Modeling Topic Hierarchies with the Recursive Chinese Restaurant Process

Joon Hee Kim, Dongwoo Kim, Suin Kim, and Alice Oh
Department of Computer Science
Korea Advanced Institute of Science and Technology
Daejeon, Republic of Korea
{joon.kim,dw.kim,suin.kim}@kaist.ac.kr, alice.oh@kaist.edu

ABSTRACT

Topic models such as latent Dirichlet allocation (LDA) and hierarchical Dirichlet processes (HDP) are simple solutions to discover topics from a set of unannotated documents. While they are simple and popular, a major shortcoming of LDA and HDP is that they do not organize the topics into a hierarchical structure which is naturally found in many datasets. We introduce the recursive Chinese restaurant process (rCRP) and a nonparametric topic model with rCRP as a prior for discovering a hierarchical topic structure with unbounded depth and width. Unlike previous models for discovering topic hierarchies, rCRP allows the documents to be generated from a mixture over the entire set of topics in the hierarchy. We apply rCRP to a corpus of New York Times articles, a dataset of MovieLens ratings, and a set of Wikipedia articles and show the discovered topic hierarchies. We compare the predictive power of rCRP with LDA, HDP, and nested Chinese restaurant process (nCRP) using held-out likelihood to show that rCRP outperforms the others. We suggest two metrics that quantify the characteristics of a topic hierarchy to compare the discovered topic hierarchies of rCRP and nCRP. The results show that rCRP discovers a hierarchy in which the topics become more specialized toward the leaves, and topics in the immediate family exhibit more affinity than topics beyond the immediate family.

Categories and Subject Descriptors

G.3 [**Probability and Statistics**]: Nonparametric Statistics; H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval—*clustering*; I.2.6 [**Artificial Intelligence**]: Learning

Keywords

Hierarchical Topic Modeling, Bayesian Nonparametric models

CIKM'12, October 29–November 2, 2012, Maui, HI, USA.
Copyright 2012 ACM 978-1-4503-1156-4/12/10 ...$15.00.

1. INTRODUCTION

Probabilistic topic models [5, 20] are important tools for discovering the latent semantic patterns in various data including text [11, 13], users [12, 24] and movie ratings [18]. A major limitation of these basic topic models and many of their extensions is that they discover topics in flat structures without organizing them into groups or hierarchies. This is a significant limitation because in many domains, topics can be naturally organized into hierarchies where the root topic of each hierarchy is the most general topic, and the topics become more specific toward the leaves. Consider, for example, the domain of movies, where there are genres (e.g., *action*) and sub-genres (e.g., *martial arts*)[1]. One branch of the tree may have the genre *action* as a topic, and children topics *martial arts* and *James Bond series*, and another branch may have the genre *comedy* and as children *slapstick* and *black comedy.* A user with preferences, then, should be allowed to be associated with both the *action* topic to indicate that his preferences span a wide variety of action movies, as well as the *black comedy* sub-genre to indicate that his preferences for comedy are limited to that sub-genre.

There have been previously proposed topic models that look at correlations among the topics [4] and hierarchical topic structure [2, 16], but these models do not fully exhibit the following three characteristics of an intuitive and flexible topic structure. First, the number of topics should be unbounded, and an optimal number should be automatically determined by the model. Second, topics should be structured in a hierarchy of unbounded depth from general to specific, and similar topics should form groups within the hierarchical structure. Third, a document should be composed of multiple topics from anywhere in the hierarchy of topics, not just a single topic, the topics of a single path, or the topics at the bottom of the hierarchy. Detailed comparisons with other related models will be presented in the next section.

We propose a novel prior, recursive Chinese restaurant process (rCRP), and a hierarchical topic model with rCRP as a prior that can handle such flexible hierarchical topic modeling. The topic hierarchies found by rCRP are consistent with the general intuition that the topics start out quite general at the root level and become more specialized toward the leaves. Additionally, the topics within the immediate family (i.e., a parent topic and its direct children) are much more similar than the topics outside the family boundary. This characteristic of the topic hierarchy is more natural and fitting for many domains where each data point

[1] http://visual.ly/complete-list-film-sub-genres

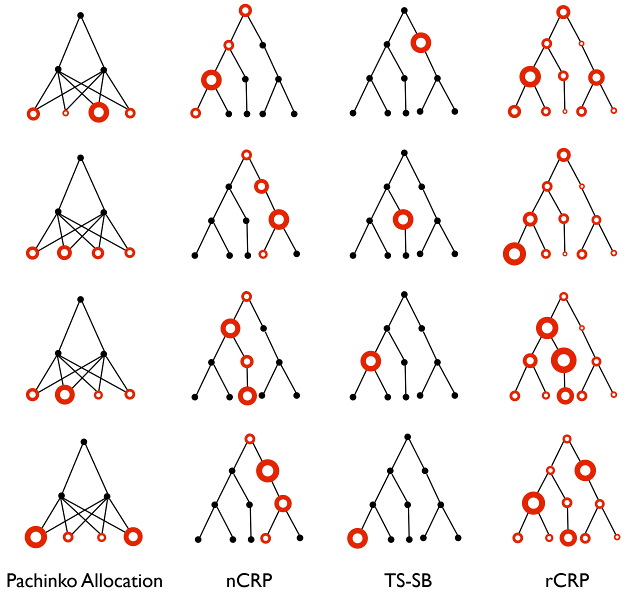


Figure 1: In PAM [14], a document is modeled as a distribution over the topics at the leaves of the topic hierarchy. In nested CRP [2], a document is modeled as a distribution over a single path from the root to the leaf node. In TS-SB [1], a document is modeled by a single node of the tree. In rCRP, a document has a distribution over all of the nodes of the hierarchy.

is best explained by a variety of topics, general to specific and placed anywhere within the hierarchy. In addition to the flexibility of the model, rCRP outperforms LDA, HDP, and nCRP [2] on the predictive metric of heldout likelihood.

The rest of this paper is organized as follows. In Section 2, we discuss existing hierarchical topic models and how they differ from rCRP. In Section 3, we describe our model with the novel nonparametric prior, the recursive Chinese restaurant process. In Section 4, we present a Markov chain Monte Carlo inference algorithm for approximating the posterior probability. In Section 5, we describe the three datasets used for experiments and visualize the topic hierarchies found for those datasets. We also compare our model against LDA, HDP, and nCRP on heldout likelihood. In Section 6, we propose two new metrics for quantifying the characteristics of a topic hierarchy and show the results of our model and nCRP. In Section 7, we conclude the paper with discussions and future directions.

2. TOPIC HIERARCHIES

Two classes of previously proposed models address hierarchical structures: ones that cluster each of the documents into the nodes of the hierarchy [9, 17, 23], and ones that place each of the topics into the nodes of the hierarchy [2, 14, 16]. Within the latter class of models, none are flexible enough to accommodate the intuition that a document exhibits multiple topics, and those topics can come from anywhere in the hierarchy of topics, from the general root-level topic down to the most specific leaf node topic, and along any path of the tree.

