Adapting LDA Model to Discover Author-Topic Relations for Email Analysis

Geng, Liqiang; Wang, Hao; Wang, Xin; Korba, Larry

https://doi.org/10.1007/978-3-540-85836-2_32

National Research Council Canada
Conseil national de recherches Canada

Institute for Information Technology
Institut de technologie de l'information

NRC · CNRC

Adapting LDA Model to Discover Author-Topic Relations for Email Analysis *

Geng, L., Wang, H., Wang, X., Korba, L.
September 2008

* published in the Proceedings of the 10th International Conference on Data Warehousing and Knowledge Discovery (DaWak 2008). Turin, Italy. September 1-5, 2008. NRC 50384.

Canada

Adapting LDA Model to Discover Author-Topic Relations for Email Analysis

Liqiang Geng[1], Hao Wang[1], Xin Wang[2], Larry Korba[1]

[1]Institute of Information Technology, National Research Council of Canada
Fredericton, New Brunswick, Canada
{liqiang.geng, hao.wang, larry.korba}nrc-cnrc.gc.ca
[2]Department of Geomatics Engineering, University of Calgary
Calgary, Alberta, Canada
xcwang@ucalgary.ca

Abstract. Analyzing the author and topic relations in email corpus is an important issue in both social network analysis and text mining. The Author-Topic model is a statistical method that identifies the author-topic relations. However, in its inference process, it ignores the information at the document level, i.e., the co-occurrence of words within documents are not taken into account in deriving topics. This may not be suitable for email analysis. We propose to adapt the Latent Dirichlet Allocation model for analyzing email corpus. This method takes into account both the author-document relations and the document-topic relations. We use the Author-Topic model as the baseline method and propose measures to compare our method against the Author-Topic model. We did empirical analysis based on experimental results on both simulated data sets and real Enron email data set to show that our method obtains better performance than the Author-Topic model.

1. Introduction

Identifying topics and author-topic relations in emails is an important issue in social network analysis. It adds semantics to social network analysis and provides additional perspectives for role analysis. Both supervised and unsupervised text mining techniques have been used for topic identification in emails.

When supervised learning methods are applied to identify email topics, email messages need to be labeled before the classification model is built [2, 6]. This is not a trivial task, especially without domain knowledge and context. Also generally speaking, email messages can involve any topics and it is very difficult to predefine the email topics. Clustering on "a bag of words" representation is an unsupervised learning method and thus does not require labeled training data set. However, it only assigns one email into one cluster or topic [5, 7]. Furthermore, none of the above-mentioned methods can identify topics and author-topic relation at the same time.

Statistical models for document modeling have attracted a lot of attentions in the recent years. Latent Dirichlet Allocation (LDA) was first proposed to extract topics from large text corpora [1]. LDA is a generative model that represents each document

as a mixture of probabilistic topics and represents each topic as a probabilistic distribution over words. One of the advantages of the LDA model is that this generative probabilistic model can be scaled up to introduce more levels of structure for inference [1]. Author-topic (AT) model can be considered as an extension of the LDA model by incorporating a layer of authors [9, 10]. It is the first probabilistic model to identify the topics and author-topic relations simultaneously. To tackle the efficiency issues of LDA and AT models, Gibbs sampling was proposed to estimate the parameters of the models. However, in the Gibbs sampling process for the AT model, the relations between documents and the words are not taken into account. This results in some information loss, i.e., the co-occurrence of words in the same document will be ignored in the algorithm. This is especially true when each document only involves one or very few topics, which is common in email messages. For example, if an author wrote two emails each consisting of two words as follows.

Email 1: *Computer Science*

Email 2: *Civil Engineering*

In the Gibbs sampling algorithm, these two documents will be mixed together. Co-occurrence between *computer* and *science* and that between *civil* and *engineering* will be ignored.

