HDPauthor: A New Hybrid Author-Topic Model using Latent Dirichlet Allocation and Hierarchical Dirichlet Processes

Ming Yang
Computing and Information Sciences
Kansas State University
yangming@ksu.edu

William H. Hsu
Computing and Information Sciences
Kansas State University
bhsu@ksu.edu

ABSTRACT
We present a new approach towards capturing topic interests corresponding to all the observed latent topics generated by an author in documents to which he or she has contributed. Topic models based on Latent Dirichlet Allocation (LDA) have been built for this purpose but are brittle as to the number of topics allowed for a collection and for each author of documents within the collection. Meanwhile, topic models based upon Hierarchical Dirichlet Processes (HDPs) allow an arbitrary number of topics to be discovered and generative distributions of interest inferred from text corpora, but this approach is not directly extensible to generative models of authors as contributors to documents with variable topical expertise. Our approach combines an existing HDP framework for learning topics from free text with latent authorship learning within a generative model using author list information. This model adds another layer into the current hierarchy of HDPs to represent topic groups shared by authors, and the document topic distribution is represented as a mixture of topic distribution of its authors. Our model automatically learns author contribution partitions for documents in addition to topics.

Keywords
Topic Modeling, Hierarchical Dirichlet Process

1. INTRODUCTION
While topic modeling has long been used to characterize the topic distributions of documents, there is also a growing need for learning the topic interests of authors in order to model their expertise, scope as collaborators and readers, and general roles as generators of documents. Moreover, the contribution of different authors to a single document is also a learning problem that needs to be studied. We seek to develop a generative mixture model capable of capturing these facets of text document collections, by extending current topic models to simultaneously learn and identify: topic interests of authors, topic distribution in documents, and author contributions to documents.

In real-world applications, the number of global topics across whole corpora may not be fixed or boundable. However, each author usually only works on and is good at a small set of topics, and each document written by a group of authors is also usually written about a small set of topics. The nonparametric Bayesian features of HDP for topic modeling can help us to solve this problem and derive a better learning algorithm compared to existing LDA-based author-topic learning models.

In this paper we present a statistical generative mixture model called HDPauthor for scientific articles with authors, which extends the existing HDP model with authorship information. It retains benefits of traditional HDP models in that the global number of topics is unbounded. Each author of one or more documents in a text collection also shares an unbounded number of topics from the global topic pool.

2. RELATED WORK
Several research advances have already incorporated co-authorship into topic modeling. One such significant development is the Author-Topic model [11] [10]. This model extends LDA to include authorship information. It makes it possible to simultaneously learn both the relevance of different global topics in document, and the interests of topics for authors. In similar fashion to the LDA model, the total number of topics for the whole corpus must be predetermined in advance, with no flexibility over the number of topics generated. This model also learns distribution of each topic in large global group of topics for each document and each author.

Models proposed by Dai [3] [4] are based on a nonparametric HDP model for the topic-author problem. This group defines a Dirichlet process (DP) over author entities and topics, which in turn is then drawn from a global author and topic DP. This model is mainly geared towards disambiguation of author entities. However, this model combines authors and topics in the same DP, which fails to decouple topics from authors. Therefore, it lacks the ability to share the same topics between different authors, and also makes it difficult to infer author contributions to these documents.

3. MODEL INTRODUCTION
Our HDPauthor model is a nonparametric Bayesian hierarchical model for author-topic generation. In this model we assume that each token in the document is written by one
and only one of the authors in the author list of this document, associated with the topic distribution of this author.

By using an HDP framework, we also assume that each author is associated with a topic distribution which is drawn based on a global topic distribution in whole corpora, with different variability. The global topic atoms are shared by all authors, but each author only occupies a small subset of these global topic components, with different stick-breaking weights. This local probability measure of each author represents the topic interests of this author.

The topic distribution of each document is not drawn from the global topic distribution directly, but represented by this mixture model of all its authors indirectly. Therefore, each document is represented by a union of all topics contributed by each of its authors.