The different assumptions of the related models, pachinko allocation model (PAM) [14, 16], nested Chinese restaurant process (nCRP) [2], tree-structured stick-breaking process (TS-SB) [1], as well as our rCRP model are illustrated in Figure 1. The figure shows the topic assignments for four fictitious documents highlighting the different assumptions of the four models. Each of the rows is a document, modeled by each of the four models in the columns. The different sized thick circles represent proportions of the document generated by a topic represented by that node. PAM [14] is a generalization of the LDA that enables learning of topics in a form of a directed acyclic graph. In [16], the model is further extended to explicitly identify the word distribution of super-topics and sub-topics. PAM and its extension assume that the documents are generated by only the leaf node topics. The nCRP [2] extends the original CRP representation and constructs a tree-structured hierarchy of topics. This model assumes that a topic is represented by a path from the root to a particular leaf node, and each document is generated by a single path. TS-SB [1] can be used to discover a hierarchy of mixture components with each data point belonging to a component. This model assumes that each document is generated by only a single node which has its unique topic distribution. Our proposed model with recursive CRP (rCRP) as a nonparametric prior enables a document to have a distribution over the entire topic tree.

3. RECURSIVE CHINESE RESTAURANT PROCESS

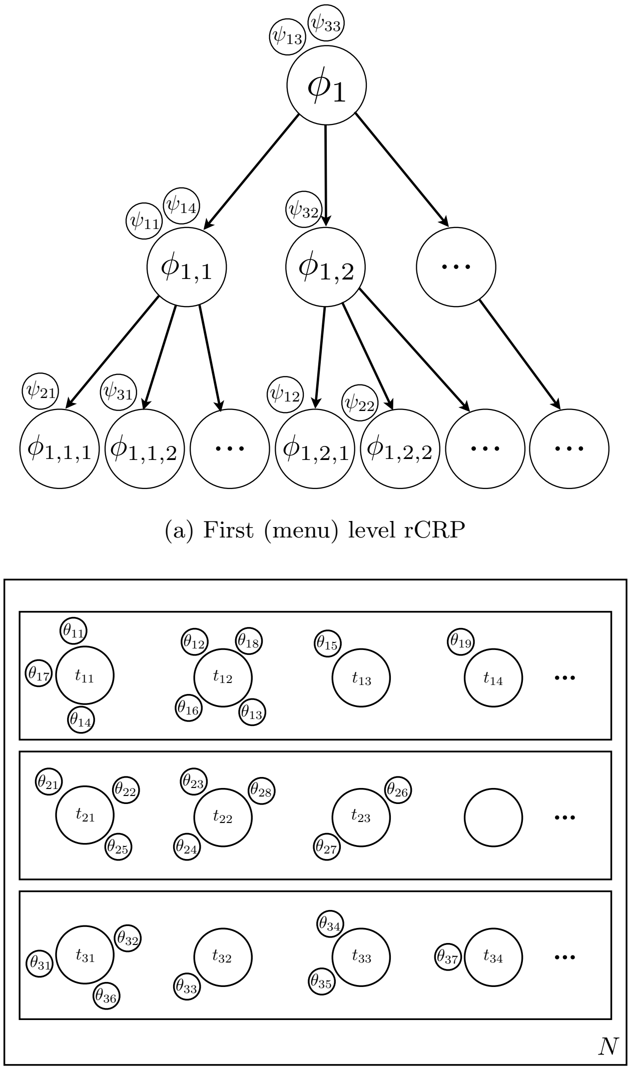
Chinese Restaurant Process (CRP) is a stochastic process that generates an exchangeable partition of data points. Due to its flexibility and extendability, CRP is widely used in nonparametric topic models. One of the most widely used model is hierarchical Dirichlet process mixture model [20]. It combines two levels of CRP to construct a mixture model of grouped data. The second level CRP partitions data points into homogeneous groups, while the first level CRP associates each group with a mixture component.

Our model also employs two levels of partitioning process. However, we discover the hierarchical structure of the mixture components by replacing the first level CRP with a new stochastic process called recursive Chinese Restaurant Process (rCPR). In rCRP, mixture components are organized in an infinite tree, and each group of data points can be associated with any node in the tree. As in [20], second level CRP is used to partition data points into groups.

With this setting, we propose a nonparametric Bayesian model that is capable of uncovering the hierarchical structure of the mixture components. In Section 3.1 we present the metaphor and notations, and review CRP. In Section 3.2 we provide a detailed description of rCRP. In Section 3.3 we formalize the generative process of our model.

3.1 Table Assignment with CRP

We maintain most of the basic assumptions and metaphors used in previous nonparametric probabilistic models as they have been proven to have strong explanatory power over data. A document is represented as a restaurant, and words are represented as customers in the restaurants. Words that convey homogeneous semantic theme are grouped together as customers of similar taste sit at the same table. A dish is chosen for each table from the global menu of dish tree,



(a) First (menu) level rCRP

(b) Second (customer) level CRP

Figure 2: Our proposed model consists of two levels of partitioning process. In the diagram, ϕ represents a dish, ψ represents a table, and θ represents a customer. The first level rCRP associates each table to a dish. The second level CRP associates each customer to a table.

which corresponds to topic assignment for each group of words. In our proposed model, the assignment of each customer to a table is determined by CRP. The association between tables and dishes is governed by rCRP.

CRP is a stochastic process that generates a random partition of discrete data. The table assignment probability distribution for a particular customer is as follows. The first customer always sits on the first table. The ith customer sits on a table depending on a draw from the following distribution

$$p(t^{th} \text{ table} \mid \text{previous assignments}) = \frac{n_t}{\alpha + i - 1} \quad (1)$$

$$p(\text{new table} \mid \text{previous assignment}) = \frac{\alpha}{\alpha + i - 1} \quad (2)$$

where n_t is the number of customers already sitting at t^{th} table of the restaurant, and α is a parameter governing the likelihood of choosing of a new table.

3.2 Dish Assignment with recursive CRP

rCRP is an extension to CRP that assumes an infinite tree structure of the mixture components. Sticking to the metaphor, rCRP assigns a dish from the global menu for each table. The menu is an infinite tree of dishes unbounded in both branching factor and height. Indexing dishes in the infinite tree is nontrivial, as the number of potential dishes is uncountably infinite whereas the number of integers used for ordinary indexing scheme is only countably infinite. Therefore we utilize the index set of strings of integers. The root dish has an index of 1. Let k be an index string of a particular dish on the menu, i'th child of dish ϕ_k is named $\phi_{k,i}$. This is visualized in Figure 2(a).

To find a dish for a particular table, we perform a recursive search beginning from the root dish. Let ϕ_k be the dish that is currently under examination. Then we make one of the three choices. The recursive search stops only when the first choice is made.

