Citation Author Topic Model in Expert Search

Yuancheng Tu, Nikhil Johri, Dan Roth, Julia Hockenmaier
University of Illinois at Urbana-Champaign
{ytu,njohri2,danr,juliahmr}@illinois.edu

Abstract

This paper proposes a novel topic model, Citation-Author-Topic (CAT) model that addresses a semantic search task we define as expert search – given a research area as a query, it returns names of experts in this area. For example, *Michael Collins* would be one of the top names retrieved given the query *Syntactic Parsing*.

Our contribution in this paper is two-fold. First, we model the cited author information together with words and paper authors. Such extra contextual information directly models linkage among authors and enhances the author-topic association, thus produces more coherent author-topic distribution. Second, we provide a preliminary solution to the task of expert search when the learning repository contains exclusively research related documents authored by the experts. When compared with a previous proposed model (Johri et al., 2010), the proposed model produces high quality author topic linkage and achieves over 33% error reduction evaluated by the standard MAP measurement.

1 Introduction

This paper addresses the problem of searching for people with similar interests and expertise, given their field of expertise as the query. Many existing people search engines need people's names to do a "keyword" style search, using a person's name as a query. However, in many situations, such information is insufficient or impossible to know beforehand. Imagine a scenario where the statistics department of a university invited a world-wide known expert in Bayesian statistics and machine learning to give a keynote speech; how can the organizer notify all the people on campus who are interested without spamming those who are not? Our paper proposes a solution to the aforementioned scenario by providing a search engine which goes beyond "keyword" search and can retrieve such information semantically. The organizer would only need to input the research domain of the keynote speaker, i.e. *Bayesian statistics, machine learning*, and all professors and students who are interested in this topic will be retrieved and an email agent will send out the information automatically.

Specifically, we propose a **C**itation-**A**uthor-**T**opic (CAT) model which extracts academic research topics and discovers different research communities by clustering experts with similar interests and expertise. CAT assumes three steps of a hierarchical generative process when producing a document: first, an author is generated, then that author generates topics which ultimately generate the words and cited authors. This model links authors to observed words and cited authors via latent topics and captures the intuition that when writing a paper, authors always first have topics in their mind, based on which, they choose words and cite related works.

Corpus linguists or forensic linguists usually

identify authorship of disputed texts based on stylistic features, such as vocabulary size, sentence length, word usage that characterize a specific author and the general semantic content is usually ignored (Diederich et al., 2003). On the other hand, graph-based and network based models ignore the content information of documents and only focus on network connectivity (Zhang et al., 2007; Jurczyk and Agichtein, 2007). In contrast, the model we propose in this paper fully utilizes the content words of the documents and combines them with the stylistic flavor contextual information to link authors and documents together to not only identify the authorship, but also to be used in many other applications such as paper reviewer recommendation, research community identification as well as academic social network search.

The novelty of the work presented in this paper lies in the proposal of jointly modeling the cited author information and using a discriminative multinomial distribution to model the co-author information instead of an artificial uniform distribution. In addition, we apply and evaluate our model in a semantic search scenario. While current search engines cannot support interactive and exploratory search effectively, our model supports search that can answer a range of exploratory queries. This is done by semantically linking the interests of authors to the topics of the collection, and ultimately to the distribution of the words in the documents.

In the rest of this paper, we first present some related work on author topic modeling and expert search in Sec. 2. Then our model is described in Sec. 3. Sec. 4 introduces our expert search system and Sec. 5 presents our experiments and the evaluation. We conclude this paper in Sec. 6 with some discussion and several further developments.

2 Related Work

Author topic modeling, originally proposed in (Steyvers et al., 2004; Rosen-Zvi et al., 2004), is an extension of Latent Dirichlet Allocation (LDA) (Blei et al., 2003), a probabilistic generative model that can be used to estimate the properties of multinomial observations via unsupervised learning. LDA represents each document as a mixture of probabilistic topics and each topic as a multinomial distribution over words. The Author topic model adds an author layer over LDA and assumes that the topic proportion of a given document is generated by the chosen author.