4. MODEL DEFINITION

The document representation in our model also follows our definition stated in HDPsent [17][16]. We assume \( D = \{d_1, d_2, \ldots \} \) is a collection of scientific articles, composed of a series of words from vocabulary \( V \) as \( x_j = \{x_{j1}, x_{j2}, \ldots \} \).

We assume that each document has a set of authors \( a_j = \{a_{j1}, a_{j2}, \ldots \} \) who cooperated in writing this document \( d_j \).

Here we associate one latent author label \( q \) from the author set \( a_j \) for each token in document \( d_j \) along with original latent topic label \( k \).

We generate \( G_0 \) as the corpus-level set of topics as a Dirichlet Process with \( H \) as base measure and \( \gamma \) as its concentration parameter. The topic components are denoted as \( \phi_\gamma \). Each author \( a \) that exists in the entire corpus corresponds to a Dirichlet process \( G_a \) that shares the same global base distribution of topics \( G_0 \), with concentration parameter \( \eta \).

\[
G_0|\gamma, H \sim DP(\gamma, H) \\
G_a|\eta, G_0 \sim DP(\eta, G_0) 
\]

(1)

Unlike in the traditional HDP model, we set up a mixture of components from probability measures of all authors of each document. We then denote the mixing proportion vector as \( \pi \sim \{\pi_j, \ldots, \pi_j|a_j\} \). Since each document is written by a fixed group of authors, we can just assume that \( \pi_j \) is drawn from a symmetric Dirichlet distribution with concentration parameter \( \epsilon \).

\[
\pi_j \sim Dir(\epsilon) 
\]

(2)

For a mixing proportion vector \( \pi_j \), there are two ways of drawing \( G_j \) from a Dirichlet process for the mixture of the probability measures of all its authors, designated \( \{G_a|a \in a_j\} \). The first method is to combine the probability measures \( G_a \) of authors as a new base measure first, then draw a DP with this base measure for document \( d_j \). We call this HDPAuthor mixture model (1), which can be denoted as:

\[
G_j \sim DP(\alpha_0, \sum_{a \in a_j} \pi_j a \cdot G_a) 
\]

(3)

Another method is to first draw separate DPs from each of the authors of the document \( d_j \) with the author’s own probability measure \( G_a \) as the base measure, and then calculate the probability measure of \( d_j \) as a mixture of these DPs. We call this HDPAuthor mixture model (2), and the mathematical formula for this method can be denoted as:

\[
G_j \sim \sum_{a \in a_j} \pi_j a \cdot DP(\alpha_0, G_a) 
\]

(4)

Each observation \( x_{ji} \) in document \( d_j \) is associated with a combination of two parameters \( <a_{ji}, \theta_{ji}, \gamma> \) sampled from this mixture \( G_j \). In this combination, \( \theta_{ji} \) is author label, \( \gamma \) is the parameter specifying the one of the author’s topic component for \( x_{ji} \). Therefore, this \( \theta_{ji} \) is associated with table \( t_{ji} \), which is an instance of mixture component \( \omega_{ah} \) from author \( a = a_{ji}; \omega_{ah} \) is then associated with one global topic component \( g \). Given global topic component \( g \), the token \( x_{ji} \) arises from a Dirichlet distribution over the whole vocabulary based on this topic label \( g \):

\[
< a_{ji}, \theta_{ji}, \gamma > | G_j \sim G_j \\
x_{ji} | \theta_{ji} \sim F(\theta_{ji}) 
\]

(5)

Here we can simply use \( \phi_h \) to denote the word distribution for topic \( g \). Therefore, the conditional density of each observation \( x_{ji} \) under this particular \( \phi_h \) given all other observations can be derived similarly to [15] equation (30):

\[
\int f(x_{ji}|\phi_h) \prod_{i' \neq j} f(x_{i'|c} | \phi_g) h(\phi_h) d\phi_g \\
f^g_{x_{ji}}(x_{ji}) = \frac{1}{\prod_{i' \neq j} f(x_{i'|c} | \phi_g) h(\phi_g) d\phi_g} 
\]

(6)

Furthermore, the conditional probability of data item \( x_{ji} \) being assigned to a new topic \( g^{new} \) is also only dependent on the conjugate prior \( H \). This can be represented as follows:

\[
f_{x_{ji}}^{g^{new}}(x_{ji}) = \int f(x_{ji}|\phi_g) h(\phi_g) d\phi_g 
\]

(7)

Here in Figure 1 we illustrate the graphical plate model for our HDPAuthor model with one more layer of author probability measures injected into original HDP model:

Figure 1: Plate Model for HDP model with authors

5. INFERENCES

Our model is based on a Gibbs sampling-based implementation of the Chinese restaurant franchise process (CRFP).