1. Choose ϕ_k
2. Choose one of the existing child dish of ϕ_k
3. Create a new child dish of ϕ_k, and choose it

We introduce the notations that will be used in the formal definition of the conditional probability of dish assignment. Let n_{jtk} be the number of customers at table t of restaurant j eating dish k. We replace an index with dot to signify that the count is marginalized. For example, $n_{jt\cdot}$ is the number of customers at table t of restaurant j, and $n_{j\cdot k}$ is the number of customers at restaurant j eating dish k. We use m_{jk} to count the number of tables at restaurant j serving dish k. Likewise, $m_{j\cdot}$ is the number of tables at restaurant j, and $m_{\cdot k}$ is the total number of tables serving dish k at any restaurant. Finally, we use $M_{\cdot k}$ to denote the cumulative counts of $m_{\cdot k}$ summed over for all dishes that are descendants of ϕ_k including ϕ_k itself. The need for this cumulative counts is illustrated shortly.

Now we formalize the probability of three choices. To find the dish for table t in restaurant j, we perform recursive search from the root dish. Let ϕ_k be the current dish, then we draw from the following distribution

$$\begin{aligned}
p(\phi_k|\text{previous tables}) &= \frac{m_{\cdot k}^{-jt}}{M_{\cdot k}^{-jt} + \gamma^n} \\
p(\phi_{k'}|\text{previous tables}) &= \frac{M_{\cdot k'}^{-jt}}{M_{\cdot k}^{-jt} + \gamma^n} \\
p(\phi_{k_{\text{new}}}|\text{previous tables}) &= \frac{\gamma^n}{M_{\cdot k}^{-jt} + \gamma^n}
\end{aligned}$$

where $\phi_{k'}$ is a direct descendent of ϕ_k, and ϕ_{new} is a new child dish of ϕ_k.

Dishes are equivalent to topic distributions used in the generation of documents, and each dish is drawn from a level-specific Dirichlet distribution. Let ϕ_k be the dish indexed by k, then it is generated as follows:

$$\phi_k \sim Dir(\beta^{\delta(k)}),$$

where $\delta(k)$ is a depth of current dish. We use $\beta^{\delta(k)}$ as a prior of Dirichlet distribution. Because symmetric Dirichlet distribution generates more sparse distribution with small values of parameter, we can expect more sparse topics with increasing depth of k when β is less than one.

3.3 Generative Process

We employ the two stages of generative process as described in Section 3.1 and Section 3.2. Now we formally describe the generative process.

Topic Tree Generation The measure G_{tree} of the global topic tree is drawn from the rCRP.

$$G_{tree} \sim rCRP(\alpha)$$

Document Generation G_j, the topic distribution of jth document, is distributed according to G_{tree}. θ_{ji} denotes the topic of ith word in the jth document, and x_{ji} denotes the word generated from the topic.

$$\begin{aligned} G_j &\sim DP(G_{tree}) \\ \theta_{ji} &\sim G_j \\ x_{ji} &\sim F(\theta_{ji}) \end{aligned}$$

4. POSTERIOR INFERENCE

In this section, we develop a Markov Chain Monte Carlo algorithm for posterior sampling of table and dish assignments. Generally, computing an exact posterior of DP and its related models is intractable. Several approaches have been employed to compute the approximate posterior. Possible approaches include (1) a Pólya urn scheme based on the marginalization of unknown infinite-dimensions [15, 7], (2) a truncation approximation which limits the complexity of the model from infinite dimensions to finite dimensions [10], (3) a variational inference which converts inference algorithms into optimization problems [3, 21]. In this work, we employ the Pólya urn scheme by incorporating the CRP metaphor for approximate inference.

Before we discuss the posterior inference algorithm, let us define variables of interest. x_{ji} indicates the ith observed word of jth document, and θ_{ji} denotes the topic of x_{ji}. ϕ_k is an atom of G_{tree}. ψ_{jt}, which denotes the topic of tth table in the jth document, is an atom of G_j. Note that each θ_{ji} is associated with one ψ_{jt} since the topic assigned to x_{ji} must correspond to the topic assigned to the table in which x_{ji} is seated. Likewise, each ψ_{jt} is associated with one ϕ_k.

For the posterior inference, we marginalize out ϕ_k, ψ_{jt}, and θ_{ji}. Therefore we need to sample the assignment relationship between these variables rather than sampling the quantities of variables themselves. For this purpose, we introduce two index variables. t_{ji} is the index variable of tables such that $\psi_{jt_{ji}} = \theta_{ji}$, and k_{jt} is the index variable of topics such that $\phi_{k_{jt}} = \psi_{jt}$.

First we write out the conditional density of x_{ji} given dish k in the inference steps for convenience. Each x_{ji} is drawn from some ϕ_k, and ϕ_k is drawn from its level distribution, $\text{Dir}(\beta^{|k|})$. Therefore, by marginalizing out ϕ_k, we can simply compute the conditional density. Letting $\mathbf{x}_k = \{x_{j'i'}; k_{jt_{j'i'}} \in \Lambda(k)\}$, the conditional density of x_{ji} only depends on the other $\mathbf{x}_k$ already assigned to that dish and its decendents, and can be computed as follows

$$\begin{aligned} p(x_{ji}|\mathbf{x}^{-ji}, \mathbf{t}, \mathbf{k}) &= \frac{p(x_{ji}, \mathbf{x}^{-ji}|\mathbf{t}, \mathbf{k})}{p(\mathbf{x}^{-ji}|\mathbf{t}, \mathbf{k})} \\ &= \frac{\int p(x_{ji}|\phi_k) \prod_{x_{j'i'}}^{\mathbf{x}_k} p(x_{j'i'}|\phi_k) p(\phi_k|\beta_k) d\phi_k}{\int \prod_{x_{j'i'}}^{\mathbf{x}_k} p(x_{j'i'}|\phi_k) p(\phi_k|\beta_k) d\phi_k}. \end{aligned}$$

We can further simplify the above equation by utilizing the Dirichlet-multinomial conjugacy as

$$p(x_{ji}|\mathbf{x}^{-ji}, \mathbf{k}) = \frac{\sum_{\{j'i'; x_{j'i'} \in \mathbf{x}_k\}} 1[x_{j'i'} = x_{ji}] + \beta^{|k|}}{\sum_{\{j'i'; x_{j'i'} \in \mathbf{x}_k\}} 1 + V\beta^{|k|}},$$

where V is the size of the vocabulary. With the Pólya urn based sampling scheme, we can efficiently sample from the above distribution by marginalizing out unknown infinite dimensional distributions. Thus, our variables of interest are index variable of tables and dishes, namely t_{ji}, and k_{jt}. It is natural to sample table t_{ji} before sampling dish k_{jt} with the CRF metaphor, so we start with sampling t_{ji}.

