Language Model for Assessing the Author Similarity

Jiayu Chen
Engineering Research Centre of Ministry of Education on Enterprise Digitalization Technology,Tongji University
China
1988_chenjiayu@tongji.edu.cn
Junli Wang
Engineering Research Centre of Ministry of Education on Enterprise Digitalization Technology,Tongji University
China
junliwang@tongji.edu.cn

ABSTRACT: *Currently, it is crucial for researchers to know if others have similar research objective. Nevertheless, the identification of authors sharing the same motivations and interests may be complex, especially as the amount of research publications is growing rapidly. Furthermore, information about research paper is often fragmented and incomplete. The incomplete information, or metadata, in this paper refers to abstracts, keywords, journals, organizations and so on. Thus, this paper analyzes the metadata information relative to an author, in order to find out similar authors. Author Similarity Model, a novel language model which is evolved from Author Topic Model, has been developed in this paper. For author similarity modeling, a four-dimensional vector has been set up to describe every author. Therefore author's neighbors (a group of people who have the same direction of research) can be found out by calculating similarity between vectors.*

Keywords: Author Similarity Model, Metadata, Author Topic Model

Received: 18 November 2015, Revised 20 December 2015, Accepted 17 January 2016

1. Introduction

Due to the rapidly growing number of researchers, it can be complex to find similar authors in different journals or conferences. That is, the task to find the right people to work together with to solve a scientific problem or to form a collaborative team is not so easy. Common interest is the first thing that need to be considered when making an author similarity assessment. Among all the resources available to do such an assessment, author's paper can be utilized to figure out his/her interests. Different paper parameters have already been used to establish links between authors. For example, co-citation and bibliographic coupling are standard measurements in scientometrics for detecting author similarity [1]. Author recommendation [2] has been proposed by adopting collaborative filtering (CF). The latter uses data from the social bookmark service CiteULike as well as multi-discipline information services Web of Science and Scopus to recommend authors as potential collaborators for a target scientist.

The first step to know the interests of an author is to find his/her papers. But the papers published by a single researcher are likely to be disseminated in multiple locations. It's far easier to get some incomplete, scattered information than to get complete ones.n this paper, metadata related to an author refers to abstracts, keywords, journals, and author's organization. Nowadays, most of the researches pay attention to full texts and don't make use of metadata. Furthermore, according to a few researches[3-4],metadata improved the accuracy of results for paper similarity assessing. So metadata is an interesting choice and this paper

takes metadata instead of comprehensive texts as sources of information.

Probabilistic language model has been used in many natural language processing applications such as speech recognition, machine translation and information retrieval. Ref.[5] used language model to improve retrieval accuracies of some search engines. In order to find how similar two authors are, this paper also chooses to employ a novel set of semantic similarity methods relying on language modeling method Author Topic Model [6-7]. Topic model has been put forward by David Blei, especially Latent Dirichlet Allocation (LDA) [8]. LDA models document as topical distribution and topics as distributions over words in the vocabulary. Each word has a certain contribution to a topic. After the introduction of LDA, many people have extended it by including other information. Author Topic Model consists of authorship information. Each author is associated with a multinomial distribution over topics and each topic is involved in a multinomial distribution over words. The main task in this paper is to verify how to use metadata to compare authors in the vector space model (VSM) [9] and the Author Similarity model. VSM applying to author's abstracts is a basic method. For Author Similarity modeling, this paper includes keywords, journals and organizations to estimate the similarities between authors. Furthermore, Author Topic Model is a part of the model and will be utilized to discover latent topics of authors and documents.The structure of this paper is as follows. The next section describes related work and background. In section 3, the Author Similarity Model is introduced. Section 4 presents a running example to explain the details of experiments. In Section 5 the results of experiments are analyzed. Finally we conclude and recommend further work that could be done.

2. Related Work

Author Topic Model [6-7] simultaneously models the contents and the interests of authors. This generative model represents each document as a mixture of probabilistic topics, which is similar to LDA. It extends LDA by using probabilistic topics-to-author modeling by allowing the mixture weights for different topics to be determined by the authors of the document. Conditioned on the set of authors and their distributions over topics, the process by which a document is generated can be summarized as follows: Assume we have T topics. We can parameterize the multinomial distribution over topics for each author using matrix θ of size $T \times A$, with elements θ_{ta} that stand for the probability of assigning topic t to a word generated by author t. Thus $\sum_{t=1}^{T} \theta_{ta} = 1$and $\theta_a$stands for the ath column of the matrix. The multinomial distributions over words associated with each topic are parameterized by a matrix $\emptyset$ of size $W \times T$, with elements φ_{wt} that stand for the probability of generating word w from topic t. Thus $\sum_{w=1}^{W} \varphi_{wt} = 1$ and stands φ_t for the t^{th}column of the matrix. These multinomial distributions are assumed to be generated from symmetric Dirichlet priors with hyper-parameters α and β respectively.

