PREDICTING PROTEIN-PROTEIN RELATIONSHIPS FROM LITERATURE USING LATENT TOPICS

TATSUYA ASO[1] KOJI EGUCHI[1]
dango-r@cs25.scitec.kobe-u.ac.jp eguchi@port.kobe-u.ac.jp

[1] *Department of Computer Science and Systems Engineering, Kobe University, 1-1 Rokkoudai, Nada-ku, Kobe, 657-8501, Japan*

This paper investigates applying statistical topic models to extract and predict relationships between biological entities, especially protein mentions. A statistical topic model, Latent Dirichlet Allocation (LDA) is promising; however, it has not been investigated for such a task. In this paper, we apply the state-of-the-art Collapsed Variational Bayesian Inference and Gibbs Sampling inference to estimating the LDA model. We also apply probabilistic Latent Semantic Analysis (pLSA) as a baseline for comparison, and compare them from the viewpoints of log-likelihood, classification accuracy and retrieval effectiveness. We demonstrate through experiments that the Collapsed Variational LDA gives better results than the others, especially in terms of classification accuracy and retrieval effectiveness in the task of the protein-protein relationship prediction.

Keywords: Biomedical text mining; probabilistic topic models.

1. Introduction

There have been increasing demands for organizing knowledge accumulated in documents and then generating potential hypotheses in biomedical fields. This paper focuses on the task to predict relationships between biological entities. Research trends on the biomedical relationship extraction can be categorized into: (1) methods using manually or automatically generated templates, (2) methods based on natural language processing, and (3) statistical co-occurrence-based methods [1, 2]. This paper focuses on the third approaches targeting a specific type of biomedical entities, proteins. While the natural language processing-based approaches usually extract entity relationships within a document, statistical methods are based on co-occurrence of biomedical entities or their related statements in a set of documents to extract relationships between the entities. Statistical topic models are promising for this objective.

Statistical topic models (e.g., [3, 4]) are based on the idea that documents are mixtures of topics, where a topic is a probability distribution over words, in order to capture semantics or to achieve dimensionality reduction. "Probabilistic Latent Semantic Analysis" (pLSA) [5], proposed by Hoffman, can model underlying topics for given documents; however, it cannot model the topics for *unseen* documents that were not used for parameter estimation. Blei et al. [3] proposed one of the

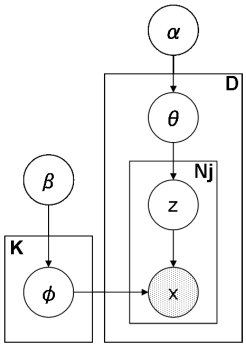


Fig. 1. The graphical model of LDA.

topic models called "Latent Dirichlet Allocation" (LDA) in an extension of pLSA, introducing a Dirichlet prior on a multinomial distribution over topics for each document. This makes the model applicable to unseen documents. The LDA model has been accepted in various fields; however, it has not been investigated for predicting biological entity relationships, to our knowledge. In this paper, we investigate applying the LDA model to extract and predict protein-protein relationships from biomedical literature. In the statistical topic modeling, a set of topics are usually assumed to be unobserved in a document collection, and so we need to infer such unknown distributions from the documents. To estimate the LDA model, "Collapsed Gibbs Sampling inference"[a] method can be used [4]. "Collapsed Variational Bayesian inference" (CVB) [6] is alternative approach to estimate the LDA model.

The focus of this paper is to investigate how to apply the LDA model to the task of protein-protein relationship prediction from biomedical literature, and to evaluate, in an extrinsic manner, the effectiveness over different model estimation methods.