Author topic analysis has attracted much attention recently due to its broad applications in machine learning, text mining and information retrieval. For example, it has been used to predict authors for new documents (Steyvers et al., 2004), to recommend paper reviewers (Rosen-Zvi et al., 2004), to model message data (Mccallum et al., 2004), to conduct temporal author topic analysis (Mei and Zhai, 2006), to disambiguate proper names (Song et al., 2007), to search academic social networks (Tang et al., 2008) and to generate meeting status analyses for group decision making (Broniatowski, 2009).

In addition, there are many related works on expert search at the TREC enterprise track from 2005 to 2007, which focus on enterprise scale search and discovering relationships between entities. In that setting, the task is to find the experts, given a web domain, a list of candidate experts and a set of topics [1]. The task defined in our paper is different in the sense that our topics are hidden and our document repositories are more homogeneous since our documents are all research papers authored by the experts. Within this setting, we can explore in depth the influence of the hidden topics and contents to the ranking of our experts. Similar to (Johri et al., 2010), in this paper we apply CAT in a semantic retrieval scenario, where searching people is associated with a set of hidden semantically meaningful topics instead of their personal names.

In recent literature, there are three main lines of work that extend author topic analyses. One line of work is to relax the model's "bag-of-words" assumption by automatically discovering multi-word phrases and adding them into the original model (Johri et al., 2010). Similar work has also been proposed for other topic models such as Ngram topic models (Wallach, 2006; Wang and McCallum, 2005; Wang et al., 2007; Griffiths et al., 2007).

[1] http://trec.nist.gov/pubs.html

Another line of work models authors information as a general contextual information (Mei and Zhai, 2006) or associates documents with network structure analysis (Mei et al., 2008; Serdyukov et al., 2008; Sun et al., 2009). This line of work aims to propose a general framework to deal with collections of texts with an associated networks structure. However, it is based on a different topic model than ours; for example, Mei's works (Mei and Zhai, 2006; Mei et al., 2008) extend probabilistic latent semantic analysis (PLSA), and do not have cited author information explicitly.

Our proposal follows the last line of work which extends author topic modeling with specific contextual information and directly captures the association between authors and topics together with this contextual information (Tang et al., 2008; Mccallum et al., 2004). For example, in (Tang et al., 2008), publication venue is added as one extra piece of contextual information and in (Mccallum et al., 2004), email recipients, which are treated as extra contextual information, are paired with email authors to model an email message corpus. In our proposed method, the extra contextual information consists of the cited authors in each documents. Such contextual information directly captures linkage among authors and cited authors, enhances author-topic associations, and therefore produces more coherent author-topic distributions.

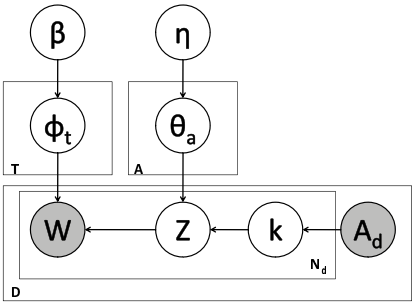

Figure 1: Plate notation of the previously proposed author topic models (Rosen-Zvi et al., 2004; Johri et al., 2010).

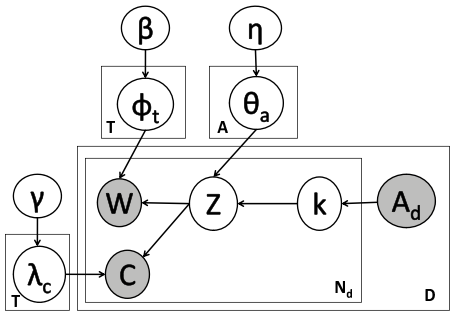

Figure 2: Plate notation of our current model: CAT generates words W and cited authors C independently given the topic.