Inference for mixture model (1)

Here we compute the marginal of \( G_j \) under this author mixture Dirichlet process model with \( G_0 \) and \( G_a \) integrated out. We want to compute the conditional distribution of \( \theta_{ji} \) given all other variables, and thus we extend [15] equation (24) to fit our author mixture model (1), to obtain:

\[
\theta_{ji}| \theta_{j1}, \ldots, \theta_{ji-1}, \alpha_0, G_j, G_0, G_a, \ldots \\
\sim \sum_{t=1}^n \frac{n_{ji}}{n_j + \alpha_0} \delta_{\theta_{ji}} + \frac{\alpha_0}{n_j + \alpha_0} \sum_{a \in a_j} \pi_j a \cdot G_a 
\]

(8)

Here \( \psi_{ji} \) represents the table-specific indicator that indicates the component choice \( k_{ji} \) from author \( a_{ji} \)’s probability
measure. A draw from this mixture model can be divided into two parts. If the former summation is chosen, then \( x_{ji} \) would be assigned to an existing \( \psi_{ji} \), and we can denote \( \theta_{ji} = \psi_{ji} \). If the latter summation is chosen, we have to create a new document-specific table \( t^\text{new} \), assign it to one of the authors according to mixing proportion vector of authors for document \( d_j \), where each \( \pi_j \in \Pi_j \) represents the probability that table \( t^\text{new} \) belongs to author \( a \). Then we can draw one new \( \psi_{ji}^\text{new} \) from the probability measure of author \( a \) represented as \( G_a \).

\[ G_a \text{ for each author } a \text{ in corpus appears in all documents in which this author participates. It should be integrated out through all } \psi_{ji} \text{ that } a_{ji} = a. \text{ We use } m_{ak} \text{ to indicate the total number of tables } t \text{ such that } k_{ji} = k \text{ and } a_{ji} = a. \text{ To integrate out each } G_a, \text{ we can get:} \]

\[
\psi_{ji}|\psi_{j1},...,\psi_{j-1},\eta,G_0 \sim \sum_{k=1}^{l_{0}} \frac{m_{ak}}{m_{ak} + \eta} \delta_{\psi_{jk}} + \frac{\eta}{m_{ak} + \eta} G_0 \tag{9}
\]

This mixture is also divided into two parts. If we draw sample \( \psi_{ji} \) from the former part, then assign it to an existing component \( k \) from author \( a \), we can denote it as \( \psi_{ji} = \omega_{ak} \). If the latter part is chosen, we will create one new component \( k^\text{new} \) for author \( a \), and we draw this new \( \omega_{ak} = \psi_{ji}^\text{new} \) from global topic probability measure \( G_0 \).

Finally, we can integrate out this global probability measure \( G_0 \) by all cluster components \( \omega_{ak} \) from all existing authors in whole corpora. We here use \( l_{0} \) to indicate the total number of \( \omega_{ak} \) such that \( g_{ak} = g \). Then the integral can be represented similarly to [15] equation (25):

\[
\omega_{ak}\omega_{1\ldots k-1},\gamma,H \sim \sum_{g=1}^{G} \frac{l_{g}}{l_{1} + \gamma} \delta_{\omega_{g}} + \frac{\gamma}{l_{1} + \gamma} H \tag{10}
\]

Similarly, if the former is chosen, we assign the existing topic component \( \phi_{g} \) to \( \omega_{ak} \); if the latter is chosen, we create a new topic \( \gamma^\text{new} \) sampled from base measure \( H \). Inference for mixture model (2)

