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

Willian 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 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 in general 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 would like to develop a generative mixture model extending current topic models, which is capable of simultaneously learning and identifying topic interests of authors, topic distribution in documents, and author contributions to documents.

In real-world applications, the number 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. Therefore, the nonparametric Bayesian feature of HDP for topic modeling can help us to solve the problem, and infer 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 to incorporate authorship information. It benefits from traditional HDP model features 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

There are many works that have already incorporated co-authorship into topic modeling. One significant model is the Author-Topic model [11] [10]. This model extends the LDA model 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

ACM 978-1-4503-4144-8/16/04.
http://dx.doi.org/10.1145/2872518.2890561

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 $D = \{d_1, d_2, ...\}$ is a collection of scientific articles, composed of a series of words from vocabulary V as $x_j = \{x_{j1}, x_{j2}, ...\}$. We assume that each document has a set of authors $a_j = \{a_{j1}, a_{j2}, ...\}$ who cooperated in writing this document d_j. Here we associate one latent author label q from the author set $\boldsymbol{a}_j$ 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 γ as its concentration parameter. The topic components are denoted as ϕ_g. Each author a that exists in whole corpus holds a Dirichlet Process G_a that shares the same global base distribution of topics G_0, with concentration parameter η.

$$\begin{aligned} G_0|\gamma, H &\sim DP(\gamma, H) \\ G_a|\eta, G_0 &\sim DP(\eta, G_0) \end{aligned} \tag{1}$$

Unlike 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 $\boldsymbol{\pi_j} =< \pi_{j1}, ..., \pi_{j|\boldsymbol{a}_j|} >$. Since each document is written by a fixed group of authors, we can here simply assume that $\boldsymbol{\pi_j}$ is drawn from a symmetric Dirichlet distribution with concentration parameter ϵ.

$$\boldsymbol{\pi_j} \sim Dir(\epsilon) \tag{2}$$

For a mixing proportion vector π_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 $\{G_a | a \in \boldsymbol{a}_j\}$. 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:

$$G_j \sim DP(\alpha_0, \sum_{a \in a_j} \pi_{ja} \cdot G_a) \tag{3}$$

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:

$$G_j \sim \sum_{a \in \boldsymbol{a}_j} \pi_{ja} \cdot DP(\alpha_0, G_a) \tag{4}$$

Each observation x_{ji} in document d_j is associated with a combination of two parameters $< a_{ji}, \theta_{ji} >$ sampled from this mixture G_j. In this combination, a_{ji} is author label, θ_{ji} is the parameter specifying the one of the author's topic component for x_{ji}. Therefore, this θ_{ji} is associated with table t_{ji}, which is an instance of mixture component ω_{ak} from author $a = a_{ji}$; ω_{ak} 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:

$$\begin{aligned} < a_{ji}, \theta_{ji} > | G_j &\sim G_j \\ x_{ji}|\theta_{ji} &\sim F(\theta_{ji}) \end{aligned} \tag{5}$$

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

$$f_g^{-xji}(x_{ji}) = \frac{\int f(x_{ji}|\phi_g) \prod_{\substack{j'i' \neq ji, \\ \theta_{j'i'}=g}} f(x_{j'i'}|\phi_g) h(\phi_g) d\phi_g}{\int \prod_{\substack{j'i' \neq ji, \\ \theta_{j'i'}=g}} f(x_{j'i'}|\phi_g) h(\phi_g) d\phi_g} \tag{6}$$

And 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:

$$f_{g^{new}}^{-xji}(x_{ji}) = \int f(x_{ji}|\phi_g) h(\phi_g) d\phi_g \tag{7}$$

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:

5. INFERENCE

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 are integrated out. We want to compute the conditional distribution of θ_{ji} given all other variables, we extend [15] equation (24) to fit our author mixture model (1), we can obtain:

$$\begin{aligned} &\theta_{ji}|\theta_{j1}, ..., \theta_{ji-1}, \alpha_0, G_j, G_{a0}, G_{a1}, ... \\ &\quad \sim \sum_{t=1}^{m_{j\cdot}} \frac{n_{jt}}{n_{j\cdot}^{-ji} + \alpha_0} \delta_{\psi_{jt}} + \frac{\alpha_0}{n_{j\cdot}^{-ji} + \alpha_0} \sum_{a \in \boldsymbol{a}_j} \pi_{ja} \cdot G_a \end{aligned} \tag{8}$$