Sampling t The conditional distribution of t_{ji} given x_{ji} is proportional to the number of customers sitting at table t times the probability of x_{ji} being observed under table t, which can be written as

$$p(t_{ji} = t|rest) \propto \begin{cases} n_{jt}^{-ji} \times p(x_{ji}|\mathbf{x}^{-ji}, \mathbf{t}^{-ji}, t_{ji} = t, \mathbf{k}) \\ \alpha \times p(x_{ji}|\mathbf{x}^{-ji}, \mathbf{t}^{-ji}, t_{ji} = t_{\text{new}}, \mathbf{k}), \end{cases}$$

where the probability of sitting at a new table can be found by marginalizing over all available dishes.

$$\begin{aligned} & p(x_{ji}|\mathbf{x}^{-ji}, \mathbf{t}^{-ji}, t_{ji} = t_{\text{new}}, \mathbf{k}) \\ & \quad = \sum_{k \in K} \frac{m_{.k}}{m_{..} + \gamma} p(x_{ji}|\mathbf{x}^{-ji}, \mathbf{t}, \mathbf{k}) \\ & \quad + \frac{\gamma}{m_{..} + \gamma} p(x_{ji}|\mathbf{x}^{-ji}, \mathbf{t}, \mathbf{k}_{\text{new}}) \end{aligned}$$

Sampling k The posterior sampling of k_{jt} involves a sequence of search along the menu tree. We begin from the root dish and move down along the tree until we find the dish. Suppose we want to sample a dish for customers at table t in restaurant j. We perform a recursive search beginning from the root dish as illustrated in Algorithm 1. The conditional probability of k_{jt} is the prior probability of k times the likelihood of $\mathbf{x}_{jt}$ being observed under dish k. The prior depends on k. If $k = k_{\text{current}}$, it is proportional to the number of tables serving dish k. Otherwise, it is proportional to the number of tables serving dish k or any of

Algorithm 1 Sampling k_{jt} by recursive algorithm

```
function samplingK($k_{\text{current}}$)
$k_{\text{next}} \sim p(k_{jt} = k|t, k^{-jt}, k_{\text{current}})$
if $k_{\text{next}} = k_{\text{current}}$ then
    return k
else if $k_{\text{next}}$ = child of $k_{\text{current}}$ then
    $k_{\text{current}} \leftarrow k_{\text{next}}$
    return samplingK($k_{\text{current}}$)
else if $k_{\text{next}}$ = new child of $k_{\text{current}}$ then
    add node $k_{\text{next}}$ into tree
    $k_{\text{current}} \leftarrow k_{\text{next}}$
    return samplingK($k_{\text{current}}$)
end if
```

its descendants.

$$p(k_{jt} = k | t, k^{-jt}, k_{\text{current}})$$

$$\propto \begin{cases} m_{.k}^{-jt} \times p(\mathbf{x}_{jt} | \mathbf{x}^{-jt}, \mathbf{t}, \mathbf{k}) \text{ if } k = k_{\text{current}} \\ M_{.k}^{-jt} \times p(\mathbf{x}_{jt} | \mathbf{x}^{-jt}, \mathbf{t}, \mathbf{k}) \text{ if } k = \text{a child of } k_{\text{current}} \\ \gamma^{n} \times p(\mathbf{x}_{jt} | \mathbf{x}^{-jt}, \mathbf{t}, \mathbf{k}) \text{ if } k = \text{a new child of } k_{\text{current}}. \end{cases}$$

Sampling k_{jt} is important as it potentially changes the membership of all data sitting at table t and leads to a well-mixed MCMC.

Estimating ϕ For the rest of this paper and the experiments, we estimate $\hat{\Phi}$ with a Maximum a posteriori (MAP) estimator.

5. EXPERIMENTS

We fit our rCRP model to discover and analyze the hierarchical topic structures in both synthetic and real data sets. We chose not only text data but also user-movie ratings data to show the generality of our model with respect to the type of data.

5.1 Datasets

5.1.1 *Synthetic Data*

We generated a synthetic corpus that consists of 1,000 documents each having 1,000 word tokens. We used a three-level topic tree where the root topic has a uniform distribution over the entire vocabulary. Topics at the second level are distributed over the terms in each of the columns. Topics at the third level have full probability concentrated at a single term from the column of its parent. The topic assignment process is performed by the two-level CRP and rCRP as described in Section 3.

5.1.2 *Real Data*

New York Times The corpus consists of 1.8 million articles published between January 1, 1987 and June 19, 2007 [2]. We randomly sampled 10,000 articles. We removed non-alphabetic characters and single-character words.

MovieLens The MovieLens dataset is a collection of movie ratings from 71,567 users on 10,681 movies. Users rated the movies on a scale of 1 to 5. We turned each user into a document made up of movies that he/she rated as 4 or 5. After this process, the dataset is equivalent to a text corpus consisting of 71,567 documents with 10,681 unique words.

Wikipedia Contemporary Art The WikiArt corpus consists of 3,600 web pages crawled by taking two hops from the Wikipedia Contemporary Art page[3].

In both New York Times and Wikipedia, we applied porter stemming algorithm. We also removed words that occur too infrequently (less than 1%), and too frequently (more than 20%) in terms of the document frequency. These data statistics are illustrated in Table 1.

5.2 Topic Tree Visualization

We visualize the result of inferring topic hierarchy from the synthetic data in Figure 3. The model successfully recovers the original structure and topics. The first and second level topics are almost identical to the original topics. The third level topics show some noise, however most noise words in a topic are accounted by its direct parent topic.

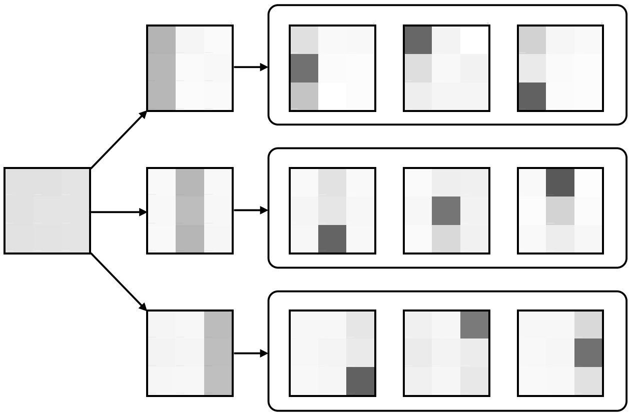


Figure 3: Topic tree inferred from synthetic data. Each cell corresponds to a single word, and is shaded with intensity proportional to the probability of each word in the topic. The first and second level topics are almost identical to the original topics. There exist some noise in the third level topics. However, they are accounted by its direct parent.

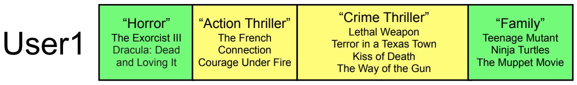


Figure 4: An example user from our MovieLens data. This user watched movies from the topics of Horror and Family from the genre level, and the topics of Action Thriller and Crime Thriller from the sub-genre level.