The graphical model corresponding to this process is given in Figure 1.

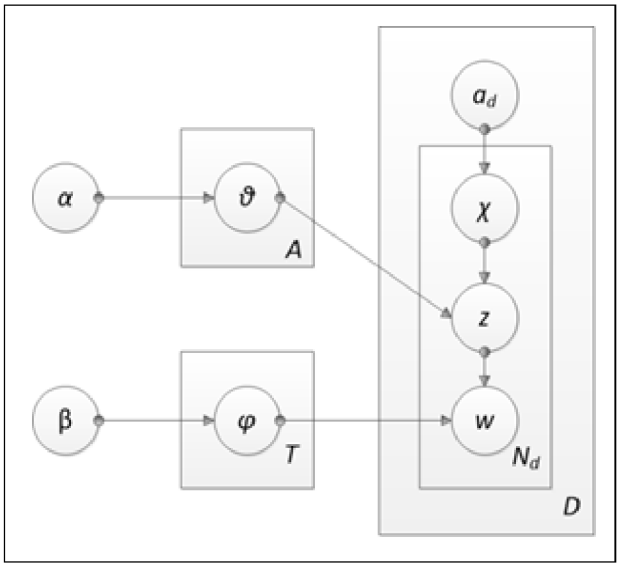

Figure 1. Graphical model for the Author Topic Model

A variety of algorithms have been applied to estimate the parameters of topic models, from basic expectation-maximization [4], to approximate inference methods like variational EM[8], expectation propagation[10], and Gibbs sampling[11-13]. The inference used throughout this paper is Gibbs sampling that based on Markov chain Monte Carlo algorithm. Our aim is to estimate the posterior distribution $P(\theta, \emptyset \mid D^{train}, \alpha, \beta)$. Sample from this distribution can be useful in many applications. Here we use the result to calculate the similarity between authors. The inference is based on the observation that:

$$P(\theta,\phi \mid D^{train}, \alpha, \beta) = \sum_{(Z,X)} P(\theta,\phi \mid z,x,Dtrain, \alpha, \beta) P(z,x \mid Dtrain, \alpha, \beta) \quad (1)$$

We obtain an approximate posterior on θ and ϕ by using the Gibbs sampler to compute the sum over *z* and *x*.

Author Topic Model has served in many areas. For example, Ref.[14] presents the Author-Recipient-Topic model for social network analysis to predict people's roles, which learns topic distributions based on the direction-sensitive messages sent among entities. The model builds on Latent Dirichlet Allocation and the Author-Topic (AT) model, adding the key attribute that distribution over topics is conditioned distinctly on both the sender and recipient. Ref.[15] uses the Author Topic clustering algorithm, trying to discern employees interests from their daily emails and find out the insider threat. Ref.[16] uses Author Topic Model to analyze topic trends over time, finding the authors who are most likely to write on a given topic, and finding the most unusual paper written by a given author.

3. Author Similarity Model

In this paper, a novel approach to calculate the similarities has been put forward and it is Author Similarity Model. Each author is a four-dimensional vector. As defined below:

$$A_i = (A_{i1}, A_{i2}, A_{i2}, A_{i4}) \quad (2)$$

Furthermore, each dimension is a vector that can be defined as follows:

$A_{i,1}$ is a vector which represents an author's distributions over words. Author-Topic model is applied on abstracts. The results are author-topic distributions matrix and topic-word distributions matrix. As formula 3 defined, it's easy to get an author's distributions over words.

$$A_{i1} = P(w_k \mid t) \times P(t \mid AD_i) \quad (3)$$

AD_i is the set of author i's papers , t is topic, k meaning the word is the *k*-th word in the dictionary.