3 The Citation-Author-Topic (CAT) Model

CAT extends previously proposed author topic models by explicitly modelling the cited author information during the generative process. Compared with these models (Rosen-Zvi et al., 2004; Johri et al., 2010), whose plate notation is shown in Fig. 1, CAT (shown in Fig. 2) adds cited author information and generates authors according to the observed author distribution.

Four plates in Fig. 1 represent topic ($\mathcal{T}$), author ($\mathcal{A}$), document ($\mathcal{D}$) and words in each document ($\mathcal{N}_d$) respectively. CAT (Fig. 2) has one more plate, cited-author topic plate, in which each topic is represented as a multinomial distribution over all cited authors (λ_c).

Within CAT, each author is associated with a multinomial distribution over all topics, $\vec{\theta_a}$, and each topic is a multinomial distribution over all words, $\vec{\phi_t}$, as well as a multinomial distribution over all cited authors $\vec{\lambda_c}$. Three symmetric Dirichlet conjugate priors, η, β and γ, are defined for each of these three multinomial distributions in CAT as shown in Fig. 2.

The generative process of CAT is formally defined in Algorithm 1. The model first samples the word-topic, cited author-topic and the author-topic distributions according to the three Dirichlet hyperparameters. Then for each word in each document, first the author k is drawn from the observed multinomial distribution and that author chooses the topic z_i, based on which word w_i and cited author c_i are generated independently.

CAT differs from previously proposed MAT (Multiword-enhanced Author Topic) model (Johri et al., 2010) in two aspects. First of all, CAT uses

Algorithm 1: CAT: $\mathcal{A}, \mathcal{T}, \mathcal{D}, \mathcal{N}$ are four plates as shown in Fig. 2. The generative process of CAT modeling.

Data: $\mathcal{A}, \mathcal{T}, \mathcal{D}, \mathcal{N}$
for *each topic* $t \in \mathcal{T}$ **do**
  draw a distribution over words:
  $\vec{\phi_t} \sim Dir_{\mathcal{N}}(\beta)$;
  draw a distribution over cited authors:
  $\vec{\lambda_c} \sim Dir_{\mathcal{C}}(\gamma)$;
for *each author* $a \in \mathcal{A}$ **do**
  draw a distribution over topics:
  $\vec{\theta_a} \sim Dir_{\mathcal{T}}(\eta)$;
for *each document* $d \in \mathcal{D}$ *and* k *authors* $\in d$ **do**
  for *each word* $w \in d$ **do**
    choose an author
    $k \sim Multinomial(A_d)$;
    assign a topic i given the author:
    $z_{k,i}|k \sim Multinomial(\theta_a)$;
    draw a word from the chosen topic:
    $w_{d,k,i}|z_{k,i} \sim Multinomial(\phi_{z_{k,i}})$;
    draw a cited author from the topic:
    $c_{d,k,i}|z_{k,i} \sim Multinomial(\lambda_{z_{k,i}})$

cited author information to enhance the model and assumes independence between generating the words and cited authors given the topic. Secondly, instead of an artificial uniform distribution over all authors and co-authors, CAT uses the observed discriminative multinomial distribution to generate authors.

3.1 Parameter Estimation

CAT includes three sets of parameters. The $\mathcal{T}$ topic distribution over words, ϕ_t which is similar to that in LDA. The author-topic distribution θ_a as well as the cited author-topic distribution λ_c. Although CAT is a relatively simple model, finding its posterior distribution over these hidden variables is still intractable due to their high dimensionality. Many efficient approximate inference algorithms have been used to solve this problem including Gibbs sampling (Griffiths and Steyvers, 2004; Steyvers and Griffiths, 2007; Griffiths et al., 2007) and mean-field variational methods (Blei et al., 2003). Gibbs sampling is a special case of Markov-Chain Monte Carlo (MCMC) sampling and often yields relatively simple algorithms for approximate inference in high dimensional models.