In this paper, we propose to adapt the LDA model in a different way to identify the author-topic relation for email analysis. The idea is that we adopt the LDA model to derive document-topic relation and then aggregate the results on authors to obtain the author-topic relation. In this way, both document-topic and author-topic relations are taken into account. We also proposed evaluation criteria for comparing the LDA and the AT models. The rest of the paper is organized as follows. Section 2 introduces the LDA and AT models and presents the adapted LDA model. In Section 3, we propose the evaluation criteria for comparing AT and the adapted LDA model. Section 4 presents the experimental results on both simulated data sets and a real data set. Section 5 concludes the paper and discusses future work.

2. Adapted LDA Model for Email Analysis

LDA is a generative statistical model that describes how words in a document might be generated on the basis of latent random variables. It assumes that a document is a multinomial distribution over topics and that a topic is a multinomial distribution over words. In the generation process, LDA first chooses a topic in terms of the probabilities of a document over topics. Then it chooses a word according to the chosen topic and the probability distribution of the topic over words. The process is repeated until the corpus is generated.

The probability of choosing a word token w_i in a particular document is

$$P(w_i) = \sum_{j=1}^{T} P(w_i \mid z_i = j) P(z_i = j) \qquad (1),$$

where $P(z_i = j)$ is the probability of topic j was sampled for word token w_i in this document; $P(w_i \mid z_i = j)$ is the probability of word w_i under topic j. T is the

number of topics. This model specifies the probability distribution over words within a document.

Let $\phi^{(j)} = P(w \mid z = j)$ refer to the multinomial distribution over words for topic j and $\theta^{(d)} = P(z)$ refer to the multinomial distribution over topics for document d. The parameters ϕ and θ indicate which words are important for a given topic and which topics are important for a particular document, respectively.

Given a document collection, the topic identification problem becomes the model fitting that finds the best estimate of the parameters ϕ and θ, i.e., the topic-word distributions and the document-topic distributions. Gibbs sampling is an efficient method to solve this model fitting problem. Gibbs sampling simulates a high-dimensional distribution by sampling on lower-dimensional subsets of variables where each subset is conditioned on the value of all other variables. The sampling is done sequentially and proceeds until the sampled values approximate the target distribution [3].

For the LDA model, the Gibbs sampling procedure considers each word token in the document collection in turn, and estimates the probability of assigning the current word token to each topic, conditioned on the topic assignments of all other word tokens. The conditioned probability is written as:

$$P(z_i = j \mid w_i, z_{-i}, w_{-i}, d_i, ...) \propto \frac{C^{WT}_{w_i j} + \beta}{\sum_{w=1}^{W} C^{WT}_{wj} + W \cdot \beta} \cdot \frac{C^{DT}_{d_i j} + \alpha}{\sum_{t=1}^{T} C^{DT}_{dt} + T \cdot \alpha} \qquad (2)$$

where $z_i = j$ represents the topic assignment of token w_i to topic j, z_{-i} refers to the topic assignments of all other word tokens, and "…" refers to all other known or observed information. T is the number of the topics, W is the number of word tokens, D is the number of documents, and α and β are prior parameters that need to be specified before the sampling process. Empirical guidelines for choosing the appropriate values for α and β are discussed in [1, 4]. A word-topic matrix C^{WT} and a topic-document matrix C^{DT} are maintained in the Gibbs sampling process to calculate the probability according to equation (2).

$$C^{WT} = \begin{bmatrix} a_{11} & a_{12} & \dots & a_{1T} \\ a_{21} & a_{22} & \dots & a_{2T} \\ \dots & \dots & & \dots \\ a_{W1} & a_{W2} & \dots & a_{WT} \end{bmatrix}_{W \times T} \qquad C^{DT} = \begin{bmatrix} b_{11} & b_{12} & \dots & b_{1T} \\ b_{21} & b_{22} & \dots & b_{2T} \\ \dots & \dots & & \dots \\ b_{D1} & b_{D2} & \dots & b_{DT} \end{bmatrix}_{D \times T}$$

Word-Topic matrix Topic-Document matrix

The word-topic matrix C^{WT} contains the number of times w_i is assigned to topic j, not including the current token of w_i; the topic-document matrix C^{DT} contains the

number of times topic j is assigned to some word token in document d, not including the current instance w_i.