2. LDA and Estimation Algorithms

2.1. *Generative Process of LDA*

Figure 1 shows the graphical model of LDA. We formally describe generative process of LDA [3], as follows,

(1) For all j documents sample $\theta_j \sim Dir(\alpha)$
(2) For all k topic sample $\phi_k \sim Dir(\beta)$
(3) For each of the N_j words x_i in document d_j
   (a) Sample a topic $z_i \sim Mult(\theta_j)$
   (b) Sample a word $x_i \sim Mult(\phi_{z_i})$

[a]It is sometimes simply called "Gibbs Sampling inference" [4].

where N_j is the total number of words in document j. θ and ϕ indicate a per-document topic distribution and a per-topic word distribution, respectively; and α and β indicate hyperparameters that specify Dirichlet priors corresponding to θ and ϕ, respectively.

Joint distribution of all the random variables and parameters in the LDA model are given by the following equation:

$$p(\mathbf{x}, \mathbf{z}, \theta, \phi | \alpha, \beta) = \prod_{j=1}^{D} \frac{\Gamma(K\alpha)}{\Gamma(\alpha)^K} \prod_{k=1}^{K} \theta_{jk}^{\alpha - 1 + n_{jk\cdot}} \times \prod_{k=1}^{K} \frac{\Gamma(W\beta)}{\Gamma(\beta)^W} \prod_{w=1}^{W} \phi_{kw}^{\beta - 1 + n_{\cdot kw}} \quad (1)$$

where n_{jkw} is the number of word w assigned to topic k in document j, and '$\cdot$' means a corresponding index is marginalized. In other words, $n_{\cdot kw} = \sum_j n_{jkw}$ and $n_{jk\cdot} = \sum_w n_{jkw}$. D, K and W indicate the number of documents, the number of topics, and the size of the entire vocabulary, respectively.

Given the observed words $\mathbf{x} = \{x_{ij}\}$, the task of Bayesian inference is to compute the posterior distribution over the latent topic variable $\mathbf{z} = \{z_{ij}\}$, the per-document topic distribution $\theta = \{\theta_j\}$ and per-topic word distribution $\phi = \{\phi_k\}$.

2.2. *Collapsed Gibbs Sampling Inference*

"Collapsed" Gibbs Sampling inference [4] uses the marginalized distribution over $\mathbf{x}$ and $\mathbf{z}$, as follows:

$$p(\mathbf{z}, \mathbf{x} | \alpha, \beta) = \prod_{j} \frac{\Gamma(K\alpha)}{\Gamma(K\alpha + n_{j\cdot\cdot})} \prod_{k} \frac{\Gamma(\alpha + n_{jk\cdot})}{\Gamma(\alpha)} \times \prod_{k} \frac{\Gamma(W\beta)}{\Gamma(W\beta + n_{\cdot k\cdot})} \prod_{w} \frac{\Gamma(\beta + n_{\cdot kw})}{\Gamma(\beta)} \quad (2)$$

Given the current state of all except one topic assignment to a word x_{ij}, the conditional probability of z_{ij} is given by:

$$p(z_{ij} = k | \mathbf{z}^{\neg ij}, \mathbf{x}, \alpha, \beta) = \frac{(\alpha + n_{jk\cdot}^{\neg ij})(\beta + n_{\cdot kx_{ij}}^{\neg ij})(W\beta + n_{\cdot k\cdot}^{\neg ij})^{-1}}{\sum_{k'=1}^{K} (\alpha + n_{jk'\cdot}^{\neg ij})(\beta + n_{\cdot k'x_{ij}}^{\neg ij})(W\beta + n_{\cdot k'\cdot}^{\neg ij})^{-1}} \quad (3)$$

where $n^{\neg ij}$ corresponds to variables or counts excluding x_{ij} and z_{ij}. The conditional probability specified by Eq. (3) can be used to carry out the Collapsed Gibbs Sampling inference.