In our CAT modeling, we use a collapsed Gibbs sampler for our parameter estimation. In this Gibbs sampler, we integrated out the hidden variables θ, ϕ and λ using the Dirichlet delta function (Heinrich, 2009). The Dirichlet delta function with an M dimensional symmetric Dirichlet prior δ is defined as:

$$\Delta_M(\delta) = \frac{\Gamma\left(\delta^M\right)}{\Gamma\left(M\delta\right)}$$

Based on the independence assumptions defined in Fig. 2, the joint distribution of topics, words and cited authors given all hyperparameters which originally represented by integrals can be transformed into the delta function format and formally derived in Equation 1.

$$\begin{aligned} & P(\vec{z}, \vec{w}, \vec{c}|\beta, \eta, \lambda) \\ = \; & P(\vec{z}|\beta, \eta, \lambda) P(\vec{w}, \vec{c}|\vec{z}, \beta, \eta, \lambda) \\ = \; & P(\vec{z}) P(\vec{w}|\vec{z}) P(\vec{c}|\vec{z}) \\ = \; & \prod_{a=1}^{A} \frac{\Delta(n_A+\eta)}{\Delta(\eta)} \prod_{z=1}^{T} \frac{\Delta(n_{Z_w}+\beta)}{\Delta(\beta)} \prod_{z=1}^{T} \frac{\Delta(n_{Z_c}+\lambda)}{\Delta(\lambda)} \end{aligned} \tag{1}$$

The updating equation from which the Gibbs sampler draws the hidden variable for the current state j, i.e., the conditional probability of drawing the k^{th} author K_j^k, the i^{th} topic Z_j^i, and the c^{th} cited author C_j^c tuple, given all the hyperparameters and all the observed documents and authors, cited authors except the current assignment (the exception is denoted by the symbol $\forall \neg j$), is defined in Equation 2.

$$\begin{aligned} & P(Z_j^i, K_j^k, C_j^c | W_j^w, \forall \neg j, A_d, \beta, \eta, \gamma) \\ \propto \; & \frac{\Delta(n_Z+\beta)}{\Delta\left(n_{Z,\neg j}+\beta\right)} \frac{\Delta(n_K+\eta)}{\Delta\left(n_{K,\neg j}+\eta\right)} \frac{\Delta(n_C+\gamma)}{\Delta\left(n_{C,\neg j}+\gamma\right)} \\ = \; & \frac{n^w_{i,\neg j}+\beta_w}{\sum_{w=1}^{V} n^w_{i,\neg j}+V\beta_w} \frac{n^i_{k,\neg j}+\eta_i}{\sum_{i=1}^{T} n^i_{k,\neg j}+T\eta_i} \frac{n^c_{i,\neg j}+\lambda_c}{\sum_{c=1}^{C} n^c_{i,\neg j}+C\lambda_c} \end{aligned} \tag{2}$$

The parameter sets ϕ and θ, λ can be interpreted as sufficient statistics on the state variables of the Markov Chain due to the Dirichlet conjugate priors we used for the multinomial distributions.

These three sets of parameters are estimated based on Equations 3 , 4 and 5 respectively, in which n_i^w is defined as the number of times the word w is generated by topic i; n_k^i is defined as the number of times that topic i is generated by author k and n_c^i is defined as the number of times that the cited author c is generated by topic i. The vocabulary size is V, the number of topics is T and the cited-author size is C.

$$\phi_{w,i} = \frac{n_i^w + \beta_w}{\sum_{w=1}^{V} n_i^w + V\beta_w} \tag{3}$$

$$\theta_{k,i} = \frac{n_k^i + \eta_i}{\sum_{i=1}^{T} n_k^i + T\eta_i} \tag{4}$$

$$\lambda_{c,i} = \frac{n_i^c + \lambda_c}{\sum_{c=1}^{C} n_i^c + C\lambda_c} \tag{5}$$

The Gibbs sampler used in our experiments is adapted from the Matlab Topic Modeling Toolbox [2].

