sets and alpha values. The overall difference between ATHDP and the author-HDP, and between ATHDP and author-sect-HDP, is statistically significant at the $p = 0.05$ level: for both comparisons we have $p = 0.03125$ from the exact binomial test over the six data sets. Note that since ATHDP outperforms the alternatives regardless of alpha value, the choice of alpha value used to represent each method does not affect the result of the test.

Author prediction. We compare the ability of different models to predict the author of a previously unseen document, that is, to classify new documents to correct authors. To predict the author for each test document, we compute perplexity for the test document under the model for each author, and assign the document to the author that yields the lowest perplexity. We report the author prediction accuracy results in Figure 5. We observe that ATHDP outperforms author-HDP and author-sect-HDP by a large margin. The overall difference between ATHDP and the author-HDP, and between ATHDP and author-sect-HDP, is again statistically significant at the $p = 0.05$ level: for both comparisons we have $p = 0.03125$ from the exact binomial test over the six data sets. As ATHDP outperforms the alternatives regardless of alpha value, the choice of alpha value used to represent each method again does not affect the result of the test.

The author prediction accuracies achieved by ATHDP are good, especially considering the large number of potential candidate authors. ATHDP accuracy results are up to about 50% accuracy, which although not a flawless score is practically usable for attribution (note that when there are numerous potential authors random guessing yields far worse accuracies than 50%). In contrast, the alternative systems perform poorly; a possible explanation is that the author-HDP model is unable to take hierarchical section-based variation of the authors' interests properly into account, whereas the author-sect-HDP model does not make full use of the hierarchical relationships between sections and hence has too little data per author-section combination to learn good models of authors' interests in each section. In contrast, ATHDP learns the topics and their variation across the hierarchy together, allowing successful modeling of author interests.

6 Conclusions

We introduced the Author Tree-structured Hierarchical Dirichlet process (ATHDP), a nonparametric probabilistic model of documents and their authors in a deep tree-structured hierarchical discussion venue where documents can arise at any tree node. ATHDP can to extract topics across the documents and sections in the hierarchy, and automatically computes the number of topics required to model the authors and text content across the hierarchical sections. ATHDP does not restrict content of topics to strictly match predefined sections, but infers them in a data driven way to describe users' interests. In experiments, ATHDP outperformed HDP based alternative models in modeling unseen documents (measured by perplexity), and author prediction of unseen documents (measured by accuracy).

In this first work ATHDP already proved a very well-performing and flexible model. In future work, its performance could be evaluated by a larger set of different measures, and the flexibility of the model could be further increased by, for example, modeling within-topic correlations between authors and word content, or by other such extensions. We also plan to integrate ATHDP in systems that can make use of the topic models, i.e., utilizing ATHDP topics in different applications such as recommendation [8], [19], and interactive exploratory search [12].

References

1. Adams, R., Ghahramani, Z., Jordan, M.: Tree-structured stick breaking for hierarchical data. In: Proc. NIPS. pp. 19–27. Curran Associates Inc. (2010)
2. Ahmed, A., Ho, Q., Teo, C.H., Eisenstein, J., Smola, A.J., Xing, E.P.: Online inference for the infinite topic-cluster model: Storylines from streaming text. In: Proc. AISTATS. pp. 101–109 (2011)
3. Alam, M.H., Ryu, W.J., Lee, S.: Joint multi-grain topic sentiment. Inf. Sci. **339**(C), 206–223 (Apr 2016)
4. Blei, D., Griffiths, T., Jordan, M.: The nested Chinese restaurant process and Bayesian nonparametric inference of topic hierarchies. J ACM **57**, 7:1–7:30 (2010)
5. Blei, D., Ng, A., Jordan, M.: Latent Dirichlet allocation. J Mach Learn Res **3**, 993–1022 (2003)
6. Erosheva, E., Fienberg, S., Lafferty, J.: Mixed-membership models of scientific publications. Proceedings of the National Academy of Sciences **101**(suppl 1), 5220–5227 (2004)
7. He, R., McAuley, J.: Ups and downs: Modeling the visual evolution of fashion trends with one-class collaborative filtering. In: Proc. WWW. pp. 507–517 (2016)
8. Jiang, S., Qian, X., Shen, J., Fu, Y., Mei, T.: Author topic model-based collaborative filtering for personalized poi recommendations. IEEE Trans. Multimedia **17**(6), 907–918 (2015)
9. Kim, H., Sun, Y., Hockenmaier, J., Han, J.: ETM: entity topic models for mining documents associated with entities. In: Proc. ICDM. pp. 349–358. IEEE Computer Society (2012)
10. Kim, J., Kim, D., Kim, S., Oh, A.: Modeling topic hierarchies with the recursive Chinese restaurant process. In: Proc. CIKM. pp. 783–792. ACM (2012)
11. Li, W., McCallum, A.: Pachinko allocation: DAG-structured mixture models of topic correlations. In: Proc. ICML. pp. 577–584. ACM (2006)
12. Peltonen, J., Belorustceva, K., Ruotsalo, T.: Topic-relevance map: Visualization for improving search result comprehension. In: Proc. IUI. pp. 611–622. ACM (2017)
13. Poddar, L., Hsu, W., Lee, M.L.: Author-aware aspect topic sentiment model to retrieve supporting opinions from reviews. In: Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing. pp. 472–481. Association for Computational Linguistics (2017)
14. Rosen-Zvi, M., Griffiths, T., Steyvers, M., Smyth, P.: The author-topic model for authors and documents. In: Proc. UAI. pp. 487–494. AUAI Press (2004)
15. Teh, Y., Jordan, M., Beal, M., Blei, D.: Hierarchical Dirichlet processes. J Am Stat Assoc **101**, 1566–1581 (2006)
16. Xuan, J., Lu, J., Zhang, G., Xu, R.Y., Luo, X.: A Bayesian nonparametric model for multi-label learning. Mach. Learn. **106**(11), 1787–1815 (Nov 2017)
17. Yang, L., Qiu, M., Gottipati, S., Zhu, F., Jiang, J., Sun, H., Chen, Z.: CQArank: Jointly model topics and expertise in community question answering. In: Proc. CIKM. pp. 99–108. ACM (2013)
18. Yang, M., Hsu, W.H.: HDPauthor: A new hybrid author-topic model using latent Dirichlet allocation and hierarchical Dirichlet processes. In: Proc. WWW. pp. 619–624. ACM (2016)
19. Zhang, S., Zhang, S., Yen, N.Y., Zhu, G.: The recommendation system of micro-blog topic based on user clustering. Mob. Netw. Appl. **22**(2), 228–239 (Apr 2017)