2.3. *Collapsed Variational Bayesian Inference*

Very recently, Collapsed Variational Bayesian inference (CVB) was proposed and applied to estimate the LDA model [6]. According to [6], this section briefly describes the CVB method. The CVB method is an algorithm that improves the estimation accuracy in an extension of Variational Bayesian inference (VB) [3]. The CVB method models the dependence of the parameters on the latent variables, instead of assuming independence. The only assumption made in the CVB method

is that the latent variables z are mutually independent, thus the posterior can be approximated as:

$$q(\mathbf{z}, \theta, \phi) = q(\theta, \phi|\mathbf{z}) \prod_{ij} q(z_{ij}|\gamma_{ij}) \tag{4}$$

where $q(z_{ij}|\gamma_{ij})$ is multinomial distribution with parameter γ_{ij}. The variational free energy can be simplified to:

$$F(q(\mathbf{z})) = \min_{q(\theta,\phi|\mathbf{z})} F(q(\mathbf{z})q(\theta, \phi|\mathbf{z})) = E_{q(\mathbf{z})}[-\log p(\mathbf{x}, \mathbf{z}|\alpha, \beta)] - H(q(\mathbf{z})) \tag{5}$$

Minimizing Eq. (5) with respect to the variational parameters γ_{ijk}, we get:

$$\gamma_{ijk} = q(z_{ij} = k) = \frac{\exp(E_{q(\mathbf{z}^{\neg ij})}[p(\mathbf{x}, \mathbf{z}^{\neg ij}, z_{ij} = k|\alpha, \beta)])}{\sum_{k'=1}^{K} \exp(E_{q(\mathbf{z}^{\neg ij})}[p(\mathbf{x}, \mathbf{z}^{\neg ij}, z_{ij} = k'|\alpha, \beta)])} \tag{6}$$

Using Eq. (2), expanding $\log \frac{\Gamma(\eta+n)}{\Gamma(\eta)} = \sum_{l=0}^{n-1} \log(\eta + l)$ for positive real values η and positive integers n, we get:

$$\gamma_{ijk} = \frac{\exp(E_{q(\mathbf{z}^{\neg ij})}[\log(\alpha + n_{jk\cdot}^{\neg ij}) + \log(\beta + n_{\cdot kx_{ij}}^{\neg ij}) - \log(W\beta + n_{\cdot k\cdot}^{\neg ij})])}{\sum_{k'} \exp(E_{q(\mathbf{z}^{\neg ij})}[\log(\alpha + n_{jk'\cdot}^{\neg ij}) + \log(\beta + n_{\cdot k'x_{ij}}^{\neg ij}) - \log(W\beta + n_{\cdot k'\cdot}^{\neg ij})])} \tag{7}$$

3. Protein-Protein Relationship Prediction based on LDA

The LDA model can represent semantics or concepts that appear in a document. Therefore, it can be applied to compute likelihood that an entity is related to another entity. Using the LDA model, similarity between a pair of entities can be computed by the following equation.

$$Sim1(e_i, e_j) = p(e_i|e_j)/2 + p(e_j|e_i)/2. \tag{8}$$

where $p(e_i|e_j)$ is obtained using latent topic k by:

$$p(e_i|e_j) = \sum_k p(e_i|k)p(k|e_j). \tag{9}$$

In [7, 8], Eq. (8) was used to compute similarity between social entities that appear in newspaper articles; however, we believe that this similarity computation has some problems. One of the problems is that the similarity largely depends on a small number of frequently appearing entities. For instance, even when the value of $p(e_i|e_j)$ is very large and $p(e_j|e_i)$ is almost zero, the final similarity given by Eq. (8) is still large because this similarity is computed by averaging these two conditional probabilities.

To address such problems, we define a new method to compute similarity between a pair of entities, as follows:

$$Sim2(e_i, e_j) = p(e_i|e_j) \times p(e_j|e_i). \tag{10}$$

This equation indicates joint probability of $P(e_i|e_j)$ and $P(e_j|e_i)$ assuming that these are independent of each other.

4. Data and Entity Representation

In this section, we briefly explain GENIA collection and TREC collection that we used for our experiments. Table 1 shows a summary of these datasets.