4 Expert Search

In this section, we describe a preliminary retrieval system that supports *expert search*, which is intended to identify groups of research experts with similar research interests and expertise by inputting only general domain key words. For example, we can retrieve *Michael Collins* via search for *natural language parsing*.

Our setting is different from the standard TREC expert search in that we do not have a pre-defined list of experts and topics, and our documents are all research papers authored by experts. Within this setting, we do not need to identify the status of our experts, i.e., a real expert or a communicator, as in TREC expert search. All of our authors and cited authors are experts and the task amounts to ranking the experts according to different topics given samples of their research papers.

The ranking function of this retrieval model is derived through the CAT parameters. The search aims to link research topics with authors to bypass the proper names of these authors. Our retrieval function ranks the joint probability of the query words (W) and the target author (a), i.e., $P(W, a)$. This probability is marginalized over all topics, and the probability that an author is cited given the topic is used as an extra weight in our ranking function. The intuition is that an author who is cited frequently should be more prominent and ranked higher. Formally, we define the ranking function of our retrieval system in Equation 6. c_a denotes when the author is one of the cited authors in our corpus. CAT assumes that words and authors, and cited authors are conditionally independent given the topic, i.e., $w_i \perp \text{a} \perp c_a$.

$$\begin{aligned} P(W, a) &= \sum_{w_i} \alpha_i \sum_t P(w_i, a|t, c_a)P(t, c_a) \\ &= \sum_{w_i} \alpha_i \sum_t P(w_i|t)P(a|t)P(c_a|t)P(t) \end{aligned} \tag{6}$$

W is the input query, which may contain one or more words. If a multiword is detected within the query, it is added into the query. The final score is the sum of all words in this query weighted by their inverse document frequency α_i.

In our experiments, we chose ten queries which cover several popular research areas in computational linguistics and natural language processing and run the retrieval system based on three models: the original author topic model (Rosen-Zvi et al., 2004), the MAT model (Johri et al., 2010) and the CAT model. In the original author topic model, query words are treated token by token. Both MAT and CAT expand the query terms with multiwords if they are detected inside the original query. For each query, top 10 authors are returned from the system. We manually label the relevance of these 10 authors based on the papers collected in our corpus.

Two standard evaluation metrics are used to measure the retrieving results. First we evaluate the precision at a given cut-off rank, namely precision at rank k with k ranging from 1 to 10. We then calculate the average precision (AP) for each query and the mean average precision (MAP) for

[2] http://psiexp.ss.uci.edu/research/programs_data/

the queries. Unlike precision at k, MAP is sensitive to the ranking and captures recall information since it assumes the precision of the non-retrieved documents to be zero. It is formally defined as the average of precisions computed at the point of each of the relevant documents in the ranked list as shown in Equation 7.

$$AP = \frac{\sum_{r=1}^{n}(Precision(r) \times rel(r))}{\mid relevant\, documents \mid} \quad (7)$$

To evaluate the recall of our system, we collected a pool of authors for six of our queries returned from an academic search engine, ArnetMiner (Tang et al., 2008)[3] as our reference author pool and evaluate our recall based on the number of authors we retrieved from that pool.

5 Experiments and Analysis

In this section, we describe the empirical evaluation of our model qualitatively and quantitatively by applying our model to the expert search we defined in Sec. 4. We compare the retrieving results with two other models: Multiword- enhanced Author Topic (MAT) model (Johri et al., 2010) and the original author topic model (Rosen-Zvi et al., 2004).

5.1 Data set and Pre-processing

We crawled the ACL anthology website and collected papers from ACL, EMNLP and CONLL over a period of seven years. The ACL anthology website explicitly lists each paper together with its title and author information. Therefore, the author information of each paper can be obtained accurately without extracting it from the original paper. However, many author names are not represented consistently. For example, the same author may have his/her middle name listed in some papers, but not in others. We therefore normalized all author names by eliminating middle names from all authors.

