

# Author-Topic over Time (AToT): A Dynamic Users' Interest Model

Shuo Xu<sup>1</sup>, Qingwei Shi<sup>1,2</sup>, Xiaodong Qiao<sup>3,\*</sup>, Lijun Zhu<sup>1</sup>,  
Hanmin Jung<sup>4</sup>, Seungwoo Lee<sup>4</sup>, and Sung-Pil Choi<sup>4</sup>

<sup>1</sup> Information Technology Supporting Center,  
Institute of Scientific and Technical Information of China,  
No. 15 fuxing Rd., Haidian District, Beijing 100038, P.R. China

<sup>2</sup> School of Software, Liaoning Technical University,  
No. 188 Longwan St. South, Huludao, Liaoning 125105, P.R. China

<sup>3</sup> College of Software, Northeast Normal University,  
5268 Renmin St., Changchun, Jilin 130024, P.R. China

<sup>4</sup> Department of Computer Intelligence Research,  
Korea Institute of Science and Technology Information,  
245 Daehak-ro, Yuseong-gu, Daejeon 305-806, Korea  
{xush,shiqw,qiaox,zhulj}@istic.ac.cn, {jhm,swlee,spchoi}@kisti.re.kr

**Abstract.** One of the key problems in upgrading information services towards knowledge services is to automatically mine latent topics, users' interests and their evolution patterns from large-scale S&T literatures. Most of current methods are devoted to either discover static latent topics and users' interests, or to analyze topic evolution only from intra-features of documents, namely text content without considering directly extra-features of documents such as authors. To overcome this problem, a dynamic users' interest model for documents using authors and topics with timestamps is proposed, named as Author-Topic over Time (AToT) model, and collapsed Gibbs sampling method is utilized for inferring model parameters. This model is not only able to discover latent topics and users' interests, but also to mine their changing patterns over time. Finally, the extensive experimental results on NIPS dataset with 1,740 papers indicate that our AToT model is feasible and efficient.

**Keywords:** Author-Topic (AT) Model, Topic over Time (ToT) Model, Author-Topic over Time (AToT) Model, Dynamic Users' Interest Model, Collapsed Gibbs Sampling.

## 1 Introduction

With a dynamic users' interest model, one can answer a range of important questions about the content of document collections, such as which topics each user prefers to, which users are similar to each other in terms of their interests, which users are likely to have written documents similar to an observed document, and

---

\* Corresponding author.

who are influential users at different stages of topic evolution and it also helps characterize users as pioneers, mainstream or laggards in different subject areas. Users' interests have shown their increasing importance for the development of personalized Web services and user-centric applications [1,2]. Hence, users' interest modeling has been attracting extensive attentions during the past few years, such as (a) Author-Topic (AT) model [3]; (b) Author-Recipient-Topic (ART) [4], Role-Author-Recipient-Topic (RART) [4] & Author-Persona-Topic (APT) models [5]; (c) Author-Interest-Topic (AIT) [6] & Latent-Interest-Topic (LIT) models [7], and (d) Author-Conference-Topic (ACT) model [8], etc.

In fact, in the process of entire scientific career, each researcher's interest is usually not static. However, the above models are devoted to discover static latent topics and research interests. Of course, one can perform some post-hoc or pre-hoc analysis [9,10] to discover changing patterns over time, but this misses the opportunity for time to improve topic discovery [11], and it is very difficult to align corresponding topics [12]. Currently, attention for dynamic models is mainly focused on analyzing topic evolution only from text content, such as Dynamic Topic Model (DTM) [13], continuous time DTM (cDTM) [14], Topic over Time (ToT) [11], and so on.

This article mainly focuses on the dynamic users' interest model. The organization of the rest of this paper is as follows. In Sec. 2, we discuss generative models for documents using authors and topics with timestamps, introduce the Author-Topic over Time (AToT) model in detail on the basis of AT and ToT models and describe the collapse Gibbs sampling methods used for inferring the model parameters. In Sec. 3, extensive experimental evaluations are conducted, and Sec. 4 concludes this work.