4.1. *GENIA Collection*

GENIA collection[b] is a subset of MEDLINE formatted in XML, and entities such as proteins are manually tagged. We removed standard 418 stop words used in "InQuery system" [9], and words and entities that were observed in less than 10 documents (abstracts).

4.2. *TREC Collection and GENIA Tagger*

Another data collection is TREC collection that was used in TREC Genomic Track from 2004 to 2005, formatted in XML. We extracted titles and abstracts of documents from the collection. First, we used a subset of documents (abstracts) in which values in "PubData" and "DataComplete" fields are 2002 for training, and another subset of documents in which values of these fields are 2003 for testing in Section 5.2. Second, we removed standard 418 stop-words from the training and test data, and removed words and entities that were observed less than 10 documents from the training data. In TREC collection, entities are not tagged and thus we identified biomedical entities using GENIA tagger[c].

Table 1. Datasets extracted from GENIA and TREC collection.

	Data item name	GENIA	TREC(2002)	TREC(2003)
D	The number of documents (abstracts)	2000	33000	31000
W	The number of vocabulary words	1959	16879	183710
E	The number of vocabulary entities	229	1897	58430
W_{freq}	The total frequency of words	107532	3501405	3863255
E_{freq}	The total frequency of entities	9733	48457	171056

5. Experiments

5.1. *Log-Likelihood*

We describe the way to compute the log-likelihood of each estimated model. The larger per-word test-set log-likelihood of estimated model is, the higher the performance of the model is. In this experiment, we computed the log-likelihood of two models. One model was estimated by Collapsed Gibbs Sampling inference, and the other was estimated by Collapsed Variational Bayesian Inference (CVB). In the

[b]http://www-tsujii.is.s.u-tokyo.ac.jp/GENIA/home/wiki.cgi
[c]http://www-tsujii.is.s.u-tokyo.ac.jp/GENIA/tagger/

experiments, we used GENIA collection and 2002's dataset extracted from TREC collection. We randomly divided each dataset into 90% and 10%, and we used the former as training data, and the latter as test data. When we compute the log-likelihood, we set the number of topic of LDA as $K = 10$. For the hyperparameters of Dirichlet priors, we set $\alpha = 0.1$ and $\beta = 0.1$ [6] for GENIA collection, and $\alpha = 50/K$ and $\beta = 0.1$ [8] for the dataset from TREC collection.

For Collapsed Gibbs Sampling inference (GS), given S samples from the posterior, the log-likelihood is given by:

$$p(\mathbf{x^{test}}) = \prod_{ij} \sum_{k} \frac{1}{|S|} \sum_{s=1}^{S} \theta^s_{jk} \phi^s_{kx^{test}_{ij}} \tag{11}$$

where θ^s_{jk} and ϕ^s_{kw} are computed by:

$$\theta^s_{jk} = \frac{\alpha + n^s_{jk\cdot}}{K\alpha + n^s_{j\cdot\cdot}} \quad \phi^s_{kw} = \frac{\beta + n^s_{\cdot kw}}{W\beta + n^s_{\cdot k\cdot}}. \tag{12}$$

For the CVB method, the log-likelihood is computed by:

$$p(\mathbf{x^{test}}) = \prod_{ij} \sum_{k} \bar{\theta}_{jk} \bar{\phi}_{kx^{test}_{ij}} \tag{13}$$

