Infinite Coauthor Topic Model (Infinite coAT): A Non-Parametric Generalization for coAT model

Han Zhang
Information Technology Support Center,
Institute of Scientific and Technical
Information of China(ISTIC)
No.15 Fuxing Rd., Haidian District,
Beijing 100038, P.R. China
zhanghan2012@istic.ac.cn

Shuo Xu
（corresponding author）
Information Technology Support Center,
Institute of Scientific and Technical
Information of China(ISTIC)
No.15 Fuxing Rd., Haidian District,
Beijing 100038, P.R. China
xush@istic.ac.cn

Xiaodong Qiao
Information Technology Support Center,
Institute of Scientific and Technical
Information of China(ISTIC)
No.15 Fuxing Rd., Haidian District,
Beijing 100038, P.R. China
qiaox@istic.ac.cn

Zhaofeng Zhang
Information Technology Support Center,
Institute of Scientific and Technical
Information of China(ISTIC)
No.15 Fuxing Rd., Haidian District,
Beijing 100038, P.R. China
zhangzf@istic.ac.cn

Hongqi Han
Information Technology Support Center,
Institute of Scientific and Technical
Information of China(ISTIC)
No.15 Fuxing Rd., Haidian District,
Beijing 100038, P.R. China
hanhq@istic.ac.cn

ABSTRACT

Inspired by the hierarchical Dirichlet process (HDP), we present a generalized coAT (coauthor Topic) model, also called infinite coAT model, in this paper. The infinite coAT model is a non-parametric extension of the coAT model. And this model can automatically determine the number of topics which are regarded for the probabilistic distribution of words. One does not need to provide prior information about the number of topics. In order to keep the consistency with the coAT model, the Gibbs sampling is utilized to infer the parameters. Finally, experimental results on the US patents dataset from US Patent Office indicate that our infinite-coAT model is feasible and efficient.

Categories and Subject Descriptors

H.3.3 [**Information Search and Retrieval**]:

General Terms

Algorithms, Performance

Keywords

coauthor topic (coAT) model, infinite coauthor topic (infinite-coAT) model, stick-breaking prior, hierarchical Dirichlet processes, collapsed Gibbs sampling.

Published at Ceur-ws.org

Proceedings of the First International Workshop on Patent Mining and Its Applications (IPAMIN) 2014. Hildesheim. Oct. 7th. 2014.

At KONVENS'14, October 8–10, 2014, Hildesheim, Germany.

1. INTRODUCTION

A social network is a social structure made up of a set of social actors (such as individuals or organizations) and a set of the dyadic ties between these actors [1] [2]. It can simulate various social relationships among people, such as shared interests, activities, backgrounds or real-life connections. And therefore social network analysis is very useful in measuring social characteristics and structure [2-6]. However most existing methods of social network analysis just consider the links between actors and ignore the attributes of links which may lead to several serious problems, for example, misdeeming some obvious wrong links for correct ones merely according to the number of collaborations between authors [7] and so on. Hence some methods considering both links and their attributes have been proposed [8-11], including our previous work—coauthor topic (coAT) model which can identify actors with similar interests from social networks.

But in the coAT model, users have to input the prior information about the number of topics ahead of time. In fact, users don't know the exact number of topics and therefore they can just guess an approximation. Hence how to choose the number of topics is a frequently raised question. Inspired by hierarchical Dirichlet processes (HDP) [12] [13], in this article, we introduce stick-breaking prior in the coAT model to propose an infinite coAT model. Thus, the infinite coAT model can not only discover the shared interests between authors, but also infer the adequate number of topics automatically.

The organization of the rest of this paper is as follow. In Section 2, we briefly introduce the coAT model and its inference. And then the non-parametric coAT model is proposed in Section 3, and the Gibbs sampling method is utilized to infer the model parameters in that section. In Section 4, experimental evaluations are conducted on US patents and Section 5 concludes this work.

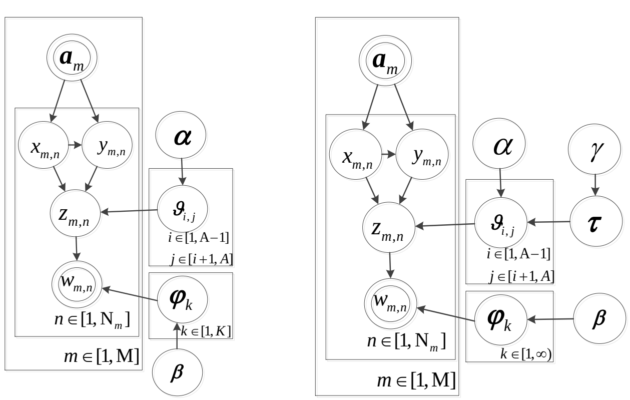
Notations For the convenience of depiction we summarize the notations in Table 1.