A_{i2} is a vector which represent an author's distribution over words. But here we use Organization-Topic model instead of Author Topic model and pseudo documents other than abstracts. The Organization-Topic model is derived from Author-Topic model, just replace the author to organization. Here the pseudo document for each organization combines all the papers that come from the same organization. First get the organization-topic distributions matrix and topic-word distributions matrix. Then by using formula 4, a distribution of each organization in the dictionary word entries will be easy to calculate:

$$A_{i2} = P(w_k|t) \times P(t|AS_l) \quad (4)$$

AS_l refers to all the papers of organization *l*, t is topic, k means word is the k-th word in the dictionary.
A_{i3} is a vector of keywords related to the author.A_{i3} combine all the keywords that appeared in author s papers. For example, an author wrote three papers. In the original paper, there are keywords k_1,k_2. In the second paper, there are keywords k_1,k_3. In the third paper, there are keywords k_2,k_3. So the author i's keywords vector is like this(k_1,k_2,k_3...). It's more convenient to convert this vector into a numeric representation and then compute it. That is, if a keyword appeared, the corresponding vector dimension labeled as 1, if not, marked as 0.

A_{i4} is a vector of journals related to the author i. A_{i4} will combine all the journals where author i has published his/her papers. For example, the authors wrote two papers. The first paper published on journal $j1$, the second paper on journal $j2$. The author's journal vector is like this $(j1, j2, \dots)$. It's more convenient to convert this vector into a numeric representation and then compute it. That is, if published in the journal, the corresponding vector dimension should be labeled as 1, if not, marked as 0.
Next step is to define a formula to assess the similarity of two authors. μ_i is the weight of each factor. As each vector is defined differently, we have to calculate the similarities separately as follows:

$$\begin{aligned} &\mathrm{sim}(A_i, A_j) = \\ &\mu_1 sim(A_{i1}, A_{j1}) + \mu_2 sim(A_{i2}, A_{j2}) + \\ &\mu_3 sim(A_{i3}, A_{j3}) + \mu_4 sim(A_{i4}, A_{j4}) \end{aligned} \tag{5}$$

$$sim(A_{i1}, A_{j1}) = \frac{KLD\left(A_{i1} \middle| \frac{A_{i1}+A_{j1}}{2}\right) + KLD\left(A_{j1} \middle| \frac{A_{i1}+A_{j1}}{2}\right)}{2} \tag{6}$$

$$sim(A_{i2}, A_{j2}) = \frac{KLD\left(A_{i2} \middle| \frac{A_{i2}+A_{j2}}{2}\right) + KLD\left(A_{j2} \middle| \frac{A_{i2}+A_{j2}}{2}\right)}{2} \tag{7}$$

$$\mathrm{sim}(\overrightarrow{A_{i3}}, \overrightarrow{A_{j3}}) = \frac{\overrightarrow{A_{i3}} \cdot \overrightarrow{A_{j3}}}{\|\overrightarrow{A_{i3}}\| \cdot \|\overrightarrow{A_{j3}}\|} \tag{8}$$

$$\mathrm{sim}(\overrightarrow{A_{i3}}, \overrightarrow{A_{j3}}) = \frac{\overrightarrow{A_{i3}} \cdot \overrightarrow{A_{j3}}}{\|\overrightarrow{A_{i3}}\| \cdot \|\overrightarrow{A_{j3}}\|} \tag{9}$$

While the formula 6 and 7 is trying to use symmetric KL distance to calculate the similarities. Formula 8 and 9 are trying to use cosine similarity to calculate the similarities.
In the last step, final results will show up through Formula 5.

4. Running example

There is a running example to illustrate how to calculate the similarities between authors. Suppose four papers and six authors in the corpus. Organizations for each author are demonstrated in the last line.

$$\begin{aligned} &d1 = (abs = (a, b, c, d, a), kws = (k1, k2), aut = (u1, u2), jou = (j1)) \\ &d2 = (abs = (a, a, d, a, b, a), kws = (k1, k3), aut = (u1, u3), jou = (j1)) \\ &d3 = (abs = (a, b, a), kws = (k2, k3), aut = (u2, u4), jou = (j2)) \\ &d4 = (abs = (a, b, b, e), kws = (k4), aut = (u5, u6), jou = (j2)) \\ &\quad u1, u2, u4 \in org1, \quad u3 \in org2, \quad u5, u6 \in org3 \end{aligned}$$

First Vector Space Model (VSM) is applied to run a contrast experiment. Then Author Similarity Model is the second method. Last step is to compare the results of different method.

