Extracting Representative Words of a Topic Determined by Latent Dirichlet Allocation

Toshiaki Funatsu
Graduate School of Information Science
and Electrical Engineering,
Kyushu University
funatsu@nlp.inf.kyushu-u.ac.jp

Yoichi Tomiura
Faculty of Information Science and Electrical Engineering
Kyushu University,
tom@inf.kyushu-u.ac.jp

Emi Ishita
Research and Development Division,
Kyushu University Library,
Kyushu University
Ishita.emi.982@m.kyushu-u.ac.jp

Kosuke Furusawa
Graduate School of Information Science
and Electrical Engineering,
Kyushu University
furusawa@nlp.inf.kyushu-u.ac.jp

***Abstract*—Determining the topic of a document is necessary to understand the content of the document efficiently. Latent Dirichlet Allocation (LDA) is a method of analyzing topics. In LDA, a topic is treated as an unobservable variable to establish a probabilistic distribution of words. We can interpret the topic with a list of words that appear with high probability in the topic. This method works well when determining a topic included in many documents having a variety of contents. However, it is difficult to interpret the topic just using conventional LDA when determining the topic in a set of article abstracts found by a keyword search, because their contents are limited and similar. We propose a method to estimate representative words of each topic from an LDA result. Experimental results show that our method provides better information for interpreting a topic than LDA does.**

Keywords-LDA; topic analysis; Gibbs sampling.

I. INTRODUCTION

Web search engines are very widely used. Users are able to access different information resources easily using keywords for a search. Academic information retrieval systems have also become common and popular. As academic research disciplines have become more specific or more interdisciplinary, users who search related documents need narrow or focused topics. However, a keyword search is often not able to address this need. When users use very specific words as search terms, they generally obtain only a few search results. On the other hand, when they use general words as search terms, they obtain many search results. In this case, it is time-consuming to select relevant documents from search results.

Therefore, the following retrieval support system is useful when a user searches academic papers related to narrow or focused topics: (1) the user retrieves academic papers with generalized keywords, (2) the system does a topic analysis of the abstracts found in the search and presents some information about their topics to the user, (3) the user chooses a particular topic among them, and (4) the system narrows down the search results to academic papers that mainly contain that topic. Some methods perform a keyword article search using the feedback of the latent topic [2] or search with a novel topic model that organizes articles using the author information [3].

Latent Dirichlet Allocation (LDA) is a well-known method for topic analysis. In LDA, a topic is treated as a latent variable for determining probabilities of words. The user is able to understand a topic based on a list of words that appear with high probability in the topic. However, when a keyword search yields results with similar content, it may be difficult to understand a topic with the word list presented by LDA. The word list contains many unnecessary words for expressing a topic. Then, we consider that there are two types of words in the list. One is a word expressing the content of a topic, and the other is a word attendant to the first type of word. We call the first type a representative word of a topic. In this paper, we propose a method for identifying representative words of a topic from the word list acquired by LDA to help the user to understand the topic.

Our method first constructs a set of documents for each

 ISBN: 978-1-61208-329-2

topic that contains only words that LDA assigns the topic to, and next identifies a representative word for each topic and each document. We assume that the representative word of a document generates the other words in that document. We use Gibbs sampling to identify the representative word of each document. The higher the probability that the word *w* represents a document of topic *t*, the more representative *w* is of *t*.

In Section II, we discuss some related studies and explain the model underlying our method. In Section III, we propose the model of our method. In Section IV, we discuss an experiment that compares the results of LDA and to those of our method.

II. RELATED STUDIES

LDA is a generative probabilistic model of a corpus [1]. LDA assumes that each document has a probability distribution over topics and each topic has a probability distribution over words. Its generative process for a document in a corpus is as follows:

For each word in the document,

a) choose a topic *t* according to the probability distribution over topics that the document has;

b) choose a word *w* according to the probability distribution over words that the topic *t* has.

Blei et al. estimate the parameters using the variational Bayesian method [1]. Griffiths et al. analyze topics in a document based on LDA, but they use Gibbs sampling in parameter estimation [4].