## 2 Author-Topic over Time (AToT) Model

The notation is summarized in Table 1, and the graphical model representations of the AToT model is shown in Fig. 1. The AToT model can be viewed as a generative process, which can be described as follows.

**Table 1.** Notation used in the AToT model

SYMBOL	DESCRIPTION
$K$	Number of topics
$M$	Number of documents
$V$	Number of unique words
$A$	Number of unique authors
$N_m$	Number of word tokens in document $m$
$A_m$	Number of authors in document $m$
$\mathbf{a}_m$	Authors in document $m$
$\vartheta_a$	Multinomial distribution of topics specific to the author $a$ . And let $\Theta = \{\vartheta_a\}_{a=1}^A$
$\varphi_k$	Multinomial distribution of words specific to the topic $k$ . And let $\Phi = \{\varphi_k\}_{k=1}^K$
$\psi_k$	Beta distribution of timestamp specific to the topic $k$ . And let $\Psi = \{\psi_k\}_{k=1}^K$
$z_{m,n}$	Topic associated with the $n$ -th token in the document $m$
$w_{m,n}$	$n$ -th token in document $m$
$x_{m,n}$	Chosen author associated with the word token $w_{m,n}$
$t_{m,n}$	Timestamp associated with the $n$ -th token in the document $m$
$\alpha$	Dirichlet priors (hyperparameter) to the multinomial distribution $\vartheta$
$\beta$	Dirichlet priors (hyperparameter) to the multinomial distribution $\varphi$

1. For each topic  $k \in [1, K]$  and each author  $a \in [1, A]$ , draw a  $\varphi_k \sim \text{Dirichlet}(\beta)$  and  $\theta_a \sim \text{Dirichlet}(\alpha)$ , respectively;
2. For each word  $n \in [1, N_m]$  in document  $m \in [1, M]$ :
  - Draw an author assignment  $x_{m,n} \sim \text{Uniform}(\mathbf{a}_m)$ ;
  - Draw a topic assignment  $z_{m,n} \sim \text{Multinomial}(\boldsymbol{\theta}_{x_{m,n}})$ ;
  - Draw a word  $w_{m,n} \sim \text{Multinomial}(\boldsymbol{\varphi}_{z_{m,n}})$ ;
  - Draw a timestamp  $t_{m,n} \sim \text{Beta}(\psi_{z_{m,n},1}, \psi_{z_{m,n},2})$ ;

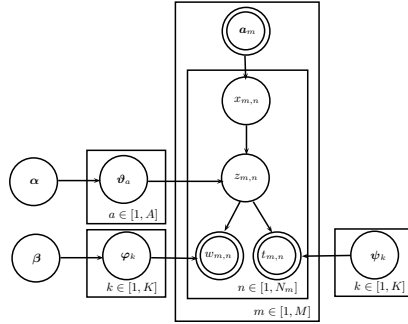


Fig. 1. The graphical model representation of the AToT model

For inference, the task is to estimate the sets of following unknown parameters in the AToT model: (1)  $\Phi, \Theta$  and  $\Psi$ ; (2) the corresponding topic and author assignments  $z_{m,n}, x_{m,n}$  for each word token  $w_{m,n}$ . In fact, inference can not be done exactly in this model. In this work, collapsed Gibbs sampling algorithm [15] is used, since it provides a simple method for obtaining parameter estimates under Dirichlet priors and allows combination of estimates from several local maxima of the posterior distribution.