4.1 VSM
Firstly, each author's pseudo paper will be put as below.

$$\begin{aligned} A_1 &= (a, b, a, c, d, a, a, a, d, a, b, a) \\ A_2 &= (a, b, a, c, d, a, a, b, a) \end{aligned}$$

$$A_3 = (a, a, d, a, b, a)$$

$$A_4 = (a, b, a)$$

$$A_5 = (a, b, b, e)$$

$$A_5 = (a, b, b, e)$$

Using the TF-IDF to compute the weights:

$$tf_a = \frac{7}{12}, idf_a = log\frac{6}{6} = 0, tf_a \times idf_a = 0$$

$$tf_b = \frac{2}{12}, idf_b = log\frac{6}{6} = 0, tf_b \times idf_b = 0;$$

$$tf_c = \frac{1}{12}, idf_c = log\frac{6}{2} = log3, tf_c \times idf_c = \frac{log3}{12};$$

$$tf_d = \frac{2}{12}, idf_d = log\frac{6}{3} = log2, tf_d \times idf_d = \frac{log2}{6};$$

$$tf_e = 0, idf_e = log\frac{6}{2} = log3, tf_e \times idf_e = 0;$$

$$A_1 = (0,0,\frac{log3}{12},\frac{log2}{6},0)$$

$$A_2 = (0,0,\frac{log3}{9},\frac{log2}{9},0)$$

.....

Then cosine similarity is used for calculate and set a threshold to pick put the parallel pairs.

$$sim(A_i, A_j) = \frac{A_i \cdot A_j}{\|A_i\|\|A_j\|} \quad (10)$$

4.2 Author Similarity Model

For each author, the four-dimensional vector is: $A_i = (A_{i1}, A_{i2}, A_{i3}, A_{i4})$

Suppose in all the paper, only two topics G, H.

As depicted before, this approach tries to use the Author-Topic model to get the author-topic distribution matrix and topic-word distribution matrix. Here we take author 1 to give an example.

$$A_{11} = P(w_k|t) \times P(t|A_1)$$

$$P(a|A_1) = P(a|G) \times P(G|A_1) + P(a|H) \times P(H|A_1)$$

$$P(b|A_1) = P(b|G) \times P(G|A_1) + P(b|H) \times P(H|A_1)$$

$$P(c|A_1) = P(c|G) \times P(G|A_1) + P(c|H) \times P(H|A_1)$$

$$P(d|A_1) = P(d|G) \times P(G|A_1) + P(d|H) \times P(H|A_1)$$

$$P(e|A_1) = P(e|G) \times P(G|A_1) + P(e|H) \times P(H|A_1)$$

Using the distribution matrixes from Author-Topic model, then put them in the formula above, it's not difficult to get the author-word distribution.

$$A_{11} = (P(a|A_1), P(b|A_1), P(c|A_1), P(d|A_1), P(e|A_1))$$

With the same method, it's straightforward to compute A_{21},...,A_{61}.

It's easy to get the result of Organization-Topic model derived from Author-Topic model, which means the calculation of A_{i2} is similar to A_{i1} .

As for A_{i3} , it has to combine all the keywords of the author *i* first, then get the vector.(The way to calculate A_{i4} is similar to A_{i3} , just have to change the keywords to journals) :

$$A_{13} = (1,1,1,0) \qquad A_{14} = (1,1,1,0)$$

$$A_{23} = (1,1,1,0) \qquad A_{24} = (1,1,1,0)$$

$$A_{33} = (1,0,1,0) \qquad A_{34} = (1,0,1,0)$$

$$A_{43} = (0,1,1,0) \qquad A_{44} = (0,1,1,0)$$

$$A_{53} = (0,1,1,0) \qquad A_{54} = (0,1,1,0)$$

$$A_{63} = (0,1,1,0) \qquad A_{64} = (0,1,1,0)$$

Then we can calculate the similarities separately by using formula 5.

The last step is to set the threshold for similarity pairs and compare the results.

5. Experimental setup and Evaluation

5.1 Experimental setup

VSM and Author Similarity Model have been applied to a corpus consisting of metadata from VLDB Journal from 1992 to 2013, Machine Learning Journal from 1986 to 2013 and Data Mining and Knowledge Discovery Journal from 1997 to 2013. These files contain 1552 abstracts, 2199 authors, 1918 organizations, 4309 keywords altogether. It corresponds to 1552 papers from these three journals and for each paper we only consider the first two authors (only one author, if the paper has a single author). For the abstracts, firstly, stopwords have been removed. StanfordParser[1] has been applied to work out the grammatical structure of sentences. Then a test collection by using VSM approach and Author Similarity Model approach has been implemented. For the Author Topic Model and Organization-Topic model, topic numbers from 10 to 70 has been test to get the best topic number and the final result is 50. For 2199 authors, we calculate all the pairs and set the threshold to 0.8/1.