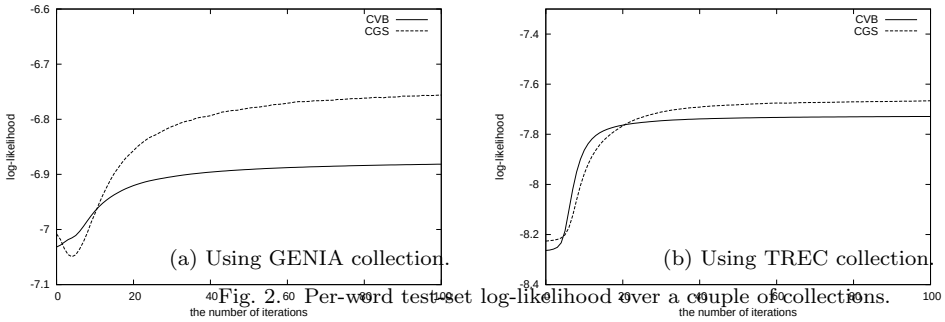
where $\bar{\theta}_{jk}$ and $\bar{\phi}_{kw}$ are computed by:

$$\bar{\theta}_{jk} = \frac{\alpha + E_q[n_{jk\cdot}]}{K\alpha + E_q[n_{j\cdot\cdot}]} \quad \bar{\phi}_{kw} = \frac{\beta + E_q[n_{\cdot kw}]}{W\beta + E_q[n_{\cdot k\cdot}]}. \tag{14}$$

The horizontal axis in Figure 2 (a) and (b) correspond to the number of iterations, the vertical axis corresponds to per-word log-likelihood. The log-likelihood in Figure 2 (a) was obtained by running each of the two inference algorithms 50 times with different random initializations ($S = 50$) and averaging the resulting test-set log-likelihood. Moreover, Figure 3 indicates histograms of per-word test-set log-likelihood across 50 random initializations, obtained by the two inference algorithms using GENIA collection. Comparing with the log-likelihood using Collapsed Gibbs Sampling (GS), that using the CVB method was found larger from 10 to 20 iterations; however, it is smaller when more iterations are performed. This indicates the Gibbs sampling method works better than the CVB method when a sufficient number of iterations are carried out, from a viewpoint of log-likelihood (or perplexity) of the estimated models. This result is similar to that was tested using non-biomedical collections [6]. However, in the following section, we demonstrate that our task-based extrinsic evaluation is not the case.

5.2. *Entity-Link Prediction*

For evaluation of protein-protein relationship prediction task, we predicted protein-protein pairs using the LDA model estimated by Collapsed Gibbs Sampling and that by the CVB method, and the pLSA model, with a couple of different similarity computations that were described in Section 3. In this experiment, we used EM algorithm to estimate pLSA.



(a) Using GENIA collection. (b) Using TREC collection.

Fig. 2. Per-word test-set log-likelihood over a couple of collections.

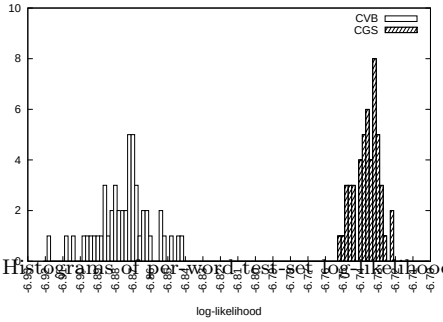


Fig. 3. Histograms of per-word test-set log-likelihood over GENIA collection.

5.2.1. *Experimental Settings*

We used the datasets extracted from TREC collection: 2002's dataset for training and 2003's dataset for testing, as described in Table 1. We generated two types of entity-entity datasets. One is "true pair" dataset, in which each pair did not co-occur within any document in the training data, but co-occurred within at least one document in the test data. We removed the entity pairs, one of which entities did not appear in any document in the training data. The other is "false pair" dataset, in which each pair never co-occurred within any document both in the training data and the test data. The number of true pairs is equal to that of false pair, as $M = 15494$.

In the experiments, we set five different numbers of topics for the LDA model and the pLSA model: $K = 10$, 50, 100 and 300. For the hyperparameters of Dirichlet prior distributions, we set $\alpha = 50/K$ and $\beta = 0.1$ in both algorithms of LDA model estimation.

5.2.2. *Task-based Evaluation*

We computed similarity of M true pairs and M false pairs, and ranked $2M$ total pairs in descending order of the similarity. We then assumed the top M entity-entity

Table 2. Classification Accuracy.