After the sampling process, the estimate of parameters ϕ and θ could be obtained from the word-topic matrix and the topic-document matrix with equations (3) and (4).

$$\hat{\phi}_i^{(j)} = \frac{C_{w_i j}^{WT} + \beta}{\sum_{w=1}^{W} C_{wj}^{WT} + W \cdot \beta} \tag{3}$$

$$\hat{\theta}_j^{(d)} = \cdot \frac{C_{d_i j}^{DT} + \alpha}{\sum_{t=1}^{T} C_{dt}^{DT} + T \cdot \alpha} \tag{4}$$

The Gibbs sampling procedure is an iterative process as follows.

1. Initialize C^{WT} and C^{DT}
2. For i = 1 to N do // N is the number of Gibbs sampling iterations
3 Randomly read a word token w from documents
3. Calculate the probabilities of assigning w to topics based on equation 2.
4. Sample a topic in terms of the estimated probabilities obtained in step 3
5. Update the matrix C^{WT} and C^{DT} with new sampling results
6. Go to step 3 until all of word tokens have been scanned.
7. Endfor

The AT model is an extension of the LDA model by substituting the variable *author* for variable *document*, which means each author is associated with a multinomial distribution over topics. In the AT model each word w in a document is associated with two latent parameters: an author x, and a topic z [9].

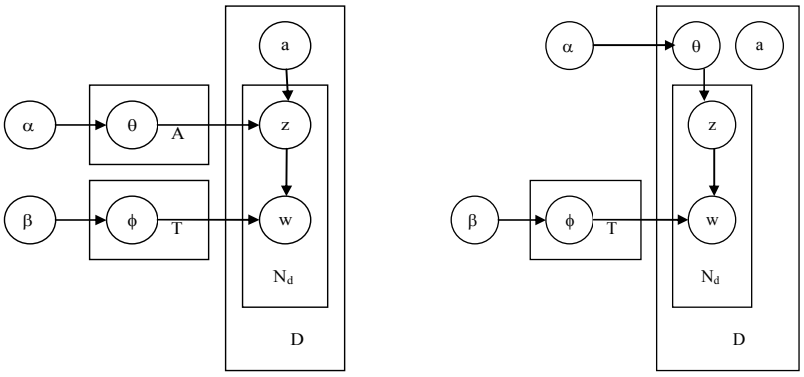
In general, one document can have more than one author. However, for email collections, usually there is only one author for one email (except for forwarded emails, where the forwarder may add more content or modify the content of the original sender). Therefore, we simplify the Gibbs sampling process for the AT model on emails by ignoring the author sampling. In this case, it is straightforward that the Gibbs sampling procedure for the AT model is equivalent to first aggregating documents on authors and then applying the LDA model on the aggregated documents.

We can see that while the AT model tries to identify the relationship between authors and topics, the relationship between documents and topics is ignored, i.e., the information about concurrence of words within a document is ignored. The author-topic relations are integrated from document-topic relation at the beginning of the Gibbs sampling process. For documents like email messages, each of which only involves one or a few topics, the ignorance of document-topic relation may deteriorate the results. We propose an adapted LDA model to derive the author and topic relationship. The idea is straightforward. First, Gibbs sampling algorithm for LDA is used to derive the document-topic relationship. Then the author-topic relationship is obtained by aggregating the document-topic matrix using the following SQL sentence:

Select *author,* sum($Topic_1$),…, *sum*($Topic_T$) from Table(C^{DT}) inner join Table(AD) on Table(C^{DT})*.document* = Table(AD)*.document* group by *author*

where Table(C^{DT}) denotes the table corresponding to the document-topic matrix. Table(AD) denotes the table representing the relationship between documents and authors.