Here ψ_{jt} represents the table-specific indicator that indicates the component choice k_{jt} from author a_{jt}'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 ψ_{jt}, and we can denote $\theta_{ji} = \psi_{jt}$. If the latter summation is chosen, we have to

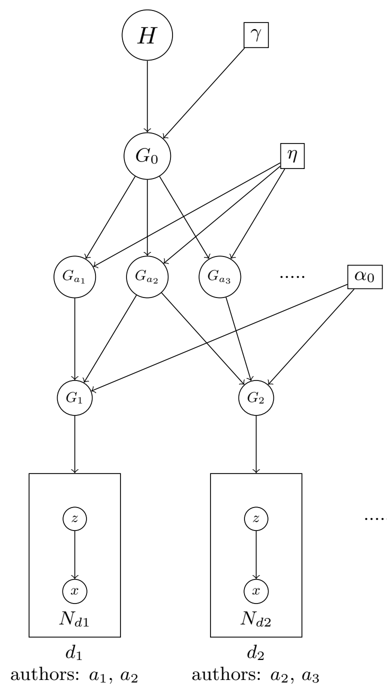

Figure 1: Plate Model for HDP model with authors

create a new document-specific table t^{new}, assign it to one of the authors according to mixing proportion vector of authors for document d_j, where each $\pi_{ja} \in \boldsymbol{\pi_j}$ represents the probability that table t^{new} belongs to author a. Then we can draw one new $\psi_{jt^{new}}$ from the probability measure of author a represented as G_a.

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

$$\psi_{jt}|\psi_{11}, ..., \psi_{jt-1}, \eta, G_0 \sim \sum_{k=1}^{l_{a\cdot\cdot}} \frac{m_{ak}}{m_{a\cdot\cdot} + \eta} \delta_{\omega_{ak}} + \frac{\eta}{m_{a\cdot\cdot} + \eta} G_0 \tag{9}$$

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

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

$$\omega_{ak}|\omega_{11}, ..., \omega_{ak-1}, \gamma, H \sim \sum_{g=1}^{G} \frac{l_{g\cdot}}{l_{\cdot\cdot} + \gamma} \delta_{\phi_g} + \frac{\gamma}{l_{\cdot\cdot} + \gamma} H \tag{10}$$

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

For mixture model (2), each document's probability measure is divided into $|\boldsymbol{a}_j|$ independent components, where the probability of each component $a \in \boldsymbol{a}_j$ to be chosen is determined by $\pi_{ja} \in \boldsymbol{\pi}_j$ from this document-specific mixing proportion vector $\boldsymbol{\pi}_j$. Once a specific author a is chosen, the probability distribution of θ_{ji} follows the Dirichlet Process $DP(\alpha_0, G_a)$ where $a \in \boldsymbol{a}_j$, 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 θ_{ji} given all other variables:

$$\theta_{ji}|\theta_{j1}, ..., \theta_{ji-1}, \alpha_0, G_j, G_{a1}, G_{a2}, ... \sim \sum_{a \in \boldsymbol{a}_j} \pi_{ja} \cdot \Big(\sum_{t=1}^{m_{ja\cdot}} \frac{n_{jt}}{n_{ja\cdot}^{-ji} + \alpha_0} \delta_{\psi_{jt}} + \frac{\alpha_0}{n_{ja\cdot}^{-ji} + \alpha_0} G_a \Big) \tag{11}$$

These two models are only different in constructing the mixture of authors with each author's own probability measure drawn from 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 ψ_{jt} and ω_{ak} for model (2) are same as model (1).

6. EXPERIMENT

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

6.1 *NIPS* Experiment

The data set we are going to use for this model is *NIPS Conference Papers*[1] Volume 0-12, provided by Sam Roweis [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:

Topic 1 and Topic 2 are general topics commonly exists in almost all the documents across the whole data set, and shared by almost all authors. Our `HDPauthor` model is able to discover a variety of more specific research areas in neuroscience. Here we also select some famous authors and list 3 most likely topics for each of them, other than Topic 1 and Topic 2, represented in Table 2:

6.2 *DBLP* abstract Experiment

We use another citation network data set [3], extracted from Digital Bibliography and Library Project (DBLP), ACM Digital Library and other sources, and provided by Arnetminer [14]. We select only publications in 5 areas in computer science category as {*Machine Learning, Information Retrieval, Artificial Intelligence, Natural Language & Speech, Data Mining*}. We then extract publications from top ranked