For mixture model (2), each document’s probability measure is divided into \( [a_j] \) independent components, where the probability of each component \( a \in [a_j] \) to be chosen is determined by \( \pi_{ja} \in \Pi_j \) from this document-specific mixing proportion vector \( \Pi_j \). Once a specific author \( a \) is chosen, the probability distribution of \( \theta_{ji} \) follows the Dirichlet Process \( DP(\alpha_0,G_0) \) where \( a \in a_j \), using the probability measure of author \( a \) denoted as \( G_a \) to be its base measure. Therefore, with \( G_0 \) and \( G_a \) integrated out, we can obtain the distribution of \( \theta_{ji} \) given all other variables:

\[
\theta_{ji}|\theta_{j1},...,\theta_{j-1},\alpha_0,G_j,G_a,G_{a1},G_{a2},... \sim \sum_{a \in a_j} \pi_{ja} \cdot \left( \sum_{l=1}^{n_{ja}} \frac{n_{jl}}{n_{ja} + \alpha_0} \delta_{\theta_{jl}} + \frac{\alpha_0}{n_{ja} + \alpha_0} G_0 \right) \tag{11}
\]

These two models are only different in how the mixture of authors is constructed with each author’s own probability measure drawn from a shared global infinite topic mixture model in one document. The constructions of each author’s probability measure and global topic measure are same. Therefore, the posterior conditional calculation of \( \psi_{ji} \) and \( \omega_{ak} \) for model (2) are same as model (1).

6. EXPERIMENT

Here we choose two data sets for conducting experiments on our \textsc{HDPauthor} model, both of which are text collections of academic papers.

6.1 NIPS Experiment

The data set we used for this model is \textit{NIPS Conference Papers} \textsuperscript{1} Volume 0-12, provided by Sam Roweis \textsuperscript{2}. We extracted a subset of papers with denser connections between authors, and finally get a dataset with 873 papers, written by 850 authors in total.

Here in Table 1 we demonstrate an example of 4 selected frequent topics with its 10 most likely words and 10 most likely authors listed in a descending order:

<table>
<thead>
<tr>
<th>Topic</th>
<th>Most Likely Words</th>
<th>Most Likely Authors</th>
</tr>
</thead>
<tbody>
<tr>
<td>Topic 1</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>Topic 2</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>

6.2 DBLP Abstract Experiment

We use another citation network data set \textsuperscript{3}, extracted from Digital Bibliography and Library Project (DBLP), ACM Digital Library and other sources, and provided by Arnetminer \textsuperscript{4}. We select only publications in five areas of computer science: {\textit{Machine Learning, Information Retrieval, Artificial Intelligence, Natural Language & Speech, Data Mining}}. We then extract publications from top-ranked conferences retrieved from Microsoft Academic Search \textsuperscript{4} for each area. These publications are labeled by area according to the category of conference in which they were published.

We generated a data set for experiment with abstracts from 3,177 papers as documents, and with a total of 2,428 authors involved. We here represent the perplexity evolution in Figure 2:

We illustrate the table of top words and top authors for these 4 selected topics as example in Table 3:

<table>
<thead>
<tr>
<th>Topic</th>
<th>Authors</th>
</tr>
</thead>
<tbody>
<tr>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>

\textsuperscript{1}http://papers.nips.cc/
\textsuperscript{2}This data set is available at http://www.cs.nyu.edu/~roweis/data.html
\textsuperscript{3}This data set is available at https://aminer.org/billboard/citation
\textsuperscript{4}http://academic.research.microsoft.com/
Table 1: Example of top topics learned from NIPS experiment