Figure 1 compares the adapted LDA model and the AT model (Note that since we focus on email analysis, we ignore the sampling process for authors). In the adapted LDA model, the author variable is isolated from other factors, and thus is not involved in the inference process. Therefore, the inference process in the model is identical to that of the LDA model, i.e., only the document-topic relation is taken into account in the inference process. The author-document relation is used in the aggregation process after the inference process to derive the author-topic relation. This is different from the AD model in that the AD model directly uses the author-topic structure in the inference process as shown in Figure 1(a).



(a) Simplified AT model (b) Adapted LDA model

Fig. 1. Graphic models for the AT model and adapted LDA model

3. Evaluation Criteria

For simulated data, the number of the topics and the probability distributions of authors over topics and those of topics over words are known. The evaluation can be done straightforwardly by comparing the degree of match between the real distributions and the discovered distributions.

Since the AT model and the adapted LDA model are unsupervised methods and the discovered topics are randomly ordered, we used a greedy algorithm to compare the discovered topics and the originally assigned topics to determine the degree of match between them. The algorithm first compares the discovered topics and the actual topics in a pair-wise fashion. The pair with the least distance will be matched if the

distance is below a threshold. Then this pair of topics is removed from topic lists and the next round begins until the current minimum distance is above a threshold, which means that the rest of the assigned topics and the discovered topics do not match any more. Here we used 1 – cosine(T, T') as the distance measure. Figure 2 shows the procedure.

Based on the degree of match between the real and the discovered topics, we also evaluated the degree of match between real and discovered authors' distribution over topics with similar experiments. The difference here is that when we match real and discovered author distributions over topics, we need to consider the discovered topics and the real topics that do not match. For example, suppose we have three real topics t_1, t_2, t_3 and we discovered three topics t_1', t_2', t_3'. If t_1 matches t_1', t_2 matches t_2', but t_3 does not matches t_3', we need to calculate the distance based on distribution over t_1, t_2, t_3, and t_3'.

Function *TopicMatch*

Input: $T[1, n]$, $T'[1, n]$ //Real topics and identified topics as probability distribution over words, n is the number of topics.

```
Count = 0      //number of topics matched
For i = 1 to n,
      (k, l) = argmax_(i,j) (Dist(T[i], T'[j])      //find the currently best matched topics
      if dist(T[k], T'[l]) > threshold      // It is a match
            remove T[k] and T'[l] from arrays respectively
            count++;
      else     //It is not a match
            break;
endfor
degreeOfMatch = count/n
```

Fig. 2. Algorithm that calculates the degree of topic match

For the real data sets, we do not know the real topics as we do for the simulated data. Therefore, we cannot use the degree of match to evaluate the models. We use different measures to compare our method with the AT model. Perplexity can be used as a measure to indicate the prediction power of the AT and LDA models [9], but here we focus on the quality of the topics in terms of the clustering results rather than the prediction power.

To measure the intra-topic quality, we use entropy to evaluate the correlation among the words of each topic. A uniform distribution of topics over words conveys no meaning to users and thus a topic of high entropy value will be considered as low quality. On the other hand, when a topic concentrates on a small group of words, which results in a lower entropy value, we say it is a topic with higher quality. The average entropy value for a topic distribution over words is defined as

$$Entropy = \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{m} p_{ij}\log_2 p_{ij}$$, where p_{ij} denotes the probability of topic i taking word j.

To measure the inter-topic quality, we use the average minimum Kullback–Leibler divergence (or KL-distance) to evaluate how close the discovered topics are to each other. The average minimum KL distance is defined as

$KL = \frac{1}{n}\sum_{i=1}^{n} \min_{j=1,n,j\neq i} kl(p_i, p_j)$, where $kl(p_i, p_j)$ denotes the symmetric KL-distance between topic i and topic j and is defined as

$$kl(p_i, p_j) = \frac{1}{2}\left(\sum_k p_i(k)\log\frac{p_i(k)}{q_i(k)} + \sum_k q_i(k)\log\frac{q_i(k)}{p_i(k)}\right).$$

A greater KL-distance value means the topics are far away from each other and thus are desired.