In the Gibbs sampling procedure, we need to calculate the conditional distribution  $P(z_{m,n}, x_{m,n} | \mathbf{w}, \mathbf{z}_{-(m,n)}, \mathbf{x}_{-(m,n)}, \mathbf{t}, \mathbf{a}, \boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\Psi})$  with  $\mathbf{z}_{-(m,n)}, \mathbf{x}_{-(m,n)}$  represents the topic, author assignments for all tokens except  $w_{m,n}$ , respectively. We begin with the joint distribution  $P(\mathbf{w}, \mathbf{z}, \mathbf{x}, \mathbf{t} | \mathbf{a}, \boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\Psi})$  of a dataset, and using the chain rule, we can get the conditional probability conveniently as

$$\begin{aligned}
 & P(z_{m,n} = k, x_{m,n} = a | \mathbf{w}, \mathbf{z}_{-(m,n)}, \mathbf{x}_{-(m,n)}, \mathbf{t}, \mathbf{a}, \boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\Psi}) \\
 \propto & \frac{n_k^{(w_{m,n})} + \beta_{w_{m,n}} - 1}{\sum_{v=1}^V (n_k^{(v)} + \beta_v) - 1} \times \frac{n_a^{(k)} + \alpha_k - 1}{\sum_{k=1}^K (n_a^{(k)} + \alpha_k) - 1} \times \text{Beta}(\psi_{z_{m,n},1}, \psi_{z_{m,n},2}) (1)
 \end{aligned}$$

where  $n_k^{(v)}$  is the number of times tokens of word  $v$  is assigned to topic  $k$ , and  $n_a^{(k)}$  represents the number of times author  $a$  is assigned to topic  $k$ .

During parameter estimation, the algorithm keeps track of two large data structures: an  $A \times K$  count matrix  $n_a^{(k)}$  and an  $K \times V$  count matrix  $n_k^{(v)}$ .

From these data structures, one can easily estimate the  $\Phi$  and  $\Theta$  as follows:

$\varphi_{k,v} = \frac{n_k^{(v)} + \beta_v}{\sum_{v=1}^V (n_k^{(v)} + \beta_v)}$  and  $\vartheta_{a,k} = \frac{n_a^{(k)} + \alpha_k}{\sum_{k=1}^K (n_a^{(k)} + \alpha_k)}$ . As for  $\Psi$ , for simplicity and speed we update it after each Gibbs sample by the method of moments:  $\psi_{k,1} = \bar{t}_k \left( \frac{\bar{t}_k(1-\bar{t}_k)}{s_k^2} - 1 \right)$  and  $\psi_{k,2} = (1-\bar{t}_k) \left( \frac{\bar{t}_k(1-\bar{t}_k)}{s_k^2} - 1 \right)$ , where  $\bar{t}_k$  and  $s_k^2$  indicate the sample mean and biased sample variance of the timestamps belonging to topic  $k$ , respectively. The readers are invited to consult [16] for details. Note that the time range of the data is normalized to [0.01, 0.99].

### 3 Experimental Results and Discussions

NIPS proceeding dataset is utilized to evaluate the performance of our model, which consists of the full text of the 13 years of proceedings from 1987 to 1999 Neural Information Processing Systems (NIPS) Conferences. In addition to downcasing and removing stopwords and numbers, we also removed the words appearing less than five times in the corpus. The dataset contains 1,740 research papers, 2,037 unique authors, 13,649 unique words, and 2,301,375 word tokens in total. Each document’s timestamp is determined by the year of the proceedings. In our experiments,  $K$  is fixed at 100, and the symmetric Dirichlet priors  $\alpha$  and  $\beta$  are set at 0.5 and 0.1, respectively. Gibbs sampling is run for 2000 iterations.

#### 3.1 Examples of Topic, Author Distributions and Topic Evolution

Fig. 2 illustrates examples of 8 topics learned by AToT model. The topics are extracted from a single sample at the 2000th iteration of the Gibbs sampler. Each topic is illustrated with (1) the top 5 words most likely to be generated conditioned on the topic; (b) the top 5 authors which have the highest probability conditioned on the topic; and (c) histograms and fitted beta PDFs which show topics evolution patterns over time.