<table>
<thead>
<tr>
<th>Topic 1</th>
<th>Word</th>
<th>Prob</th>
<th>Author</th>
<th>Prob</th>
</tr>
</thead>
<tbody>
<tr>
<td>network</td>
<td>0.107</td>
<td></td>
<td>Sejnowski_T</td>
<td>0.056</td>
</tr>
<tr>
<td>input</td>
<td>0.045</td>
<td></td>
<td>Mozer_M</td>
<td>0.035</td>
</tr>
<tr>
<td>neural</td>
<td>0.028</td>
<td></td>
<td>Hinton_G</td>
<td>0.022</td>
</tr>
<tr>
<td>learning</td>
<td>0.028</td>
<td></td>
<td>Bengio_Y</td>
<td>0.022</td>
</tr>
<tr>
<td>unit</td>
<td>0.027</td>
<td></td>
<td>Jordan_M</td>
<td>0.020</td>
</tr>
<tr>
<td>output</td>
<td>0.027</td>
<td></td>
<td>Chen_H</td>
<td>0.016</td>
</tr>
<tr>
<td>weight</td>
<td>0.023</td>
<td></td>
<td>Moody_J</td>
<td>0.016</td>
</tr>
<tr>
<td>training</td>
<td>0.019</td>
<td></td>
<td>Stork_D</td>
<td>0.016</td>
</tr>
<tr>
<td>time</td>
<td>0.014</td>
<td></td>
<td>Munro_P</td>
<td>0.014</td>
</tr>
<tr>
<td>system</td>
<td>0.013</td>
<td></td>
<td>Sun_G</td>
<td>0.013</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Topic 2</th>
<th>Word</th>
<th>Prob</th>
<th>Author</th>
<th>Prob</th>
</tr>
</thead>
<tbody>
<tr>
<td>set</td>
<td>0.015</td>
<td></td>
<td>Sejnowski_T</td>
<td>0.032</td>
</tr>
<tr>
<td>result</td>
<td>0.015</td>
<td></td>
<td>Jordan_M</td>
<td>0.025</td>
</tr>
<tr>
<td>figure</td>
<td>0.014</td>
<td></td>
<td>Hinton_G</td>
<td>0.022</td>
</tr>
<tr>
<td>number</td>
<td>0.013</td>
<td></td>
<td>Koch_C</td>
<td>0.020</td>
</tr>
<tr>
<td>data</td>
<td>0.011</td>
<td></td>
<td>Dayan_P</td>
<td>0.019</td>
</tr>
<tr>
<td>function</td>
<td>0.010</td>
<td></td>
<td>Moody_I</td>
<td>0.015</td>
</tr>
<tr>
<td>based</td>
<td>0.008</td>
<td></td>
<td>Mozer_M</td>
<td>0.014</td>
</tr>
<tr>
<td>model</td>
<td>0.008</td>
<td></td>
<td>Tishby_N</td>
<td>0.014</td>
</tr>
<tr>
<td>method</td>
<td>0.008</td>
<td></td>
<td>Barto_A</td>
<td>0.013</td>
</tr>
<tr>
<td>case</td>
<td>0.008</td>
<td></td>
<td>Viola_P</td>
<td>0.013</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Topic 98</th>
<th>Word</th>
<th>Prob</th>
<th>Author</th>
<th>Prob</th>
</tr>
</thead>
<tbody>
<tr>
<td>image</td>
<td>0.049</td>
<td></td>
<td>Koch_C</td>
<td>0.119</td>
</tr>
<tr>
<td>visual</td>
<td>0.028</td>
<td></td>
<td>Horiuchi_T</td>
<td>0.106</td>
</tr>
<tr>
<td>field</td>
<td>0.023</td>
<td></td>
<td>Ruderman_D</td>
<td>0.088</td>
</tr>
<tr>
<td>system</td>
<td>0.020</td>
<td></td>
<td>Biakle_W</td>
<td>0.068</td>
</tr>
<tr>
<td>pixel</td>
<td>0.017</td>
<td></td>
<td>Dimitrov_A</td>
<td>0.05</td>
</tr>
<tr>
<td>filter</td>
<td>0.015</td>
<td></td>
<td>Bair_W</td>
<td>0.038</td>
</tr>
<tr>
<td>signal</td>
<td>0.013</td>
<td></td>
<td>Indiveri_G</td>
<td>0.035</td>
</tr>
<tr>
<td>object</td>
<td>0.013</td>
<td></td>
<td>Viola_P</td>
<td>0.030</td>
</tr>
<tr>
<td>center</td>
<td>0.012</td>
<td></td>
<td>Zee_A</td>
<td>0.030</td>
</tr>
<tr>
<td>local</td>
<td>0.011</td>
<td></td>
<td>Miyake_S</td>
<td>0.027</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Topic 110</th>
<th>Word</th>
<th>Prob</th>
<th>Author</th>
<th>Prob</th>
</tr>
</thead>
<tbody>
<tr>
<td>word</td>
<td>0.053</td>
<td></td>
<td>Tebelskis_J</td>
<td>0.107</td>
</tr>
<tr>
<td>speech</td>
<td>0.042</td>
<td></td>
<td>Franco_H</td>
<td>0.089</td>
</tr>
<tr>
<td>recognition</td>
<td>0.037</td>
<td></td>
<td>Bourlard_H</td>
<td>0.086</td>
</tr>
<tr>
<td>training</td>
<td>0.025</td>
<td></td>
<td>De-Mori_R</td>
<td>0.084</td>
</tr>
<tr>
<td>frame</td>
<td>0.020</td>
<td></td>
<td>Rahim_M</td>
<td>0.069</td>
</tr>
<tr>
<td>system</td>
<td>0.017</td>
<td></td>
<td>Waibel_A</td>
<td>0.055</td>
</tr>
<tr>
<td>error</td>
<td>0.014</td>
<td></td>
<td>Hild_H</td>
<td>0.043</td>
</tr>
<tr>
<td>hmm</td>
<td>0.013</td>
<td></td>
<td>Chang_E</td>
<td>0.038</td>
</tr>
<tr>
<td>level</td>
<td>0.012</td>
<td></td>
<td>Singer_E</td>
<td>0.036</td>
</tr>
<tr>
<td>output</td>
<td>0.012</td>
<td></td>
<td>Bengio_Y</td>
<td>0.035</td>
</tr>
</tbody>
</table>