[1] http://papers.nips.cc/

[2] This data set is available at http://www.cs.nyu.edu/~roweis/data.html

[3] This data set is available at https://aminer.org/billboard/citation

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

Topic 2			
Word	Prob	Author	Prob
set	0.015	Sejnowski_T	0.032
result	0.015	Jordan_M	0.025
figure	0.014	Hinton_G	0.022
number	0.013	Koch_C	0.020
data	0.011	Dayan_P	0.019
function	0.010	Moody_J	0.015
based	0.008	Mozer_M	0.014
model	0.008	Tishby_N	0.014
method	0.008	Barto_A	0.013
case	0.008	Viola_P	0.013

Topic 98			
Word	Prob	Author	Prob
image	0.049	Koch_C	0.119
visual	0.028	Horiuchi_T	0.106
field	0.023	Ruderman_D	0.088
system	0.020	Bialek_W	0.068
pixel	0.017	Dimitrov_A	0.05
filter	0.015	Bair_W	0.038
signal	0.013	Indiveri_G	0.035
object	0.013	Viola_P	0.030
center	0.012	Zee_A	0.030
local	0.011	Miyake_S	0.027

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

Table 1: Example of top topics learned from *NIPS* experiment

Hinton_G (Geoffrey Hinton)		
Topic 154	Topic 132	Topic 98
model	expert	image
image	task	visual
unit	mixture	field
hidden	network	system
hinton	architecture	pixel
code	gating	filter
digit	weight	signal
vector	nowlan	object
energy	soft	center
space	competitive	local

Bengio_Y (Yoshua Bengio)		
Topic 90	Topic 110	Topic 28
model	word	gate
data	speech	unit
parameter	recognition	input
mixture	training	threshold
distribution	frame	circuit
likelihood	system	polynomial
algorithm	error	output
probability	hmm	layer
density	level	parameter
gaussian	output	machine

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

conferences retrieved from Microsoft Academic Search [4] from each of the area. These publications are labeled by the 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:

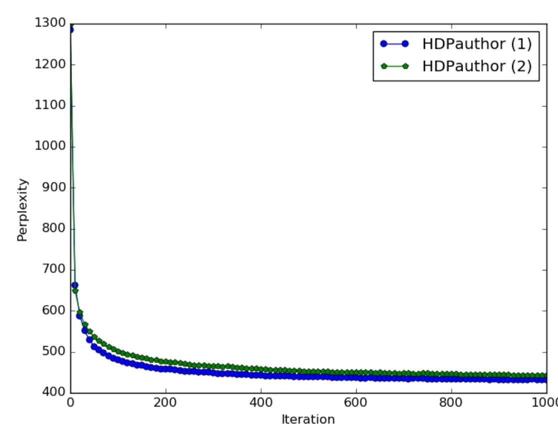

Figure 2: Perplexity evolution for *DBLP* experiments

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

[4]http://academic.research.microsoft.com/

Topic 3				Topic 11			
Word	Prob	Author	Prob	Word	Prob	Author	Prob
data	0.21	Charu C. Aggarwal	0.070	agent	0.147	Nicholas R. Jennings	0.076
stream	0.072	Jimeng Sun	0.046	mechanism	0.027	Sarit Kraus	0.056
mining	0.037	Philip S. Yu	0.035	system	0.018	Jeffrey S. Rosenschein	0.045
change	0.021	Kenji Yamanishi	0.034	negotiation	0.017	Kagan Tumer	0.036
time	0.020	Hans-Peter Kriegel	0.031	strategy	0.016	Kate Larson	0.036
application	0.012	Wei Wang	0.030	multi	0.014	Michael Wooldridge	0.035
real	0.012	Qiang Yang	0.028	problem	0.014	Moshe Tennenholtz	0.030
online	0.0094	Yong Shi	0.025	show	0.014	Vincent Conitzer	0.029
detect	0.008	Xiang Lian	0.019	multiagent	0.013	Sandip Sen	0.028
detection	0.008	Pedro P. Rodrigues	0.018	design	0.011	Victor R. Lesser	0.025