5 experts were invited to verify the truth of the results. For each approach, experts analyze the author pairs to calculate the accuracy. Formula 13 defined below is to evaluate these pairs.

$$Ac = \frac{Pre\ (p)}{|p|} \qquad (11)$$

Pre (***p***) is the number of right pairs verified by experts and $|p|$ is the number of all the pairs result from different approach.

5.2 Results

Table 1 shows the experimental results obtaining from the Author Similarity Model and the VSM. As we compare the first line and the last line, Author Topic model applied to abstracts (we say ATA for short in the following part) outperform than traditional VSM. This suggests the Author Similarity model achieves better performance when comparing pairs.

As presented before, Author Similarity Model is based on a four-dimensional vector. The four dimensions of this vector are involved with Author Topic model, Organization Topic model, keywords information and journal information. We use configuration described in Table 1 to control the weights of different factors. In Table 1, A_{i1} indicates Author Topic model applied to abstracts, A_{i2} indicates Organization Topic model applied to abstracts, A_{i3} indicates keywords and A_{i4} indicates journal information.

The first four lines only consider one factor (A_{i1},A_{i2},A_{i3}, A_{i4})and it turns out that abstract using Author Topic model has the best result. When we only use keywords, the accuracy is higher than the line only using organization information. However, 0.213 is obtained when only journal information is considered. We can conclude that journal information is too general.

The rest lines we consider two factors or four factors. When abstract combines with other metadata, here it mean A_{i2},A_{i3}, A_{i4}, the accuracy promotes compared to VSM as we observe line 12 and line 5-11 in Table 1. So it's not hard to draw a conclusion that the addition of metadata is useful and should take them into consideration.

ConfigurationLine	A_{i1}	A_{i2}	A_{i3}	A_{i4}	Accuracy
1	0.9	0	0	0	0.638
2	0	0.9	0	0	0.431
3	0	0	0.9	0	0.547
4	00	0	0.9	0.213	0.123
5	0.45	0.45	0	0	0.625
6	0.45	0	0.45	0	0.697
7	0.45	0	0	0.45	0.614
8	0.4	0.4	0.05	0.05	0.636
9	0.4	0.3	0.1	0.1	0.715
10	0.5	0.2	0.15	0.05	0.758
11	0.5	0.15	0.2	0.05	0.78
12	traditional VSM				0.534

Table 1. The experimental results

It's not difficult to find out when using A_{i1} and A_{i3} together, the performance is a little better than using alone. The combination of Author Topic model and keywords can perform slightly better than other two factors combination. The last four lines (here we suppose the last line, that is VSM, is not included) give weight to every factor. We can observe that the line 11 has the best result among the four, even among all the lines. When give a lion share to A_{i1} , second is A_{i3} , A_{i2} is the third, the result standing out. When comparing the last two lines and find out that the keywords have a bigger influence than A_{i2}. In a word, Keywords improve the performance.

In Figure 2, there is an example of line 11. As it depicted in the Figure 2, the result demonstrated the authors linked to topic 9. In the middle, the first 8 words distributed in Topic 9 have been picked out. As we can see, this topic is about mining frequent pattern from large datum set. The authors with the same topic are around the Topic 9 and we pick out 6 authors, with their name and keywords from one of their papers. When compare the keywords, almost 1/2 of the items are same.