Table 1. Notation used in the models

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
$\boldsymbol{\varphi}_k$	The multinomial distribution of words specific to the topic k
$\boldsymbol{\vartheta}_{i,j}$	The multinomial distribution of topics specific to the coauthor relationship (i , j).
$z_{m,n}$	The topic assignment associated with the nth token in the document m
$w_{m,n}$	The nth token in document m
$x_{m,n}$	One chosen author associated with the word token $w_{m,n}$
$y_{m,n}$	Another chosen author associated with the word token $w_{m,n}$
$\boldsymbol{\alpha}$	Dirichlet priors (hyper-parameter) to the multinomial distribution $\boldsymbol{\vartheta}$ in coAT model
$\boldsymbol{\beta}$	Dirichlet priors(hyper-parameter) to the multinomial distribution $\boldsymbol{\varphi}$
$\boldsymbol{\tau}$	The root distribution of the hierarchical Dirichlet processes in infinite coAT model
α	scalar precision to the multinomial distribution $\boldsymbol{\vartheta}$ in infinite coAT model
γ	Dirichlet priors to the root distribution $\boldsymbol{\tau}$

2. Coauthor Topic (coAT) model

In this section, we introduce the coAT model with a fixed number of topics briefly, and the graphical model representation of the coAT model is shown in Fig. 1 a).



a) coAT b) infinite coAT

Fig.1. Admixture models for documents and coauthor relationship: a) The coAT model, b) the non-parametric coAT model—infinite coAT model.

The coAT model [11] can be viewed as the following generative process:

(1) For each topic $k \in [1,K]$:

(i) draw a multinomial φ_k from Dilichlet ($\boldsymbol{\beta}$);

(2) for each author pair (i, j) with $i \in [1,A\text{-}1]$, $j \in [i+1, A]$:

(i) draw a multinomial $\vartheta_{i,j}$ from Dirichlet ($\boldsymbol{\alpha}$);

(3) for each word $n \in [1, N_m]$ in document $m \in [1, M]$:

(i) draw an author $x_{m,n}$ uniformly from the group of authors $\mathbf{a}_m$;

(ii) draw another author $y_{m,n}$ uniformly from the group of authors $\mathbf{a}_m \backslash x_{m,n}$;

(iii) if $x_{m,n} > y_{m,n}$, to swap $x_{m,n}$ with $y_{m,n}$;

(iv) draw a topic assignment $z_{m,n}$ from multinomial ($\vartheta_{x_{m,n}, y_{m,n}}$);

(v) draw a word $w_{m,n}$ from multinomial ($\varphi_{z_{m,n}}$).

Based on the generative process above, the coAT model has two sets of unknown parameters: (1) $\Phi = \{\varphi_k\}_{k=1}^{K}$ and $\Theta = \{\{\vartheta_{i,j}\}_{i=1}^{A-1}\}_{j=i+1}^{A}$;(2) the corresponding topic and author pair assignments $z_{m,n}$ and ($x_{m,n}$, $y_{m,n}$) for each word token $w_{m,n}$. And the full conditional probability is as follow [11]:

$$P(z_{m,n}=k, x_{m,n}=i, y_{m,n}=j \mid \boldsymbol{w}, \boldsymbol{z}_{\neg(m,n)}, \boldsymbol{x}_{\neg(m,n)}, \boldsymbol{y}_{\neg(m,n)}, \mathbf{a}, \boldsymbol{\alpha}, \boldsymbol{\beta}) \propto \frac{n_{i,j}^{(k)}+\alpha_k-1}{\sum_{k=1}^{K}(n_{i,j}^{(k)}+\alpha_k)-1} \times \frac{n_k^{(v)}+\beta_v-1}{\sum_{v=1}^{V}(n_k^{(v)}+\beta_v)-1} \tag{1}$$

where $n_k^{(v)}$ is the number of times tokens of word v is assigned to topic k and $n_{i,j}^{(k)}$ represent the number of times author pair (i, j) is assigned to topic k.Then we get the parameter estimations with their definitions and Bayes' rules as follow [11]:

$$\varphi_{k,v} = \frac{n_k^{(v)}+\beta_v}{\sum_{v=1}^{V}(n_k^{(v)}+\beta_v)} \tag{2}$$

$$\vartheta_{i,j,k} = \frac{n_{i,j}^{(k)}+\alpha_k}{\sum_{k=1}^{K}(n_{i,j}^{(k)}+\alpha_k)} \tag{3}$$