Topic 24				Topic 39			
Word	Prob	Author	Prob	Word	Prob	Author	Prob
document	0.093	ChengXiang Zhai	0.11	learn	0.093	Matthew E. Taylor	0.090
retrieval	0.066	Iadh Ounis	0.073	learning	0.084	Shimon Whiteson	0.079
query	0.055	Maarten de Rijke	0.020	reinforcement	0.034	Andrew Y. Ng	0.059
term	0.035	W. Bruce Croft	0.020	policy	0.033	Peter Stone	0.054
information	0.027	Laurence A. F. Park	0.020	task	0.032	Bikramjit Banerjee	0.051
model	0.026	James P. Callan	0.019	algorithm	0.029	Sherief Abdallah	0.040
relevance	0.021	Donald Metzler	0.017	transfer	0.019	Sridhar Mahadevan	0.039
feedback	0.020	Guihong Cao	0.017	action	0.019	Michael H. Bowling	0.036
collection	0.019	C. Lee Giles	0.016	function	0.018	Kagan Tumer	0.033
language	0.017	Oren Kurland	0.016	domain	0.016	David Silver	0.022

Table 3: Example of top topics learned from *DBLP* experiment

We also compare our `HDPauthor` model to other models as Okapi BM25[7], HDP modeling, Author-Topic (AT) model[11], by conducting retrieval tasks for queries constructed from academic documents outside training data set. We retrieved 100 papers from data set, and construct list of query word tokens from query paper by four methods: title only; content only; title with author; content with author.

We follow the steps from [10], add author names to each document as additional word tokens, and use author names of each query paper as additional query tokens for retrieval for Okapi BM25 and HDP modeling. For AT model and `HDPauthor` model, we add topic similarity score as one more measurement in retrieval score calculation, as:

$$p(q, \boldsymbol{a}_q|d_j, \boldsymbol{a}_j) = \omega \cdot p(q|d_j) + (1-\omega) \cdot similarity(\boldsymbol{a}_q, \boldsymbol{a}_j) \quad (12)$$

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 AT model, and set $K = 200$ for this experiment. We use one Python library called Gensim [9] for HDP topic learning.

7. CONCLUSION

We have presented a HDP-based hierarchical, nonparametric Bayesian generative model for author-topic hybrid learning, called `HDPauthor`. This model represents each author with a Dirichlet process of global topics, and represents each document as a mixture of these Dirichlet processes of it's authors. This model learns topic interests of authors, 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 our ability to represent each document, each author, and global topics as Dirichlet processes, or mixtures of Dirichlet processes. Therefore, none of them suffers from restrictions on the number of topic components that the user should 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.

8. FUTURE WORK

In future work, there are several directions we would like to explore:

1. A variational approximate inference [2] [6] approach can be used for our model. It is hard to infer[5], but more efficient and quicker to converge.

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

3. Combination of `HDPauthor` model with citation network [1] [14] can help us to construct a better model for author and document retrieval model.

9. REFERENCES

[1] V. Batagelj. Efficient algorithms for citation network analysis. *arXiv preprint cs/0309023*, 2003.

[2] D. M. Blei, M. I. Jordan, et al. Variational inference for dirichlet process mixtures. *Bayesian analysis*, 1(1):121–143, 2006.

[3] A. M. Dai and A. J. Storkey. Author disambiguation: a nonparametric topic and co-authorship model. In *NIPS Workshop on Applications for Topic Models Text and Beyond*, pages 1–4, 2009.