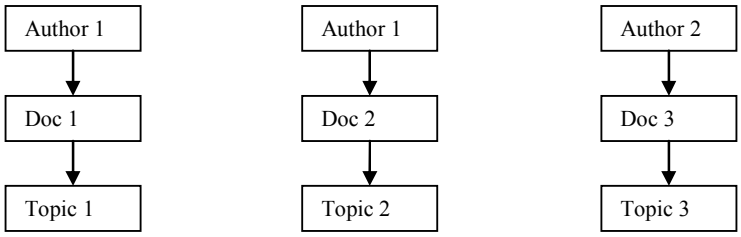
4. Experimental Results

We first did experiments on simulated data to compare the adapted LDA model and the AT model. To simplify the generating process and facilitate the comparison of the results, we assume that different topics do not share any common words and all topics have uniform distributions over the words within that topic. We assume two types of documents in terms of the document-topic structure: single-topic documents and multi-topic documents. A single-topic document is generated from words from a single topic, while a multi-topic document is generated from words from more than one topic. We also assume two types of author-topic structures: separated-author-topic structure and the mixed-author-topic structure. The separated-author-topic structure requires that any two authors either share all the topics they are involved in or share no topics at all. The mixed-author-topic structure allows authors to share some of the topics they are involved in with other authors.

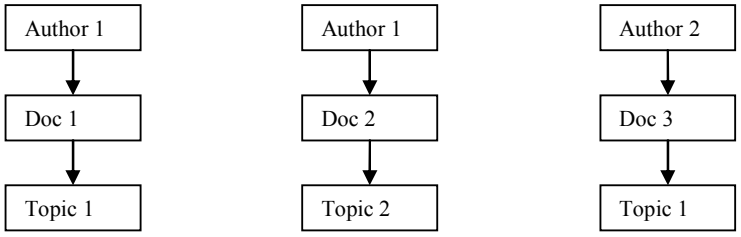
We get four combinations based on the author-topic and document-topic structures Figure 3 shows examples of the four cases.

We can see that the multi-topic documents with separated author-topic structure can be converted to multi-topic documents with separated-author-topic structure if we combine the topics under the same author to one topic. For example, in Figure 3(c), topic 1 and topic 2 can be combined as one topic and topic 3 and topic 4 can be combined as another topic. Therefore, we only consider the three other cases.

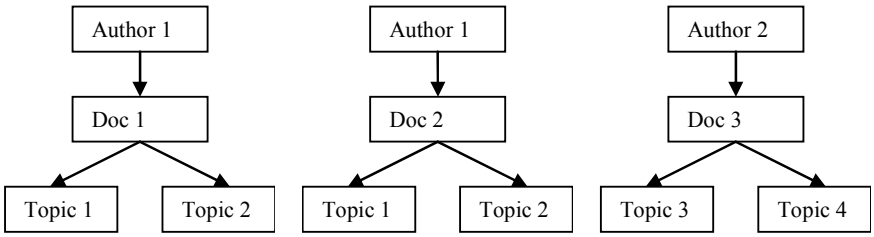
We generated three data sets to simulate the three cases respectively. Each data set consists of 5000 emails with a vocabulary of 200 words. We set 20 topics with each topic consisting of 10 words. We set 20 authors and each author has two topics. When running the AT and the adapted LDA models, we set the number of topics to 20, which means that we already know the number of topics in advance. This facilitates the comparison of the results (Some principles for choosing the appropriate number of the topics were discussed in [4]). We follow the suggestions from [9] and set $\alpha = 50/T$, and $\beta = 0.01$.