Cited authors of each paper are extracted from the reference section and automatically identified by a named entity recognizer tuned for citation extraction (Ratinov and Roth, 2009). Similar to regular authors, all cited authors are also normalized with their first name initial and their full last name. We extracted about 20,000 cited authors from our corpus. However, for the sake of efficiency, we only keep those cited authors whose occurrence frequency in our corpus is above a certain threshold. We experimented with thresholds of 5, 10 and 20 and retained the total number of 2,996, 1,771 and 956 cited authors respectively.

[3]http://www.arnetminer.org

Conf.	Year	Paper	Author	uni.	Vocab.
ACL	03-09	1,326	2,084	34,012	205,260
EMNLP	93-09	912	1,453	40,785	219,496
CONLL	97-09	495	833	27,312	123,176
Total	93-09	2,733	2,911	62,958	366,565

Table 1: Statistics about our data set. *Uni.* denotes unigram words and *Vocab.* denotes all unigrams and multiword phrases discovered in the data set.

We applied the same strategy to extract multiwords from our corpus and added them into our vocabulary to implement the model described in (Johri et al., 2010). Some basic statistics about our data set are summarized in Table 1 [4].

5.2 Qualitative Coherence Analysis

As shown by other previous works (Wallach, 2006; Griffiths et al., 2007; Johri et al., 2010), our model also demonstrates that embedding multiword tokens into the model can achieve more cohesive and better interpretable topics. We list the top 10 words from two topics of CAT and compare them with those from the unigram model in Table 2. Unigram topics contain more general words which can occur in every topic and are usually less discriminative among topics.

Our experiments also show that CAT achieves better retrieval quality by modeling cited authors jointly with authors and words. The rank of an author is boosted if that author is cited more frequently. We present in Table 3 the ranking of one of our ten query terms to demostrate the high quality of our proposed model. When compared to the model without cited author information, CAT not only retrieves more comprehensive expert list, its ranking is also more reasonable than the model without cited author information.

Another observation in our experiments is that

[4]Download the data and the software package at: http://L2R.cs.uiuc.edu/∼cogcomp/software.php.

Query term: **parsing**				
	Proposed CAT Model		Model without cited authors	
Rank	**Author**	**Prob.**	**Author**	**Prob.**
1	J._Nivre	0.125229	J._Nivre	0.033200
2	C._Manning	0.111252	R._Barzilay	0.023863
3	M._Johnson	0.101342	M._Johnson	0.023781
4	J._Eisner	0.063528	D._Klein	0.018937
5	M._Collins	0.047347	R._McDonald	0.017353
6	G._Satta	0.042081	L._Marquez	0.016003
7	R._McDonald	0.041372	A._Moschitti	0.015781
8	D._Klein	0.041149	N._Smith	0.014792
9	K._Toutanova	0.024946	C._Manning	0.014040
10	E._Charniak	0.020843	K._Sagae	0.013384

Table 3: Ranking for the query term: *parsing*. CAT achieves more comprehensive and reasonable rank list than the model without cited author information.

CAT	**Uni. AT Model**
TOPIC 49	Topic 27
pronoun_resolution	anaphor
antecedent	antecedents
coreference_resolution	anaphoricity
network	anphoric
resolution	is
anaphor	anaphora
pronouns	soon
anaphor_antecedent	determination
semantic_knowledge	pronominal
proper_names	salience
TOPIC 14	Topic 95
translation_quality	hypernym
translation_systems	seeds
source_sentence	taxonomy
word_alignments	facts
paraphrases	hyponym
decoder	walk
parallel_corpora	hypernyms
translation_system	page
parallel_corpus	logs
translation_models	extractions

Table 2: CAT with embedded multiword components achieves more interpretable topics compared with the unigram Author Topic (AT) model.

some experts who published many papers, but on heterogeneous topics, may not be ranked at the very top by models without cited author information. However, with cited author information, those authors are ranked higher. Intuitively this makes sense since many of these authors are also the most cited ones.