#### 3.2 Author Interest Evolution Analysis

In order to analyze further author interest evolution, it is interesting to calculate  $P(z, t|a) = P(z|a)p(z|t) = \vartheta_{a,z} \times \text{Beta}(\psi_{z,1}, \psi_{z,2})$ . In this subsection, we take Sejnowski as an example, who published 43 papers in total from 1987 to 1999 in the NIPS conferences, as shown Fig. 3 (a). The research interest evolution for Sejnowski is reported in Fig. 3 (b), in which the area occupied by a square is proportional to the strength of his research interest.

From Fig. 3 (b), one can see that Sejnowski’s research interest focused mainly on Topic 51 (Eye Recognition & Factor Analysis), Topic 37 (Neural Networks) and Topic 58 (Data Model & Learning Algorithm) but with different emphasis from 1987 to 1999. In the early phase (1989–1993), Sejnowski’s research interest is only limited to Topic 51, and then extended to Topic 37 in 1994 & Topic 58 in 1996 with great research interest strength, and finally back to Topic 51 after 1997. Anyway, Sejnowski did not change his main research direction, Topic 51, which is verified from his homepage again.

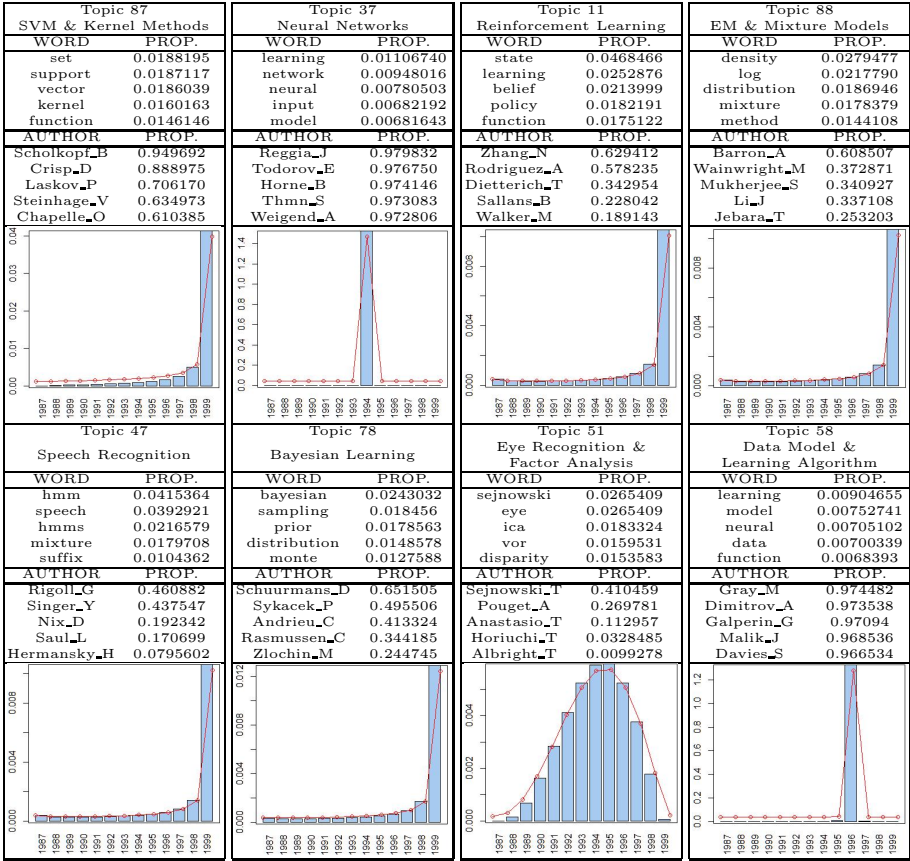
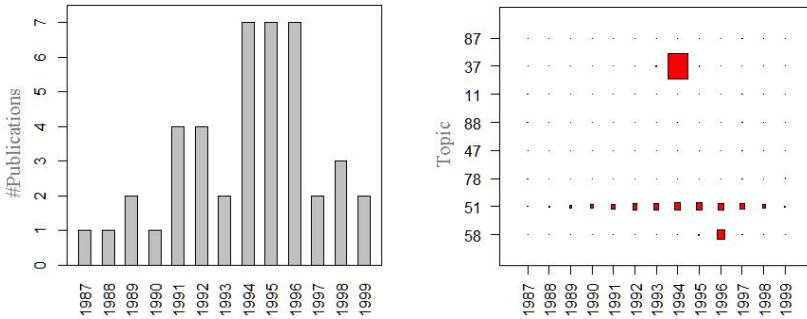