3. Infinite Coauthor Topic (infinite coAT) model—nonparametric coAT model

How to choose the number of topics in coAT model is always a troublesome question. The hierarchical Dirichlet process (HDP) [12] [13] provides a non-parametric method to solve this problem. The method allows a prior over a countably infinite number of topics of which only a few will dominate the posterior. Inspired by this method, we propose an infinite coAT model shown as Fig.1b). Based on the parametric coAT model the infinite coAT model splits the Dirichlet hyper-parameter $\boldsymbol{\alpha}$ into a scalar precision α and a base distribution $\boldsymbol{\tau} \sim \mathrm{Dir}(\gamma/K)$[13]. Taking this to the limit $K \to +\infty$, we can get the root distribution for the non-parametric coAT model. In this way, we can retain the structure of the parametric case for the Gibbs update of parameters:

$$P(z_{m,n}=k, x_{m,n}=i, y_{m,n}=j \mid \boldsymbol{w}, \boldsymbol{z}_{\neg(m,n)}, \boldsymbol{x}_{\neg(m,n)}, \boldsymbol{y}_{\neg(m,n)}, \mathbf{a}, \boldsymbol{\alpha}, \boldsymbol{\beta})$$
$$\propto \begin{cases} \dfrac{n_{i,j}^{(k)}+\alpha\tau_k-1}{\sum_{k=1}^{K} n_{i,j}^{(k)}+\alpha-1} \times \dfrac{n_k^{(v)}+\beta_v-1}{\sum_{v=1}^{V}(n_k^{(v)}+\beta_v)-1}, & \text{if } z=k \\ \dfrac{\alpha\tau_{k+1}}{\sum_{k=1}^{K} n_{i,j}^{(k)}+\alpha-1} \times \dfrac{1}{V}, & \text{if } z=k_{new} \end{cases} \quad (4)$$

Note that the sampling space has K+1dimensions because the root distribution $\boldsymbol{\tau}$ provides K+1 possible states. We use $\alpha\tau_{K+1}/V$ to present all unused topics. If $\alpha\tau_{K+1}/V$ is sampled, a new topic is created as well. In that way, we can consider no information about the number of topics and the model will output the result automatically.

According to the inference above, the importance of the root distribution $\boldsymbol{\tau}$ in the non-parametric model becomes obvious, and how to sample $\boldsymbol{\tau}$ is naturally a crucial problem. In this paper, we can sample $\boldsymbol{\tau}$ by simulating how the new components are created and we can obtain a sequence of Bernoulli trials [13]:

$$p(m_{ijkr}=1)=\frac{\alpha\tau_k}{\alpha\tau_k+r-1} \quad \forall r\in[1,n_{i,j}^{(k)}], m\in[1,M], k\in[1,K] \quad (5)$$

The posterior of the top-level Dirichlet process $\boldsymbol{\tau}$ is then sampled via [13]

$$\tau \sim \mathrm{Dirichlet}([m_1,\cdots,m_k],\gamma) \quad (6)$$

with $\quad m_k=\sum_{ijr} m_{ijrk}$.

4. Experimental results and discussions

We downloaded US patents from US Patent Office[1] with the following search strategy on Jun 25, 2014[search strategy: ICL/F02M069/48 or TTL/("gas sensor" or "air sensor") and (VOC OR CO OR formaldehyde) or ABST/("gas sensor" or "air sensor") and (VOC OR CO OR formaldehyde) or ACLM/("gas sensor" or "air sensor") and (VOC OR CO OR formaldehyde) or SPEC/("gas sensor" or "air sensor") and (VOC OR CO OR formaldehyde)].The dataset contains 4760 patent abstracts and 7540 unique inventors, which is utilized to evaluate the performance of our model.

In our experiment, the infinite coAT model calculates the number of topics automatically which is 20. Because topics consist of probabilities of words, so we list 5 topics, the top ten words belonging to these topics with their probabilities and the top ten co-inventor relationships which have the highest probability conditioned on those topics respectively in Table 2. We can easily summarize the meaning of these topics. For example, topic 1 is obviously about "engine", topic 4 is about "material" and so on.