5.3 Quantitative retrieval results

One annotator labeled the relevance of the retrieval results from our expert search system. The annotator was also given all the paper titles of each corresponding retrieved author to help make this binary judgment. We experiment with ten queries and retrieve the top ten authors for each query.

Precision@K		
K	CAT Model	Model w/o Cited Authors
1	0.80	0.80
2	0.80	0.70
3	0.73	0.60
4	0.70	0.50
5	0.68	0.48
6	0.70	0.47
7	0.69	0.40
8	0.68	0.45
9	0.73	0.44
10	0.70	0.44

Table 4: Precision at K evaluation of our proposed model and the model without cited author information.

We first used the precision at k for evaluation. We calculate the precision at k for both our proposed CAT model and the MAT model, which does not have the cited author information. The results are listed in Table 4. It can be observed that at every rank position, our CAT model works better. In order to focus more on relevant retrieval results, we also calculated the mean average precision (MAP) for both models. For the given ten queries, the MAP score for the CAT model is 0.78, while the MAT model without cited author information has a MAP score of 0.67. The CAT model with cited author information achieves about 33% error reduction in this experiment.

Query ID	Query Term
1	parsing
2	machine translation
3	dependency parsing
4	transliteration
5	semantic role labeling
6	coreference resolution
7	language model
8	Unsupervised Learning
9	Supervised Learning
10	Hidden Markov Model

Table 5: Queries and their corresponding ids we used in our experiments.

Recall for each query		
Query ID	CAT Model	Model w/o Cite
1	0.53	0.20
2	0.13	0.20
3	0.27	0.13
4	0.13	0.2
5	0.27	0.20
6	0.13	0.26
Average	0.24	0.20

Table 6: Recall comparison between our proposed model and the model without cited author information.

Since we do not have a gold standard experts pool for our queries, to evaluate recall, we collected a pool of authors returned from an academic search engine, ArnetMiner (Tang et al., 2008) as our reference author pool and evaluated our recall based on the number of authors we retrieved from that pool. Specifically, we get the top 15 returned persons from that website for each query and treat them as the whole set of relevant experts for that query and our preliminary recall results are shown in Table 6.

In most cases, the CAT recall is better than that of the compared model, and the average recall is better as well. All the queries we used in our experiments are listed in Table 5. And the average recall value is based on six of the queries which have at least one overlap author with those in our reference recall pool.

6 Conclusion and Further Development

This paper proposed a novel author topic model, CAT, which extends the existing author topic model with additional cited author information.

We applied it to the domain of expert retrieval and demonstrated the effectiveness of our model in improving coherence in topic clustering and author topic association. The proposed model also provides an effective solution to the problem of *community mining* as shown by the promising retrieval results derived in our expert search system.

One immediate improvement would result from extending our corpus. For example, we can apply our model to the ACL ARC corpus (Bird et al., 2008) to check the model's robustness and enhance the ranking by learning from more data. We can also apply our model to data sets with rich linkage structure, such as the TREC benchmark data set or ACL Anthology Network (Radev et al., 2009) and try to enhance our model with the appropriate network analysis.

Acknowledgments

The authors would like to thank Lev Ratinov for his help with the use of the NER package and the three anonymous reviewers for their helpful comments and suggestions. The research in this paper was supported by the Multimodal Information Access & Synthesis Center at UIUC, part of CCICADA, a DHS Science and Technology Center of Excellence. This research also sponsored by the Army Research Laboratory (ARL) under agreement W911NF-09-2-0053. Any opinions, findings, conclusions or recommendations are those of the authors and do not necessarily reflect the view of the ARL.

References

Bird, S., R. Dale, B. Dorr, B. Gibson, M. Joseph, M. Kan, D. Lee, B Powley, D. Radev, and Y. Tan. 2008. The acl anthology reference corpus: A reference dataset for bibliographic research in computational linguistics. In *Proceedings of LREC'08*.