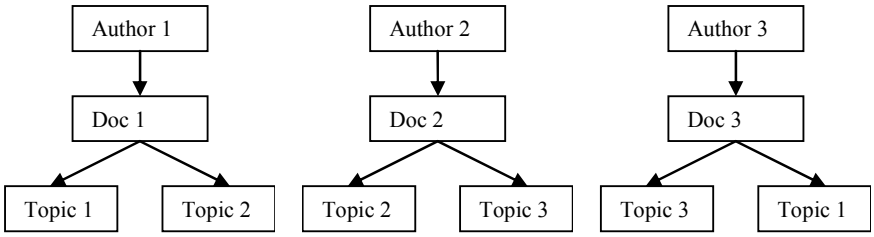
(a) Single-topic document, separated-author-topic structure



(b) Single-topic document, mixed-author topic structure



(c) Multi-topic document, single author-topic structure

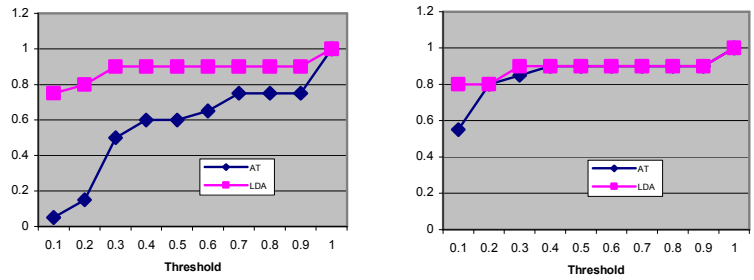


(d) Multi-topic document, mixed-author-topic structure

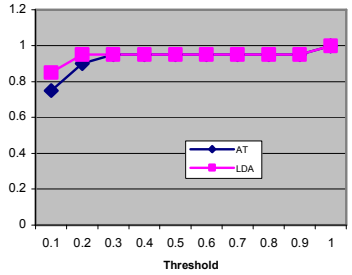
Fig. 3. Four combinations of author-topic and document-topic structure

In Figure 4, x-axis denotes the threshold for match and y-axis denotes the degree of match between real and discovered topics, as defined in Section 3. Figure 4 shows that in all three cases, the LDA model outperforms the AT model when the distance

threshold is less than 0.3. Figure 4(a) shows that for the single-topic document and separated author-topic structure, The LDA model performs much better than the AT model. This is because the AT model mixed the documents of a single author together and co-occurrence information is totally lost. Figure 4(b) shows that for the single-topic document and mixed author-topic structure, LDA still has better performance than the AT model, but not as significant as in the first case. This is because in the aggregating process in the AT model, although some co-occurrence information within a document is lost, some co-occurrence can still be embodied in the author-topic structure. Figure 4(c) shows that for the multiple-topic document and mixed author document structure, LDA just performs slightly better than the AT model. This is because in aggregation process in the AT algorithm, the loss of co-occurrence information is very limited.



(a) Single-Separated (b) Single-Mixed



(c) Multiple-Mixed

Fig. 4. Degree of match for topic-word relation for three simulated data set

We also evaluated the results based on F1 measure and get the very similar trends for all three cases. Here the precision is defined as the ratio between the number of the discovered topics whose minimum distance to the actual topics are below a threshold

and the number of the discovered topics. The recall is defined as the ratio between the number of the actual topics whose minimum distance to the discovered topics are below a threshold and the number of the actual topics.

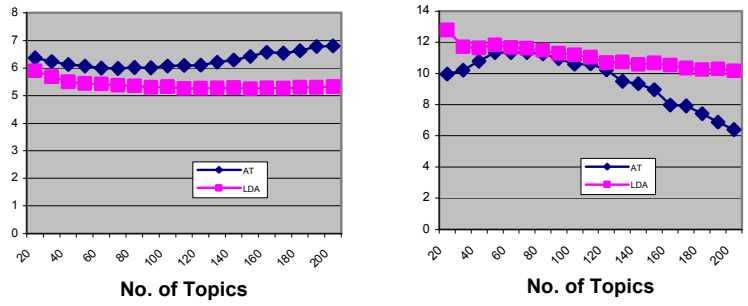
Based on the degree of match between the real and the discovered topics, we also evaluated the degree of match between real and discovered authors' distribution over topics with similar experiments. We set the distance threshold for both the topic-word distributions and the author-topic distributions to 0.3 and obtain the results as shown in Table 1.