Table 2 An illustration of 5 topics from 20-topic solutions for air sensor patent dataset

Topic 1			
Word	Prob.	Co-inventor	Prob.
engine	0.05524	(Surnilla, Gopichandra; Roth, John M.)	0.97754
fuel	0.05332	(Yasui, Yuji; Akazaki, Shusuke)	0.97625
control	0.03385	(Lewis, Donald J.; Michelini, John O.)	0.97564
exhaust	0.03152	(Pursifull, Ross Dykstra; Surnilla, Gopichandra)	0.97451
system	0.02910	(Surnilla, Gopichandra; Smith, Stephen B.)	0.97230
combustion	0.02766	(Lewis, Donald J.; Russell, John D.)	0.96848
air	0.02521	(Bidner, David Karl; Cunningham, Ralph Wayne)	0.96833
method	0.02086	(Glugla, Chris Paul; Baskins, Robert Sarow)	0.96025
ratio	0.01671	(Akazaki, Shusuke; Iwaki, Yoshihisa)	0.95992
internal	0.01658	(Leone, Thomas G.; Stein, Robert A.)	0.95740
Topic 4			
Word	Prob.	Co-inventor	Prob.
oxide	0.02411	(Den, Tohru; Iwasaki, Tatsuya)	0.97003
material	0.02376	(Baughman,Ray Henry;Zakhidov,Anvar Abdulahadovic)	0.95226
layer	0.02227	(Suh, Dong-Seok; Baughman, Ray Henry)	0.95026
metal	0.02094	(Suh, Dong-Seok; Zakhidov, Anvar Abdulahadovic)	0.94531
film	0.01970	(Taylor, Earl J.; Moniz, Gary A.)	0.93500
method	0.01897	(Ishihara, Tatsumi; Takita, Yusaku)	0.92722
substrate	0.01799	(Godwin, Harold; Whiffen, David)	0.91776
semiconductor	0.00765	(Shindo, Yuichiro; Takemoto, Kouichi)	0.91718
thin	0.00949	(Itoh, Takashi; Kato, Katsuaki)	0.91239
device	0.00905	(Ata, Masafumi; Ramm, Matthias)	0.90789
Topic 6			
Word	Prob.	Co-inventor	Prob.
air	0.04102	(Owen, Donald R.; Kravitz, David C.)	0.96377
flow	0.02549	(Burbank, Jeffrey H.; Treu, Dennis M.)	0.96215
fluid	0.01982	(Brugger, James M.; Burbank, Jeffrey H.)	0.95740
system	0.01684	(Brugger, James M.; Treu, Dennis M.)	0.94809
apparatus	0.01565	(McMillin, John R.; Strandwitz, Peter)	0.92530
pressure	0.01433	(Hess, Joseph; Muller, Myriam)	0.92202
device	0.01163	(Brassil, John; Taylor, Michael John)	0.92164
chamber	0.01117	(Yasuda, Yoshinobu; Nakazeki, Tsugito)	0.91810
method	0.01005	(Johnstone; III, Albert E.)	0.91518
heat	0.00912	(Brassil, John; Schein, Douglas)	0.90206
Topic 9			
Word	Prob.	Co-inventor	Prob.
vehicle	0.03134	(Grubbs, Michael R.; Kenny, Garry R.)	0.93642
electric	0.01319	(Ogawa, Gen; Senda, Satoru)	0.87615
oil	0.01300	(Madan, Arun; Morrison, Scott)	0.85625
motor	0.01153	(Bingham, Lynn R.; Henke, Jerome R.)	0.85484
control	0.00812	(Pursifull, Ross Dykstra; Lewis, Donald J.)	0.84167
heating	0.00763	(Yamada, Hirohiko; Kokubo, Naoki	0.84167
position	0.00734	(Hjort, Klas Anders; Lindberg, Mikael Peter Erik)	0.81897
compartment	0.00724	(Bunyard, Marc R.; Holst, Peter A.)	0.79787
assembly	0.00714	(Gibson, Alex O'Connor; Nedorezov, Felix)	0.79348
speed	0.00607	(Masuda, Satoshi; Kokubo, Naoki)	0.78889
Topic 16			
Word	Prob.	Co-inventor	Prob.
electron	0.00143	(Yokoyama, Yoshiaki; Kodama, Tooru)	0.58696
soil	0.00143	(Takagi, Hiroshi; Takase, Hiromitsu)	0.50000
elastomer	0.00143	(Boden, Mark W.; Bergquist, Robert A.)	0.36111
radiative	0.00098	(Leuthardt, Eric C.; Lord, Robert W.)	0.32143
suppressing	0.00098	(Sato, Akira; Okamura, Masami)	0.13636
halides	0.00098	(Shiroma, Iris; Tomasco, Allan)	0.12500
inhalation	0.00054	(Berretta, Francine; Roberts, Joy)	0.12500
dioxins	0.00054	(Schielinsky, Gerhard; Kubach, Hans)	0.12500
program	0.00054	(Kamen,Dean L.;Langenfeld,Christopher C.)	0.09375
realized	0.00054	(Kubo, Yasuhiro; Ikegami, Eiji)	0.09375