	CGS-LDA	CVB-LDA	pLSA
K=10	0.6318	0.6310	0.6075
K=50	0.5359	0.6434	0.5829
K=100	0.5669	0.6383	0.5648
K=300	0.5504	0.6317	0.5293

Table 3. Average Precision.

	CGS-LDA	CVB-LDA	pLSA
K=10	0.6745	0.6651	0.6347
K=50	0.6574	0.6895	0.5977
K=100	0.6443	0.6905	0.5606
K=300	0.6262	0.6890	0.5308

Table 4. Comparison of two different similarity computations (in the case of CVB-LDA).

	ClassificationAccuracy		AveragePrecision	
	Sim1	Sim2	Sim1	Sim2
K=10	0.6310	0.6351*	0.6651	0.6719*
K=50	0.6434	0.6467*	0.6895	0.6947*
K=100	0.6383	0.6398*	0.6905	0.6940*
K=300	0.6317	0.6305*	0.6890	0.6892

Note: "*" indicates that the result of Sim2 was 0.05 level significant via Wilcoxon signed rank test, compared to the result of Sim1 with the same topic number.

pairs to be positive, and the other M pairs to be negative. We computed classification accuracy that is given by the proportion of true-positive and false-negative pairs, changing the parameter K setting, and compared the classification accuracy. In the experiments, we used two methods to compute the similarity between entity pairs: one is the existing method presented in Eq. (8) and the other is our method presented in Eq. (10), and compared the results of these methods.

Table 2 shows the classification accuracy results with each parameter. In this table, we found that the classification accuracy of the CVB-based LDA worked best when $K = 50$. The LDA based on CVB worked better than that based on the Collapsed Gibbs Sampling (CGS) and than pLSA, with 0.05 level significance via Wilcoxon signed rank test in case of using the optimal topic number in each model.

We further evaluated using average precision [10], which is widely used to evaluate information retrieval. Average precision was computed by averaging precision at each rank of true-positive entity pairs. In this paper, we sometimes refer to average precision as "ranking effectiveness" in contrast to classification accuracy. Table 3 shows the evaluated results of average precision with each parameter. Ranking effectiveness of LDA based on CVB worked best when we set $K = 100$. The results of LDA based on CVB were better than the results of Collapsed Gibbs Sampling inference and than the results of pLSA, both with statistical significance in the same manner as in the last paragraph.

Table 4 shows that our proposed similarity computation worked better than the previous one in case of LDA based on CVB, at a statistically significant level of 0.05 in almost every test.

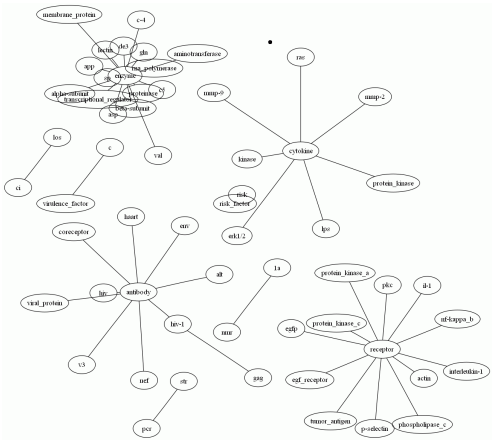


Fig. 4. An example entity-entity network

5.2.3. *Protein-Protein Relationship Network*

In the previous section, we investigated how well knowledge contained in the test data in the year of 2003 can be predicted using the LDA model that was estimated over the training data in the year of 2002. Figure 4 shows an example of the predicted protein-protein network using the LDA model with the best condition that was determined by the results in Section 5.2.2. This network consists of top-ranked 50 entity-entity pairs in order of similarity. In the network, vertices represent protein names. As for the length of edge, the tighter the relationship between a pair of vertices connected to an edge is, the shorter the edge will be.