Figure 5 shows the topic trees inferred from NYTimes, MovieLens, and Wikipedia. Each topic tree is too large to fit in the space provided, so we take a subtree from each tree to illustrate the important points of the discovered topic hierarchies. Each topic is represented by the top ten highest probability words in that topic, and words with probability lower than 0.001 are not shown.

The root topic of each tree contains the most frequently used words in each corpus, and as we move down the tree, topics become more specialized. For example, in the topic tree of the NYTimes dataset, the **Economy** topic is followed by topics about **Technology**, **Stocks**, **Prices**, and **Labor**. For the MovieLens dataset, the root topic represents the generally popular movies, the movies with the most number of high ratings such as *Star Wars* and *Forest Gump*. One level down from the root, we can see the movies being clustered into genres such as **Family, Horror, and Classics**, and the **Horror** movie topic is separated into the more typical **Horror** movies and movies in the **Zombie** sub-genre. As with any hierarchical taxonomy, some parts of the structure are arguable, for example, whether the **Drama** topic should be a subtopic of **Family** or vice versa. Such arguable anomalies in the topic trees discovered by our model reflect the unique characteristics of the data.

[2] http://archive.ics.uci.edu/ml/

[3] en.wikipedia.org/wiki/Painting#Contemporary_art

Table 1: Data statistics

	Documents	Unique Terms	Ave Doc Length
Synthetic	1,000	9	1,000
New York Times	10,000	6,841	1,886
MovieLens	71,567	10,681	56
Wikipedia	3,600	6,386	445

The rCRP model allows each document to have a topic distribution over the entire topic tree. For example, using the topic tree in Figure 5(a), a user with high ratings for “Toy Story” and “Hellraiser” can be interpreted to be interested in both the **Disney** topic and the **Horror** topic. We show, in figure 4, an example user from our data whose topic proportion includes **Horror, Family, Action Thriller**, and **Crime Thriller**, with the movie titles that belong to those topics. Although the vanilla LDA and HDP models allow this, such flexible assignment of document-topic assignments is rare in models of topic hierarchies, and this shows that our model better reflects the nature of the hierarchical topic structure than previously proposed models.

5.3 Heldout likelihood

Heldout likelihood, widely used as a comparative evaluation metric in topic modeling (cf. [5]), evaluates how well the trained model explains the heldout data. Heldout likelihood is defined as log-likelihood of the heldout data given the trained model. Formally,

$$L = \log p(W_{\text{heldout}} | M_{\text{trained}})$$

where W_{heldout} is the heldout data and M_{trained} is the trained model. We use ten-fold cross validation.

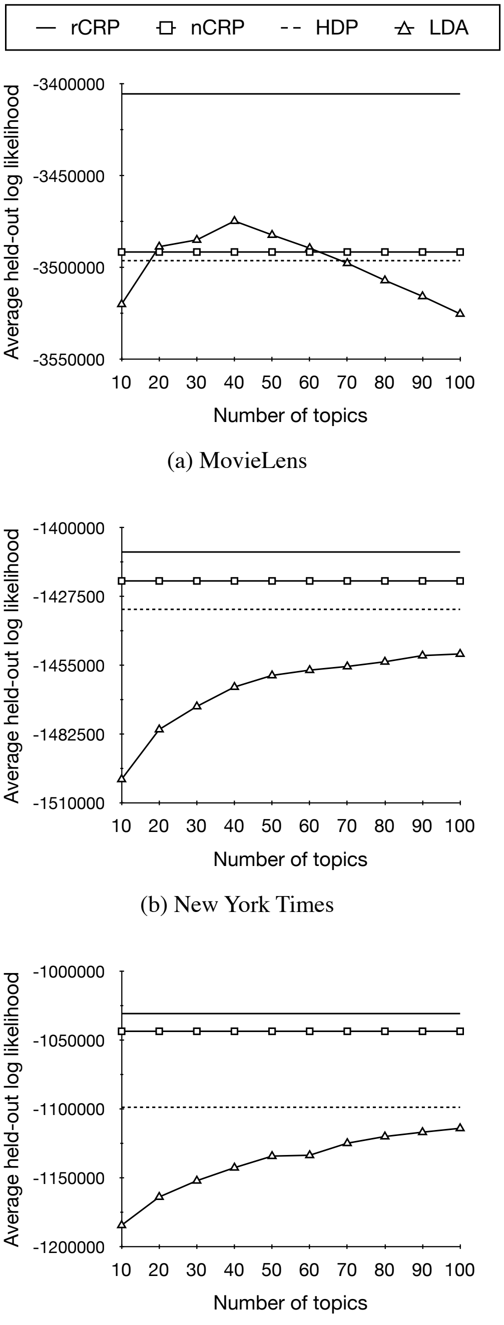
We compare heldout likelihoods of our model with the baselines of LDA, HDP, and nCRP [4]. The result is shown in Figure 6. A model with higher explanatory power produces higher heldout likelihood. Note that the result of LDA with the optimal number of topics is very close to that of the HDP, which is natural because the HDP is designed to find the optimal number of topics for LDA. Figure 6 shows that rCRP model outperforms LDA, HDP, and nCRP which confirms the intuition that the flexible hierarchical topic structure of rCRP explains the data better than the topic structures of HDP and nCRP.

6. HIERARCHY ANALYSIS

A topic model is commonly evaluated by either directly calculating the perplexity or likelihood of held-out data [22], or applying the result to related tasks such as document classification or recommendation [5]. To our knowledge, however, there is no commonly used evaluation metric for measuring the goodness of a topic hierarchy. We suggest two fundamental characteristics of a topic hierarchy and propose concrete evaluation metrics. We then use these metrics to quantitatively compare the characteristics of the topic trees discovered by our model and by nCRP.

6.1 Topic Specialization

Studies on human semantic processing [6] find that in concept trees, the most general semantic category is placed at the top of the tree, and more specific categories toward