[1] http://patft.uspto.gov/netahtml/PTO/search-adv.htm

Table 3 Co-invented patents between David Karl Bidner and Ralph Wayne Cunningham

Titles	Topic belonged to
Method and system for engine control	Topic 1
Particulate filter regeneration in an engine	Topic 1
Method and system for engine control	Topic 1
Particulate filter regeneration in an engine	Topic 1
Particulate filter regeneration in an engine	Topic 1
Particulate filter regeneration during engine shutdown	Topic 1
Particulate filter regeneration in an engine coupled to an energy conversion device	Topic 1
Method and system for engine control	Topic 1
Particulate filter regeneration during engine shutdown	Topic 1

We take David Karl Bidner and Ralph Wayne Cunningham as an example, and list their co-invented patents' titles in Table 3. From Table 3, one can easily find that their co-invented patents are all about the engine which is the meaning of topic 1. In other words, by comparing Table 3 with Table 2, it is not difficult to see that David Karl Bidner and Ralph Wayne Cunningham share interest Topic 1 with the strength of 0.96833 which illustrates that their co-invented patents all about topic 1 make sense.

In addition, in order to compare the performance of coAT and infinite coAT models, we use perplexity which is a standard measure to estimate the performance of probabilistic models to evaluate our models. And the smaller the perplexity is, the better the model performs. The perplexity is defined as the reciprocal geometric mean of the token likelihoods in the test set $\mathcal{D}$ = $\{\boldsymbol{w}_{\tilde{m}}, \boldsymbol{a}_{\tilde{m}}\}$ under the coAT or infinite coAT model:

$$perplexity^{coAT}(\boldsymbol{w}_{\tilde{m}} \mid \boldsymbol{a}_{\tilde{m}}, B) = \exp\left[-\frac{\ln P^{coAT}(\boldsymbol{w}_{\tilde{m}} \mid \boldsymbol{a}_{\tilde{m}}, B)}{N_{\tilde{m}} \times \frac{A_{\tilde{m}}(A_{\tilde{m}}-1)}{2}} \right] \quad (7)$$

$$perplexity^{icoAT}(\boldsymbol{w}_{\tilde{m}} \mid \boldsymbol{a}_{\tilde{m}}, B) = \exp\left[-\frac{\ln P^{icoAT}(\boldsymbol{w}_{\tilde{m}} \mid \boldsymbol{a}_{\tilde{m}}, B)}{N_{\tilde{m}} \times \frac{A_{\tilde{m}}(A_{\tilde{m}}-1)}{2}} \right] \quad (8)$$

where B is the set of all the prior parameters.

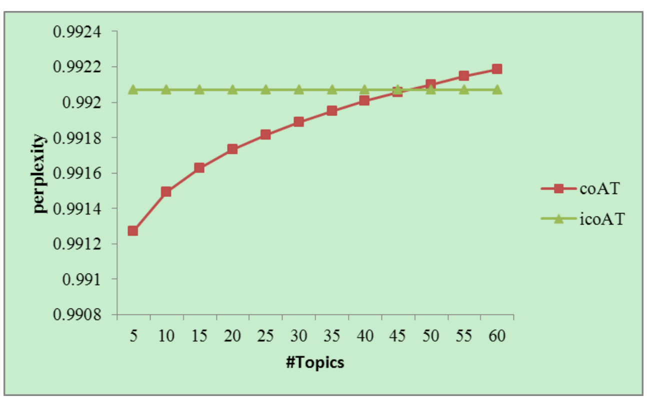


Fig.2 Perplexity of the test set $\mathcal{D}$

Fig.2 shows the results of the coAT and infinite coAT model. The perplexity increases in proportion to the number of topics, so the perplexity of the coAT model increases with the number of topics increasing and the perplexity of infinite coAT model stays stable with the dertermined number of topics 20. It is not difficult to see that when the number of topics in the coAT model is greater than 45, the perplexity of coAT model is bigger than that of infinite coAT model. But in the coAT model, we don't know choose what number of topics in advance, and what's more we prefer the bigger number such as 100. Hence, without the information of the exact number of topics, the infinite coAT model outperforms the coAT model.