Table 1. Degree of match for author-topic relation for three simulated data sets

%	Single-Separated	Single-Mixed	Multiple-Mixed
LDA	90	90	90
AT	55	80	90

We then did experiments on the Enron email data set [11] to compare the LDA and the AT models.

We variated the number of the topics from 20 to 200 and recorded the entropy and KL-distance measures in Figure 5. Figure 5(a) shows that when the specified number of topics increases, the average entropy of the results generated from the AT model increases, while the entropy of the results generated from the LDA model remain stable. This means that the intra-topic quality of the LDA model is relative stable to the specified number of topic and the AT model produce deteriorated results when the number of topics increase. Also the LDA model consistently produces better results than AT model in terms of intra-topic quality.



(a) Entropy (b) KL distance

Fig. 5. Comparison of the AT and LDA models on Enron email data set

Figure 5(b) shows that the LDA model consistently attains greater KL values than the AT model regardless of the number of specified topics. This means that the LDA model produces results with higher inter-topic quality. Also when the number of the specified topics increases, the difference between the KL values from the two models

increases. This means that when the number of topics increases, the inter-topic quality of the AT model deteriorate dramatically, while the LDA model remains stable.

5. Conclusion and Future work

We proposed a method to find topics and author-topic relations in emails based on the LDA model. Compared with the AT model, our method takes into account the word co-occurrence information within documents. Experimental results on both synthetic and real data sets show that the adapted LDA method obtains better results than the AT model for email corpus where each document has one author and involves only one or a few topics.

We will extend our work to identify author-recipient-topic relations based on the adapted LDA method and compare the results with the Author-Recipient-Topic model [8]. Another approach we are interested in involves taking into account the threading information in our method to facilitate the discovery of topics and author-topic relations.

Reference

[1] Blei, D.M., Ng, A.Y, and Jordan, M.I. Latent Dirichlet allocation. *Journal of Machine Learning Research*, 3:993 -1022, 2003.

[2] Dredze, M., Lau, T.A., and Kushmerick, N. Automatically classifying emails into activities. *Intelligent User Interfaces*. 70-77, Sydney, Australia, January, 2006.

[3] Gilks, W., Richardson, S., and Spiegelhalter, D. Markov Chain Monte Carlo in Practice. Chapman & Hall, New York, 1996.

[4] Griffiths, T.L. and Steyvers, M. Finding scientific topics. *Proceedings of the National Academy of Sciences of the United States of America*, 101:5228-5235, 2004.

[5] Huang, Y., Govindaraju, D., Mitchell, T.M., de Carvalho, V.R, and Cohen, W.W. Inferring ongoing activities of workstation users by clustering email. *Proceedings of the First Conference on Email and Anti-Spam*. Mountain View, California, USA, July, 2004.

[6] Khoussainov, R. and Kushmerick, N. Email task management: An iterative relational learning approach. *Proceedings of the Second Conference on Email and Anti-Spam*. Stanford University, California, USA, 2005.

[7] Li, H., Shen, D., Zhang, B., Chen, A., and Yang, Q. Adding semantics to email clustering. *Proceedings of the 6th IEEE International Conference on Data Mining*. 938-942, Hong Kong, China, 2006.

[8] McCallum, A., Wang, X., and Corrada-Emmanuel, A. Topic and role discovery in social networks with experiments on Enron and academic email. *Journal of Artificial Intelligence Research*. 30:249-272, 2007.

[9] Rosen-Zvi, M., Griggiths, T.L, Smyth, P., and Steyvers, M. Learning author topic models from text corpora. http://citeseer.ist.psu.edu/rosen-zvi05learning.html

[10] Steyvers, M., Smyth, P., Rosen-Zvi, M., Griffiths, T.L. Probabilistic author-topic models for information discovery. *Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 306-315, Seattle, USA, August, 2004.

[11] Enron email data set. http://www.isi.edu/~adibi/Enron/Enron.htm.