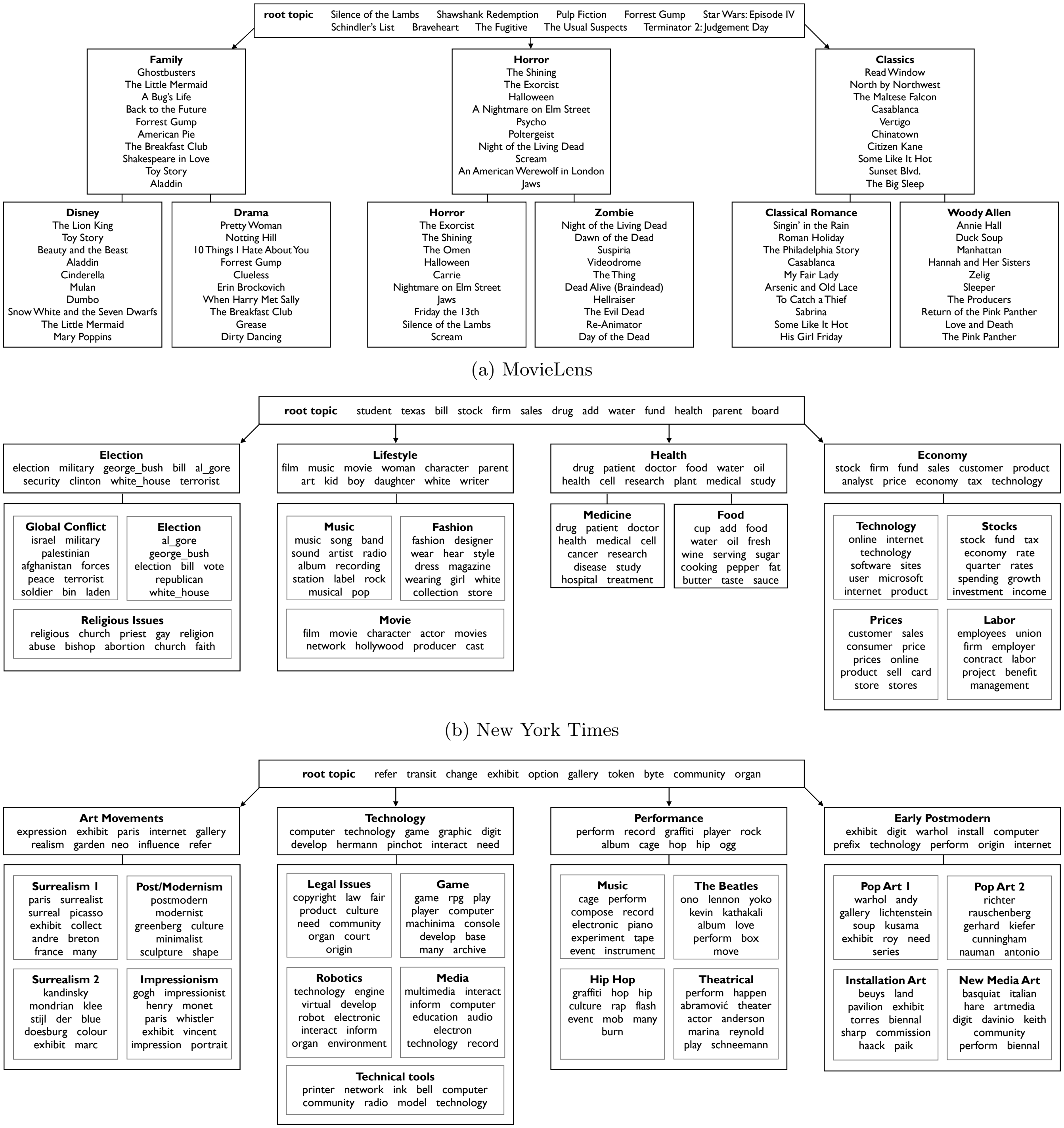
(a) MovieLens

(b) New York Times

(c) Wikipedia

Figure 6: Heldout likelihoods of LDA, HDP, and rCRP for the three datasets. A higher value indicates that the model can explain better the heldout data.

[4] http://www.cs.princeton.edu/ blei/downloads/hlda-c.tgz


root topic Silence of the Lambs Shawshank Redemption Pulp Fiction Forrest Gump Star Wars: Episode IV Schindler's List Braveheart The Fugitive The Usual Suspects Terminator 2: Judgement Day
Family Ghostbusters The Little Mermaid A Bug's Life Back to the Future Forrest Gump American Pie The Breakfast Club Shakespeare in Love Toy Story Aladdin
Horror The Shining The Exorcist Halloween A Nightmare on Elm Street Psycho Poltergeist Night of the Living Dead Scream An American Werewolf in London Jaws
Classics Read Window North by Northwest The Maltese Falcon Casablanca Vertigo Chinatown Citizen Kane Some Like It Hot Sunset Blvd. The Big Sleep
Disney The Lion King Toy Story Beauty and the Beast Aladdin Cinderella Mulan Dumbo Snow White and the Seven Dwarfs The Little Mermaid Mary Poppins
Drama Pretty Woman Notting Hill 10 Things I Hate About You Forrest Gump Clueless Erin Brockovich When Harry Met Sally The Breakfast Club Grease Dirty Dancing
Horror The Exorcist The Shining The Omen Halloween Carrie Nightmare on Elm Street Jaws Friday the 13th Silence of the Lambs Scream
Zombie Night of the Living Dead Dawn of the Dead Suspiria Videodrome The Thing Dead Alive (Braindead) Hellraiser The Evil Dead Re-Animator Day of the Dead
Classical Romance Singin' in the Rain Roman Holiday The Philadelphia Story Casablanca My Fair Lady Arsenic and Old Lace To Catch a Thief Sabrina Some Like It Hot His Girl Friday
Woody Allen Annie Hall Duck Soup Manhattan Hannah and Her Sisters Zelig Sleeper The Producers Return of the Pink Panther Love and Death The Pink Panther
root topic student texas bill stock firm sales drug add water fund health parent board
Election election military george_bush bill al_gore security clinton white_house terrorist
Lifestyle film music movie woman character parent art kid boy daughter white writer
Health drug patient doctor food water oil health cell research plant medical study
Economy stock firm fund sales customer product analyst price economy tax technology
Global Conflict israel military palestinian afghanistan forces peace terrorist soldier bin laden
Election al_gore george_bush election bill vote republican white_house
Religious Issues religious church priest gay religion abuse bishop abortion church faith
Music music song band sound artist radio album recording station label rock musical pop
Fashion fashion designer wear hear style dress magazine wearing girl white collection store
Movie film movie character actor movies network hollywood producer cast
Medicine drug patient doctor health medical cell cancer research disease study hospital treatment
Food cup add food water oil fresh wine serving sugar cooking pepper fat butter taste sauce
Technology online internet technology software sites user microsoft internet product
Stocks stock fund tax economy rate quarter rates spending growth investment income
Prices customer sales consumer price prices online product sell card store stores
Labor employees union firm employer contract labor project benefit management
root topic refer transit change exhibit option gallery token byte community organ
Art Movements expression exhibit paris internet gallery realism garden neo influence refer
Technology computer technology game graphic digit develop hermann pinchot interact need
Performance perform record graffiti player rock album cage hop hip ogg
Early Postmodern exhibit digit warhol install computer prefix technology perform origin internet
Surrealism 1 paris surrealist surreal picasso exhibit collect andre breton france many
Post/Modernism postmodern modernist greenberg culture minimalist sculpture shape
Surrealism 2 kandinsky mondrian klee stijl der blue doesburg colour exhibit marc
Impressionism gogh impressionist henry monet paris whistler exhibit vincent impression portrait
Legal Issues copyright law fair product culture need community organ court origin
Game game rpg play player computer machinima console develop base many archive
Robotics technology engine virtual develop robot electronic interact inform organ environment
Media multimedia interact inform computer education audio electron technology record
Technical tools printer network ink bell computer community radio model technology
Music cage perform compose record electronic piano experiment tape event instrument
The Beatles ono lennon yoko kevin kathakali album love perform box move
Hip Hop graffiti hop hip culture rap flash event mob many burn
Theatrical perform happen abramović theater actor anderson marina reynold play schneemann
Pop Art 1 warhol andy gallery lichtenstein soup kusama exhibit roy need series
Pop Art 2 richter rauschenberg gerhard kiefer cunningham nauman antonio
Installation Art beuys land pavilion exhibit torres biennal sharp commission haack paik
New Media Art basquiat italian hare artmedia digit davinio keith community perform biennal

(a) MovieLens

(b) New York Times

(c) Wikipedia

Figure 5: Topic trees inferred from each dataset. The bold labels at the top of the topics are manually chosen for better readability.

the leaves. We assume the hierarchy of topics should follow this general principle and propose a metric to quantify the general-to-specific characteristic. We name this *topic specialization* and compute it by the semantic distance of a topic from the norm as defined below.