Fig. 2. An illustration of 8 topics from a 100-topic solutions for the NIPS collection. The titles are our own interpretation of the topics. Each topic is shown with the 5 words and authors that have the highest probability conditioned on that topic. Histograms show how the topics are distributed over time; the fitted beta PDFs is shown also.



(a) Distribution of #publications over time (b) Research Interest Evolution

Fig. 3. The distribution of #publications and research interest evolution for Sejnowski

### 3.3 Predictive Power Analysis

Similar to [3], we further divide the NIPS papers into a training set  $\mathcal{D}^{\text{train}}$  of 1,557 papers, and a test set  $\mathcal{D}^{\text{test}}$  of 183 papers of which 102 are single-authored papers. Each author in  $\mathcal{D}^{\text{test}}$  must have authored at least one of the training papers. The perplexity is a standard measure for estimating the performance of a probabilistic model. The perplexity of a test document  $\tilde{m} \in \mathcal{D}^{\text{test}}$ , is defined as the exponential of the negative normalized predictive likelihood under the model:  $\text{perplexity}(\mathbf{w}_{\tilde{m},\cdot}, \mathbf{t}_{\tilde{m},\cdot} | \mathbf{a}_{\tilde{m}}, \boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\Psi}) = \exp \left[ -\frac{\ln P(\mathbf{w}_{\tilde{m},\cdot}, \mathbf{t}_{\tilde{m},\cdot} | \mathbf{a}_{\tilde{m}}, \boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\Psi})}{N_{\tilde{m}}} \right]$  with

$$P(\mathbf{w}_{\tilde{m},\cdot}, \mathbf{t}_{\tilde{m},\cdot} | \mathbf{a}_{\tilde{m}}, \boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\Psi}) = \frac{1}{[A_{\tilde{m}}]^{N_m}} \times \sum_{\mathbf{z}_{\tilde{m},\cdot}} \text{Beta}(\psi_{z_{\tilde{m},n},1}, \psi_{z_{\tilde{m},n},2} | \mathcal{D}^{\text{train}}) \times \int p(\boldsymbol{\Phi} | \boldsymbol{\beta}, \mathcal{D}^{\text{train}}) \sum_{\mathbf{z}_{\tilde{m},\cdot}} \varphi_{z_{\tilde{m},n}, w_{\tilde{m},n}} d\boldsymbol{\Phi} \times \int p(\boldsymbol{\Theta} | \boldsymbol{\alpha}, \mathcal{D}^{\text{train}}) \sum_{\mathbf{x}_{\tilde{m},\cdot}} \vartheta_{x_{\tilde{m},n}, z_{\tilde{m},n}} d\boldsymbol{\Theta} \quad (2)$$

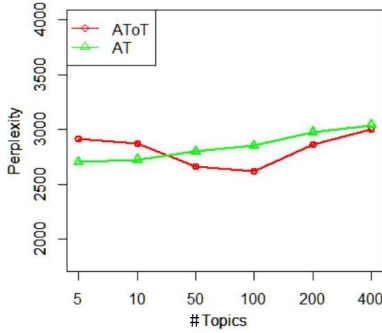


Fig. 4. Perplexity of the 102 single-authored test documents

We approximate the integrals over  $\boldsymbol{\Phi}$  and  $\boldsymbol{\Theta}$  using the point estimates obtained in Sec. 2 for each sample  $s \in \{1, 2, \dots, 10\}$  of assignments  $\mathbf{x}, \mathbf{z}$ , and then average over samples. Fig. 4 shows the results for the AToT model and AT model in a post-hoc fashion on 102 single-authored papers. It is not difficult to see that the perplexity of AToT model is smaller than that of AT model when #topics > 10, which indicates that AToT model outperforms AT model.