5. Conclusions

In this paper, we generalize the coAT model to a nonparametric counterpart--infinite coAT model, which can estimate the number of topics. In that way, the model can not only discover the shared interests between inventors but also determine the number of topics automatically. Meanwhile, the experiments on US patent illustrate that the infinite coAT model is feasible.

In ongoing work, we can consider infinite coAT model over time to discover dynamic shared interests among authors or use this nonparametric method in other extended LDA models ,such as AToT models [14][15],to mine more useful information.

6. ACKNOWLEDGMENTS

This work is funded partially by the Natural Science Foundation of China: Research on Technology Opportunity Detection based on Paper and Patent Information Resources under grant number 71403255 and Study on the Disconnected Problem of Scientific Collaboration Network under grant number 71473237; Key Technologies R&D Program of Chinese 12th Five-Year Plan (2011–2015): Key Technologies Research on Data Mining from the Multiple Electric Vehicle Information Sources under grant number 2013BAG06B01; and Key Work Project of Institute of Scientific and Technical Information of China (ISTIC): Intelligent Analysis Service Platform and Application Demonstration for Multi-Source Science and Technology Literature in the Era of Big Data under grant number ZD2014-7-1.Our gratitude also goes to the anonymous reviewers for their valuable comments.

7. REFERENCES

[1] C. C. Aggarwal. *Social network data analytics*. Springer US, 2011.

[2] M. E. J. Newman. Scientific collaboration networks. I. Network construction and fundamental results. Physical review letters, 2001, 64(1): 016131-016131~016138.

[3] M. E. J. Newman. Scientific collaboration networks. II. Shortest paths, weighted networks, and centrality. Physical Review vol. 64, pp. 016132-1~7, 2001.

[4] A. Abbasi. Exploring the Relationship between Research Impact and Collaborations for Information Science. *In Proceedings of the 45th Hawaii International Conference on Systems Science* (HICSS-45), Hawaii, USA, 2012.

[5] Z. Zhang, Q. Li, D. Zeng, et al. User community discovery from multi-relational networks. *Decision Support Systems*, vol. 54, no.2, pp. 870-879, 2013.

[6] H. Han , S. Xu, J. Gui , X. Qiao, L. Zhu, H.Zhang. *Uncovering Research Topics of Academic Communities of Scientific Collaboration Network*. International Journal of Distributed Sensor Networks. 2014,4,529842,1-14.

[7] W. Chi, J. Han, Y. Jia, et al. Mining advisor-advisee relationships from research publication networks. *KDD*' 10, 2010.

[8] B. Taskar, A. Pieter, K. Daphne. Discriminative probabilistic models for relational data. *Eighteenth Conference (2002) on Uncertainty in Artificial Intelligence*, 2002: 485-492.

[9] L. E. Sucar. Probabilistic Graphical Models and Their Applications in Intelligent Environments. *In Intelligent Environments (IE), 2012 8th International Conference on*, 2012: 11-15

[10] P. Larrañaga, H. Karshenas, C. Bielza , et al. *A review on probabilistic graphical models in evolutionary computation*. Journal of Heuristics, 2012: 1-25.

[11] X. An, S. Xu, Y. Wen, et al. *A Shared Interest Discovery Model for Coauthor Relationship in SNS*. International Journal of Distributed Sensor Networks, 2014, 2014.

[12] Y .W. Teh, M.I. Jordan, M. J. Beal, et al. *Hierarchical Dirichlet processes*. Journal of the american statistical association, 2006, 101(476).

[13] G. Heinrich. *Infinite LDA implementing the HDP with minimum code complexity*. Technical note, Feb, 170, 2011.

[14] S. Xu, Q. Shi, X. Qiao, et al. Author-Topic over Time (AToT): A Dynamic Users' Interest Model. *Mobile, Ubiquitous, and Intelligent Computing*. Springer Berlin Heidelberg, 2014: 239-245.

[15] S. Xu, Q. Shi, X. Qiao, et al. *A dynamic users' interest discovery model with distributed inference algorithm.* International Journal of Distributed Sensor Networks, 2014, 2014.