6. Conclusions

In this paper, we presented how to generate hypotheses, especially on protein-protein relationships, from the biomedical literature using the LDA model, and compared the LDA model estimated via different inference approaches and the pLSA model. For the LDA model, we investigated a couple of model estimation methods: one is Collapsed Variational Bayesian method and the other is Collapsed Gibbs Sampling method. We found that both LDA models worked better than pLSA in terms of classification accuracy and ranking effectiveness.

We further found that the LDA based on Collapsed Variational Bayesian method worked better than that based on Collapsed Gibbs Sampling method in terms of the evaluation measures mentioned above. We also proposed a new method to compute similarity between entities using the LDA model, and indicated that the proposed method worked better than the previous method. Since our approaches do not require specific knowledge on target entities, other various types of entities can also be targeted.

The above-mentioned findings with a couple of inference approaches for the LDA model were opposite to what was found in terms of test-set log-likelihood, and thus it

suggests that intrinsic evaluation, such as using test-set log-likelihood or perplexity, does not always indicate the same findings as task-based extrinsic evaluation. One of the future tasks is to investigate why this happens, theoretically. Another task is to combine the Collapsed Variational Bayesian LDA model with non-cooccurrence-based approaches, such as based on natural language processing. Examining how the topic-based method is actually useful to protein-protein interaction prediction is also left to future work.

Acknowledgements

This work was partially supported by the Grant-in-Aid for Scientific Research (B) (#20300038) from the Ministry of Education, Culture, Sports, Science and Technology of Japan.

References

[1] A. M. Cohen and W. R. Hersh. A survey of current work in biomedical text mining. *Briefings in Bioinformatics*, 6(1):57–71, 2005.

[2] Deyu Zhou, and Yulan He, Extracting Interactions between Proteins from the Literature. *Journal of Biomedical Informatics, No.41*, pp. 393–407 (2008).

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

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

[5] T.Hofmann, Probabilistic latent semantic indexing. *Proceeding of the 22nd International Conference on Research and Development in Information Retrieval, Berkeley, California, USA*, pp. 50–57 (1999).

[6] Y. W. Teh, D. Newman, and M. Welling. A collapsed variational Bayesian inference algorithm for latent Dirichlet allocation. *Advances in Neural Information Processing Systems*, 19:1353–1360, 2007.

[7] D. Newman, C. Chemudugunta, P. Smyth, and M. Steyvers. Statistical entity-topic models. In *Proc. of ACM SIGKDD 2006*, pages 680–686, 2006.

[8] M.Steyvers and T.Griffiths, Handbook of Latent Semantic Analysis. *chapter 21:Probabilistic Topic Models, Lawrence Erbaum Associates, Mahwah, New Jersey, London* (2007).

[9] James P. Callan, W. Bruce Croft and Stephen M. Harding, The INQUERY Retrieval System. *Proceedings of the 3rd International Conference on Database and Expert Systems Applications, Valencia, Spain*, pp. 78–83 (1992).

[10] Ricardo Baeza-Yates, and Berthier Ribeiro-Neto, Modern Information Retrieval. *Addison-Wesley* (1999).

[11] S. Deerwester, S. T. Dumais, G. W. Furnas, T. K. Landauer, R. Harshman, Indexing by Latent Semantic Analysis. *Journal of the Society for Information Science*, 41(6), pp. 391–407 (1990).

[12] Ramin Homayouni, Kevin Heinrich, Lai Wei, and Michael W. Berry, Gene clustering by Latent Semantic Indexing of MEDLINE Abstracts. *Bioinformatics, Vol.21, No.1*, pp. 104–115 (2005).

[13] Hyunsoo Kim, Haesun Park, and Barry L Drake, Extracting Unrecognized Gene Relationships from the Biomedical Literature via Matrix Factorizations. *BMC Bioinformatics, Vol.8 (Suppl 9), No.S6* (2007).