## 4 Conclusions

With a dynamic users' interest model, one can answer many important questions about the content of document collections. Based on AT & ToT models, this article proposes a dynamic users' interest model, Author-Topic over Time (AToT) model, for documents using authors and topics with timestamps, and collapsed Gibbs sampling is used for inferring model parameters. It combines the

merits of AT & ToT models. The results on NIPS dataset show the discovery of more salient topics and more reasonable users' interest evolution patterns. What's more, one can generalize the approach in the work to construct alternative dynamic models from other static users' interest models and ToT model.

**Acknowledgments.** This work was funded partially by Key Technologies R&D Program of Chinese 12th Five-Year Plan (2011–2015): Key Technologies Research on Large-Scale Semantic Calculation for Foreign STKOS, and Key Technologies Research on Data Mining from the Multiple Electric Vehicle Information Sources under grant number 2011BAH10B04 and 2013BAG06B01, respectively.

## References

1. Qiu, F., Cho, J.: Automatic identification of user interest for personalized search. In: WWW 2006, pp. 727–736. ACM, New York (2006)
2. Kim, J., Jeong, D.H., Lee, D., Jung, H.: User-centered innovative technology analysis and prediction application in mobile environment. *Multimed. Tools Appl.* (2013)
3. Rosen-Zvi, M., Chemudugunta, C., Griffiths, T., Smyth, P., Steyvers, M.: Learning author-topic models from text corpora. *ACM T. Inform. Syst.* 28(1), 1–38 (2010)
4. McCallum, A., Wang, X., Corrada-Emmanuel, A.: Topic and role discovery in social networks with experiments on enron and academic email. *J. Artif. Intell. Res.* 30(1), 249–272 (2007)
5. Mimno, D., McCallum, A.: Expertise modeling for matching papers with reviewers. In: KDD 2007, pp. 500–509. ACM, New York (2007)
6. Kawamae, N.: Author interest topic model. In: SIGIR 2010, pp. 887–888. ACM, New York (2010)
7. Kawamae, N.: Latent interest-topic model: Finding the causal relationships behind dyadic data. In: CIKM 2010, pp. 649–658. ACM, New York (2010)
8. Tang, J., Zhang, J., Jin, R., Yang, Z., Cai, K., Zhang, L., Su, Z.: Topic level expertise search over heterogeneous networks. *Mach. Learn.* 82(2), 211–237 (2011)
9. Steyvers, M., Smyth, P., Rosen-Zvi, M., Griffiths, T.: Probabilistic author-topic models for information discovery. In: KDD 2004, pp. 306–315. ACM, New York (2004)
10. Wang, X., Mohanty, N., McCallum, A.: Group and topic discovery from relations and their attributes. In: Weiss, Y., Schölkopf, B., Platt, J. (eds.) NIPS18, pp. 1449–1456. MIT Press, Cambridge (2006)
11. Wang, X., McCallum, A.: Topics over time: A non-Markov continuous-time model of topical trends. In: KDD 2006, pp. 424–433. ACM, New York (2006)
12. Xu, S., Zhu, L., Qiao, X., Shi, Q., Gui, J.: Topic linkages between papers and patents. In: AST 2012. SERSC, pp. 176–183. Daejeon, South Korea (2012)
13. Blei, D.M., Lafferty, J.D.: Dynamic topic models. In: ICML 2006, pp. 113–120. ACM, New York (2006)
14. Wang, C., Blei, D., Heckerman, D.: Continuous time dynamic topic models. In: UAI 2008, pp. 579–586 (2008)
15. Griffiths, T.L., Steyvers, M.: Finding scientific topics. *Proc. Natl. Acad. Sci. USA* 101(suppl. 1), 5228–5235 (2004)
16. Owen, C.B.: Parameter estimation for the Beta distribution. Master's thesis, Brigham Young University (2008)