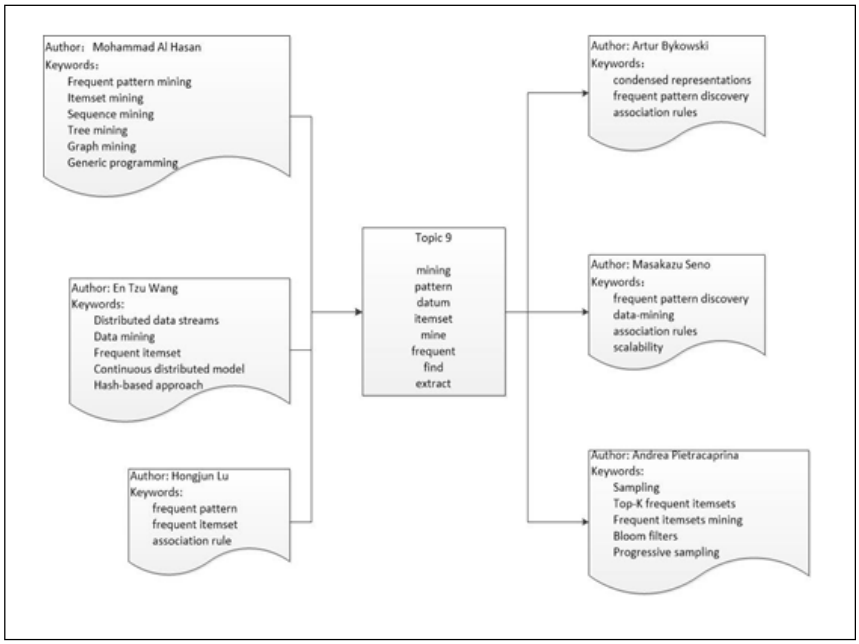

Figure 2. Authors related to Topic 9

6. Conclusion and Future work

This paper proposed a different approach to compare authors and illustrate it with experiments. First it took metadata into the model. Moreover, most of the time information is incomplete and scattered. While most of the researches focus on full texts, metadata should be made full use. it seems more relevant to exploit metadata. Clearly, Author Topic Model is more applicable to abstracts and outperforms than the traditional VSM. Furthermore, after other metadata have been added to the language model, we can see there was a noticeable change.

However, some problems need to be solved. First of all, we don't take different people of same names into consideration. In digital libraries such as Google Scholar, this type of name disambiguation is a significant and difficult problem. Second, it will be an interesting task to apply Author similarity model to social networks, such as Twitter or Facebook.

References

[1] White, H D., Griffith, B C. (1981). Author cocitation: A literature measure of intellectual structure. *Journal of the American Society for Information Science*, 32(3) 163-171.

[2] Heck ,T., Peters, I., Stock ,W G. (2011). Testing collaborative filtering against co-citation analysis and bibliographic coupling for academic author recommendation. *In*: Proceedings of the 3rd ACM RecSys' 11 Workshop on Recommender Systems and the Social Web, 16-23.

[3] Hurtado MartíN, G N., Schockaert ,S., Cornelis, C., et al (2013). Using semi-structured data for assessing research paper similarity. *Information Sciences*, 221. 245-261.

[4] Hofmann, T. (1999). Probabilistic latent semantic indexing. *In*: Proceedings of the 22nd Annual International ACM SIGIR conference on Research and development in information retrieval. *ACM*, 50-57.

[5] Tamura, Koya., Hatano, Kenji.,Yadohisa., Hiroshi (2012). A Retrieval Method Based on Language Model Considering Neighboring Contents. *Journal of Digital Information Management,* 10 (1).

[6] Rosen-Zvi, M., Griffiths, T., Steyvers, M., et al (2004). The author-topic model for authors and documents. *In*: Proceedings of the 20th Conference on Uncertainty in Artificial Intelligence. AUAI Press, 487-494.

[7] Rosen-Zvi , M, Chemudugunta, C., Griffiths, T., et al (2010). Learning author-topic models from text corpora. *ACM Transactions on Information Systems* (TOIS), 28 (1).

[8] Blei , D M., Ng, A Y., Jordan, M I . (2003). Latent dirichlet allocation. *The Journal of machine Learning Research*, 993-1022.

[9] Salton ,G,., Wong, A., Yang, C S. (1975). A Vector Space Model for Automatic Indexing. *Communications of the ACM*, 18 (11) 613-620.

[10] Minka, T P. (2001). Expectation propagation for approximate Bayesian inference. *In*: Proceedings of the Seventeenth conference on Uncertainty in artificial intelligence. Morgan Kaufmann Publishers Inc, 362-369.

[11] Griffiths, T L., Steyvers, M. (2004). Finding scientific topics, *In*: Proceedings of the National academy of Sciences of the United States of America, 101(Suppl 1): 5228-5235.

[12] Heinrich G (2005). Parameter estimation for text analysis. Web: http://www. arbylon. net/publications/text-est. pdf.

[13] Gershman, S J., Blei, D M .(2012). A tutorial on Bayesian nonparametric models. *Journal of Mathematical Psychology*, 56 (1) 1-12.

[14] McCallum A, Corrada-Emmanuel A, Wang X (2005). The author-recipient-topic model for topic and role discovery in social networks: Experiments with enron and academic email.

[15] Okolica, J S., Peterso,n G L., Mills, R F. (2006). Using author topic to detect insider threats from email traffic. Intelligence and Security Informatics. Springer Berlin Heidelberg, 642-643.

[16] Steyvers, M., Smyth, P., Rosen-Zvi M, et al (2004). Probabilistic author-topic models for information discovery. *In*: Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, 306-315.