Let ϕ_{Norm} be the norm topic such that the probability of generating a particular word x_i is proportional to the frequency of x_i in the entire corpus. Formally, let $freq(x_i)$ be the frequency of word x_i in the entire corpus, V be the set of entire vocabulary, and β be the smoothing factor. ϕ_{Norm} is a topic such that for each word x_i,

$$p(x_i|\phi_{\text{Norm}}) = \frac{freq(x_i) + \beta}{\sum_{j \in V} freq(x_j) + \beta|V|}.$$

As ϕ_{Norm} represents the word distribution of the entire corpus, we consider it to be the most general topic. For each topic ϕ_k, we measure how much it has drifted away from ϕ_{Norm} by measuring the cosine distance between the two. Formally, let $\Delta(\phi_k)$ be the topic specialization of topic ϕ_k then

$$\Delta(\phi_k) = 1 - \frac{\phi_k \cdot \phi_{\text{Norm}}}{|\phi_k||\phi_{\text{Norm}}|}$$

In rCRP, since customers at each table always visit the root topic first, the semantic distance between ϕ_{Norm} and ϕ_{Root} is zero.

We calculate and average the topic specialization of all topics at each level. From the definition of Δ, a higher value indicates that the topic has drifted farther away from ϕ_{Norm}, which implies that the topic has become more specialized. Figure 7 illustrates the concept of this topic specialization score, where we can see that the topic-word multinomial is near uniform for the root topic and becomes increasingly sparse toward the leaf topic.

In Figure 8, we summarize the topic specialization scores of rCRP and nCRP. In rCRP, topics at the second, third and fourth levels become increasingly more specialized. In nCRP, the general trend is the same, but the pattern is not so pronounced as the topics at all levels appear to be quite specialized. We conjecture this is because nCRP assumes that a document is generated only by the topics in a single path of the hierarchy, so all of the topics must be more specialized to explain the data well. On the other hand, the topics discovered by rCRP do not have that restriction, so the model can focus more on finding an appropriate hierarchy of increasingly more specialized topics, as shown in the topic trees in Figure 5.

6.2 Hierarchical Affinity

Another important characteristic we expect to find from a hierarchical structure of topics is *hierarchical affinity.* That is, topics that descend from ϕ_k must be more similar to ϕ_k than topics that descend from other topics. For clarity, we only use topics at the second level as the parent topics, and topics at the third level as the children topics, and compute the hierarchical affinity among them. Figure 9 illustrates this concept of hierarchical affinity, where the topic-word multinomial for the **Lifestyle** topic is similar to its children topics **Music** and **Movie** but quite different from its non-children topic **Stocks**.

Let ϕ_k be a topic at the second level, and let $\lambda(k)$ be the index set of all topics that have ϕ_k as a direct parent. Also let $\bar{\lambda}(k)$ be the index set of all topics at third level that do not have ϕ_k as a direct parent. In other words, $\lambda(k)$ are children of ϕ_k and $\bar{\lambda}(k)$ are non-children of ϕ_k. To measure the hierarchical affinity, we compare the average cosine similarity between ϕ_k and all topics in $\lambda(k)$ against the average cosine similarity between ϕ_k and all topics in $\bar{\lambda}(k)$.

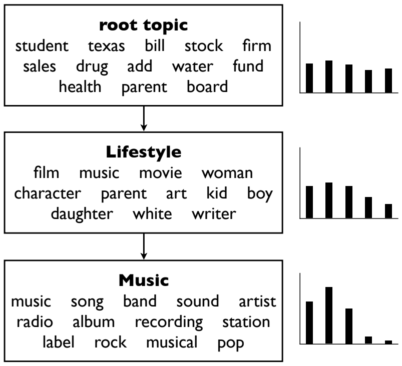
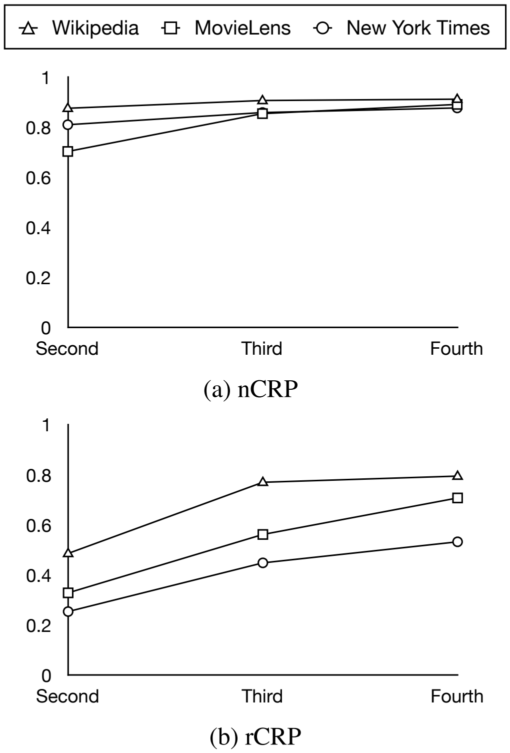

root topic
student texas bill stock firm
sales drug add water fund
health parent board
Lifestyle
film music movie woman
character parent art kid boy
daughter white writer
Music
music song band sound artist
radio album recording station
label rock musical pop

Figure 7: Topic specialization. The topic-word multinomial is near uniform for the root topic and becomes increasingly sparse toward the leaf topic.