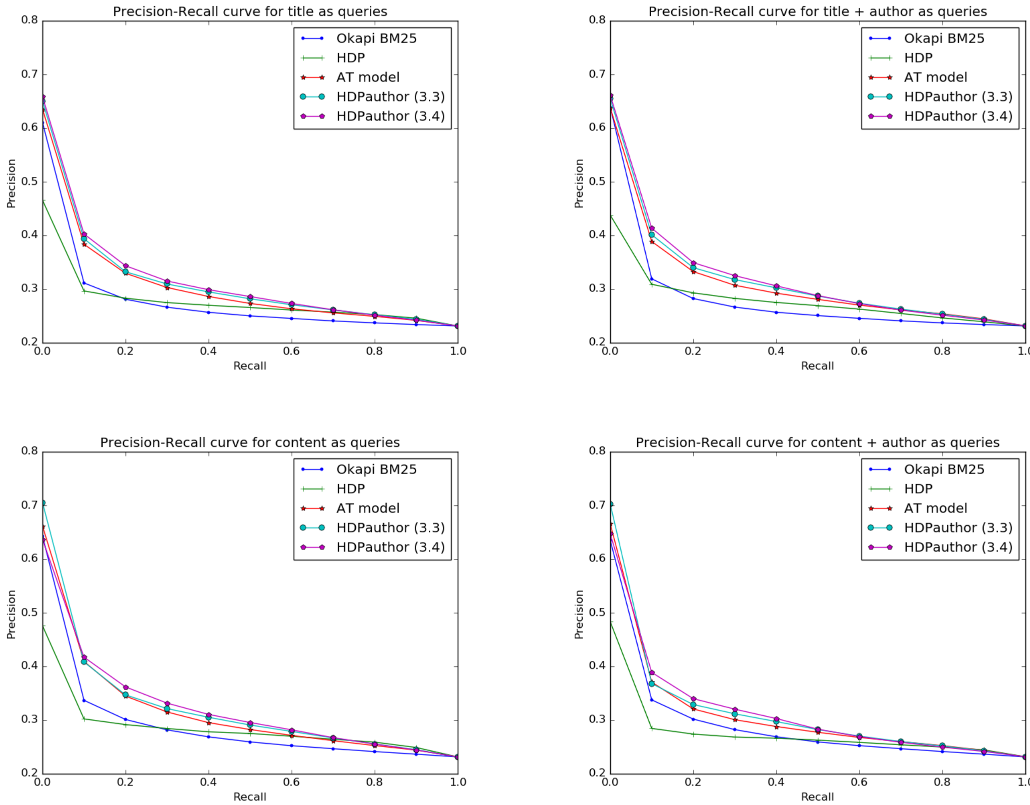

Figure 3: Precision-Recall curve for document retrieval for *DBLP* experiment

[4] A. M. Dai and A. J. Storkey. The grouped author-topic model for unsupervised entity resolution. In *Artificial Neural Networks and Machine Learning–ICANN 2011*, pages 241–249. Springer, 2011.

[5] S. Gershman, M. Hoffman, and D. Blei. Nonparametric variational inference. *arXiv preprint arXiv:1206.4665*, 2012.

[6] M. D. Hoffman, D. M. Blei, C. Wang, and J. Paisley. Stochastic variational inference. *The Journal of Machine Learning Research*, 14(1):1303–1347, 2013.

[7] K. S. Jones, S. Walker, and S. E. Robertson. A probabilistic model of information retrieval: development and comparative experiments: Part 2. *Information Processing & Management*, 36(6):809–840, 2000.

[8] C. D. Manning, P. Raghavan, H. Schütze, et al. *Introduction to information retrieval*, volume 1. Cambridge university press Cambridge, 2008.

[9] R. Řehůřek and P. Sojka. Software Framework for Topic Modelling with Large Corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, pages 45–50, Valletta, Malta, May 2010. ELRA. http://is.muni.cz/publication/884893/en.

[10] M. Rosen-Zvi, C. Chemudugunta, T. Griffiths, P. Smyth, and M. Steyvers. Learning author-topic models from text corpora. *ACM Transactions on Information Systems (TOIS)*, 28(1):4, 2010.

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

[12] A. Singhal. Modern information retrieval: A brief overview. *IEEE Data Eng. Bull.*, 24(4):35–43, 2001.

[13] Y. Song, J. Huang, I. G. Councill, J. Li, and C. L. Giles. Efficient topic-based unsupervised name disambiguation. In *Proceedings of the 7th ACM/IEEE-CS joint conference on Digital libraries*, pages 342–351. ACM, 2007.

[14] J. Tang, J. Zhang, L. Yao, J. Li, L. Zhang, and Z. Su. Arnetminer: extraction and mining of academic social networks. In *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 990–998. ACM, 2008.

[15] Y. W. Teh, M. I. Jordan, M. J. Beal, and D. M. Blei. Hierarchical dirichlet processes. *Journal of the american statistical association*, 101(476), 2006.

[16] M. Yang. *Hierarchical Bayesian Topic Modeling with Sentiment and Author Extension*. PhD thesis, Kansas State University, 2016.

[17] M. Yang and W. H. Hsu. Hdpsent: Incorporation of latent dirichlet allocation for aspect-level sentiment into hierarchical dirichlet process-based topic models.