Blei, D., A. Ng, and M. Jordan. 2003. Latent dirichlet allocation. *Journal of Machine Learning Research*.

Broniatowski, D. 2009. Generating status hierarchies from meeting transcripts using the author-

topic model. In *In Proceedings of the Workshop: Applications for Topic Models: Text and Beyond*.

Diederich, J., J. Kindermann, E. Leopold, and G. Paass. 2003. Authorship attribution with support vector machines. *Applied Intelligence*, 19:109–123.

Griffiths, T. and M. Steyvers. 2004. Finding scientific topic. In *Proceedings of the National Academy of Science*.

Griffiths, T., M. Steyvers, and J. Tenenbaum. 2007. Topics in semantic representation. *Psychological Review*.

Heinrich, G. 2009. Parameter estimation for text analysis. Technical report, Fraunhofer IGD.

Johri, N., D. Roth, and Y. Tu. 2010. Experts' retrieval with multiword-enhanced author topic model. In *Proceedings of NAACL-10 Semantic Search Workshop*.

Jurczyk, P. and E. Agichtein. 2007. Discovering authorities in question answer communities by using link analysis. In *Proceedings of CIKM'07*.

Mccallum, A., A. Corrada-emmanuel, and X. Wang. 2004. The author-recipient-topic model for topic and role discovery in social networks: Experiments with enron and academic email. Technical report, University of Massachusetts Amherst.

Mei, Q. and C. Zhai. 2006. A mixture model for contextual text mining. In *Proceedings of KDD-2006*, pages 649–655.

Mei, Q., D. Cai, D. Zhang, and C. Zhai. 2008. Topic modeling with network regularization. In *Proceeding of WWW-08:*, pages 101–110.

Radev, D., M. Joseph, B. Gibson, and P. Muthukrishnan. 2009. A Bibliometric and Network Analysis of the field of Computational Linguistics. *Journal of the American Society for Information Science and Technology*.

Ratinov, L. and D. Roth. 2009. Design challenges and misconceptions in named entity recognition. In *Proc. of the Annual Conference on Computational Natural Language Learning (CoNLL)*.

Rosen-Zvi, M., T. Griffiths, M. Steyvers, and P. Smyth. 2004. the author-topic model for authors and documents. In *Proceedings of UAI*.

Serdyukov, P., H. Rode, and D. Hiemstra. 2008. Modeling multi-step relevance propagation for expert finding. In *Proceedings of CIKM'08*.

Song, Y., J. Huang, and I. Councill. 2007. Efficient topic-based unsupervised name disambiguation. In *Proceedings of JCDL-2007*, pages 342–351.

Steyvers, M. and T. Griffiths. 2007. Probabilistic topic models. In *Handbook of Latent Semantic Analysis*. Lawrence Erlbaum Associates.

Steyvers, M., P. Smyth, and T. Griffiths. 2004. Probabilistic author-topic models for information discovery. In *Proceedings of KDD*.

Sun, Y., J. Han, J. Gao, and Y. Yu. 2009. itopicmodel: Information network-integrated topic modeling. In *Proceedings of ICDM-2009*.

Tang, J., J. Zhang, L. Yao, J. Li, L. Zhang, and Z. Su. 2008. Arnetminer: Extraction and mining of academic social networks. In *Proceedings of KDD-2008*, pages 990–998.

Wallach, H. 2006. Topic modeling; beyond bag of words. In *International Conference on Machine Learning*.

Wang, X. and A. McCallum. 2005. A note on topical n-grams. Technical report, University of Massachusetts.

Wang, X., A. McCallum, and X. Wei. 2007. Topical n-grams: Phrase and topic discoery with an application to information retrieval. In *Proceedings of ICDM*.

Zhang, J., M. Ackerman, and L. Adamic. 2007. Expertise networks in online communities: Structure and algorithms. In *Proceedings of WWW 2007*.