Table 2: Example of top topics for selected authors learned from NIPS experiment

\[ p(q, a_q|d_j, a_j) = \omega \cdot p(q|d_j) + (1 - \omega) \cdot \text{similarity}(a_q, a_j) \] (12)

We then calculate cosine similarity[12] as the similarity score for averaged topic distribution for authors from two sides. We use 11-point interpolated average precision[8] for model comparison. Here in Figure 3 we illustrate our performance compared to other models. We set \( \omega = 0.5 \) for Equation 12. We implemented an AT model, and set \( K = 200 \) for this experiment. We use one Python library called Gensim [9] for HDP topic learning.

7. CONCLUSIONS

We have presented an HDP-based hierarchical, nonparametric Bayesian generative model for author-topic hybrid learning, called HDPauthor. This model represents each author as a Dirichlet process of global topics, and each document as a mixture of these Dirichlet processes of its authors. It learns topic interests of authors and the topic distribution of documents as classical topic models, but also learns author contribution for documents in the meantime. It also preserves the benefit of nonparametric Bayesian hierarchical topic model. Our model uses a purely unsupervised learning methodology; it requires neither knowledge about documents nor data about authors.

A key novel contribution of our HDPauthor model is its ability to represent each document, each author, and global topics as Dirichlet processes (DPs) or mixtures of DPs. This eliminates restrictions on the number of topic components that the user must define beforehand for all other LDA-based hybrid models [10]. Thus, the emergence of new topic components and fading out of old topic components can be easily detected and accounted for using our framework.
Several topics we plan to explore in continuing work are as follows:

1. Variational approximate inference [2] [6] can be used for our model. Inference is more difficult to formulate [5], but converges more efficiently. We plan to develop a variational approximation approach that can be used to adapt the HDPauthor and other similar hybrid models.

2. Author disambiguation [13] [3] is also an interesting topic to explore, based on our model.

3. A combination of the HDPauthor model with citation network [1] [14] can help with author and document information retrieval.

9. ACKNOWLEDGEMENTS

This work was made possible through the support of Lawrence Livermore National Labs (LLNL). We thank David Buttler of LLNL for technical discussions and recommendations regarding evaluation methods and use cases.

10. REFERENCES

Figure 3: Precision-Recall curve for document retrieval for DBLP experiment