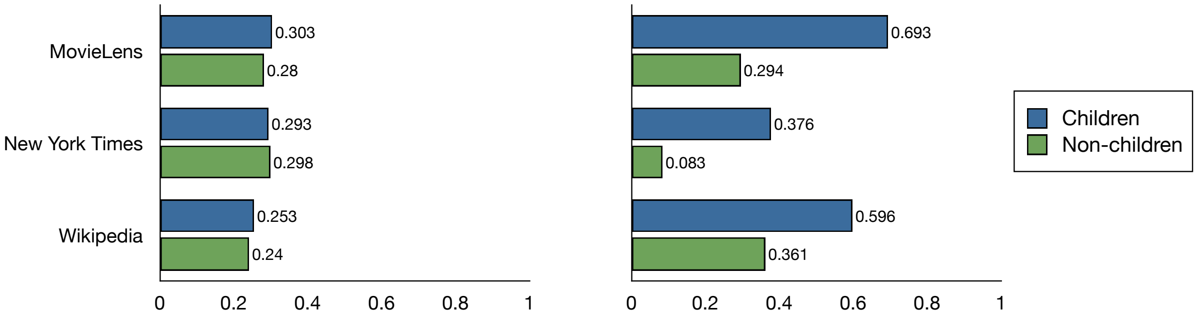


(a) nCRP

(b) rCRP

Figure 8: Topic specialization scores of rCRP and nCRP. This shows the characteristic of rCRP to find general topics at the root and increasingly more specialized topics toward the leaves.

We compute this hierarchical affinity for all topics at the second level and compare the results for rCRP and nCRP. The results are illustrated in Figure 10. For all three data sets, rCRP clearly shows stronger hierarchical affinity for



(a) nCRP

(b) rCRP

Figure 10: Hierarchical Affinity. The average cosine similarity between topics at second and their direct children is compared against their non-children topics. A higher affinity score means that the topics are more similar. For all three data sets, rCRP shows stronger hierarchical affinity for children topics compared to the non-children topics, but nCRP shows similar affinity for children and non-children.

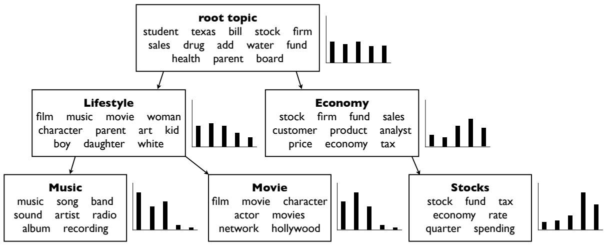


Figure 9: Hierarchical Affinity. Topics that form parent-child relationship show greater similarity in their word distribution than topics that are distant in the topic tree.

children topics compared to the *non-children* topics. However, in nCRP the affinity scores are not different for *children* compared to *non-children*.

7. DISCUSSION

We developed the recursive Chinese Restaurant Process, a new nonparametric prior that captures the hierarchical nature of mixture components. We used the rCRP to construct a nonparametric topic model that infers the hierarchical structure of topics from discrete data. We applied our model to two text corpora and a user ratings dataset and visualized the inferred topic trees to show how our model can find intuitive hierarchical topic structures. We identified topic specialization and hierarchical affinity as two important characteristics of a hierarchical topic structure, and we suggested and tested evaluation metrics to quantify them. We also showed that our model outperformed LDA, HDP, and nCRP in terms of heldout likelihood.

Our model for discovering topic hierarchies with the recursive Chinese restaurant process describes a natural procedure of finding a mixture component in a tree-structured way. This intuitive representation facilitates further extensions to our proposed model. One can relax the assumption that each table is assigned a single dish, and devise the Indian Buffet Process [8] in with topics are in a hierarchical structure. The proposed model can also be deployed in various applications that rely on the latent structure of general-to-specific themes. An example would be recommendations based on collaborative filtering, extending our results with the movie ratings data, or social network search based on topics [19].

8. ACKNOWLEDGEMENTS

This research was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education, Science and Technology (2011-0026507).

9. REFERENCES

[1] R. Adams, Z. Ghahramani, and M. Jordan. Tree-structured stick breaking for hierarchical data. *Advances in Neural Information Processing (NIPS)*, 23, 2010.

[2] 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, 2003.

[3] D. Blei and M. Jordan. Variational inference for dirichlet process mixtures. *Bayesian Analysis*, 1(1):121–144, 2006.

[4] D. Blei and J. Lafferty. Correlated topic models. In *Advances in Neural Information Processing Systems*, 2006.

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

[6] A. Collins and E. Loftus. A spreading-activation theory of semantic processing. *Psychological Review*, 82(6):407, 1975.

[7] M. Escobar and M. West. Bayesian density estimation and inference using mixtures. *Journal of the American Statistical Association*, pages 577–588, 1995.

[8] T. Griffiths and Z. Ghahramani. Infinite latent feature models and the indian buffet process. *Advances in Neural Information Processing Systems*, 18:475, 2006.

[9] K. Heller and Z. Ghahramani. Bayesian hierarchical clustering. In *Proceedings of the 22nd international conference on Machine learning*, pages 297–304, 2005.

[10] H. Ishwaran and L. James. Approximate dirichlet process computing in finite normal mixtures. *Journal of Computational and Graphical Statistics*, 11(3):508–532, 2002.

[11] M. Jeong and I. Titov. Multi-document topic segmentation. In *Proceedings of the 19th ACM International Conference on Information and Knowledge Management*, 2010.

[12] N. Kawamae. Latent interest-topic model: finding the causal relationships behind dyadic data. In *Proceedings of the 19th ACM International Conference on Information and Knowledge Management*, 2010.

[13] D. Kim and A. Oh. Accounting for data dependencies within a hierarchical dirichlet process mixture model. In *the Proceedings of the 20th ACM Conference on Information and Knowledge Managment*, 2011.

[14] 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, 2006.

[15] S. MacEachern. Estimating normal means with a conjugate style dirichlet process prior. *Communications in Statistics-Simulation and Computation*, 23(3):727–741, 1994.

[16] D. Mimno, W. Li, and A. McCallum. Mixtures of hierarchical topics with pachinko allocation. In *Proceedings of International Conference on Machine Learning*, 2007.

[17] R. Neal. Density modeling and clustering using dirichlet diffusion trees. *Bayesian Statistics*, 7:619–629, 2003.

[18] T. Rubin and M. Steyvers. A topic model for movie choices and ratings. In *Proceedings of International Conference on Cognitive Model*, 2009.

[19] J. Tang, S. Wu, B. Gao, and Y. Wan. Topic-level social network search. In *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 769–772, 2011.

[20] Y. Teh, M. Jordan, M. Beal, and D. Blei. Hierarchical dirichlet processes. *Journal of the American Statistical Association*, Jan 2006.

[21] Y. Teh, K. Kurihara, and M. Welling. Collapsed variational inference for hdp. *Advances in Neural Information Processing Systems*, 20:1481–1488, 2008.

[22] H. M. Wallach, I. Murray, R. Salakhutdinov, and D. Mimno. Evaluation methods for topic models. *Proceedings of the 26th International Conference on Machine Learning*, 2009.

[23] C. Williams. A mcmc approach to hierarchical mixture modelling. *Advances in Neural Information Processing Systems*, 12:680–686, 2000.

[24] G. Zheng, J. Guo, L. Yang, S. Xu, S. Bao, Z. Su, D. Han, and Y. Yu. A topical link model for community discovery in textual interaction graph. In *Proceedings of the 19th ACM International Conference on Information and Knowledge Management*, 2010.