We define our notation as follows:

M : number of documents,

$w_n^{(m)}$: the n-th word in the m-th document,

$\mathbf{w}^{(m)} = (w_1^{(m)}, w_2^{(m)}, \cdots, w_{N_m}^{(m)})$: the m-th document,

$\mathbf{w} = (\mathbf{w}^{(1)}, \mathbf{w}^{(2)}, \cdots, \mathbf{w}^{(M)})$: set (sequence) of documents,

$z_n^{(m)}$: latent variable expressing a topic to be assigned to the word $w_n^{(m)}$,

$\mathbf{z}^{(m)} = (z_1^{(m)}, z_2^{(m)}, \cdots, z_{N_m}^{(m)})$,

$\mathbf{z} = (\mathbf{z}^{(1)}, \mathbf{z}^{(2)}, \cdots, \mathbf{z}^{(M)})$,

K : number of topics,

$\theta_t^{(m)}$: probability of words with topic t in the m-th document,

$\theta^{(m)} = (\theta_1^{(m)}, \theta_2^{(m)}, \cdots, \theta_K^{(m)})$,

$\theta = (\theta^{(1)}, \theta^{(2)}, \cdots, \theta^{(M)})$,

V : number of words (by type),

$\phi_w^{(t)}$: occurrence probability of word w from topic t,

$\phi^{(t)} = (\phi_1^{(t)}, \phi_2^{(t)}, \cdots, \phi_V^{(t)})$,

$\phi = (\phi^{(1)}, \phi^{(2)}, \cdots, \phi^{(K)})$.

The joint probability of w and z in LDA is expressed as

$$p(\mathbf{w}, \mathbf{z} \mid \theta, \phi) = \prod_{m=1}^{M} \prod_{n=1}^{N_m} p(z_n^{(m)} \mid \theta^{(m)}) p(w_n^{(m)} \mid z_n^{(m)}, \phi) \,. \quad (1)$$

$$p(t \mid \theta^{(m)}) = \theta_t^{(m)}, \; p(w \mid t, \phi) = \phi_w^{(t)}$$

The prior distribution of $\theta^{(m)}$ is the dimensionality K-1 of the Dirichlet distribution with parameter α :

$$p(\theta^{(m)} \mid \alpha) = \frac{\Gamma(K\alpha)}{\Gamma(\alpha)^K} \prod_{k=1}^{K} \theta_k^{(m)} \,. \quad (2)$$

The prior distribution of $\phi(t)$ is the dimensionality V-1 of the Dirichlet distribution with parameter β :

$$p(\phi^{(t)} \mid \beta) = \frac{\Gamma(V\beta)}{\Gamma(\beta)^V} \prod_{w=1}^{V} \phi_w^{(t)} \,. \quad (3)$$

The probability of $\mathbf{z}$ given the set of documents $\mathbf{w}$ and the hyper parameters α and β is obtained via

$$p(\mathbf{z} \mid \mathbf{w}, \alpha, \beta) \propto \prod_{m=1}^{M} \frac{\Gamma(K\alpha)}{\Gamma(\alpha)^K} \frac{\prod_{k=1}^{K} \Gamma(n_{DZ}(m,k;\mathbf{z}) + \alpha)}{\Gamma(n_{DZ}(m,*;\mathbf{z}) + K\alpha)} \times \prod_{k=1}^{K} \frac{\Gamma(V\beta)}{\Gamma(\beta)^V} \frac{\prod_{w=1}^{V} \Gamma(n_{ZW}(k,w;\mathbf{w},\mathbf{z}) + \beta)}{\Gamma(n_{ZW}(k,*;\mathbf{w},\mathbf{z}) + V\beta)} \,, \quad (4)$$

where $n_{DZ}(m, k ; \mathbf{z})$ is the number of times a word from the m-th document is assigned to topic k in $\mathbf{z}$, $n_{ZW}(k, w ; \mathbf{w}, \mathbf{z})$ is the number of times word w is assigned to topic k in ($\mathbf{w}$, $\mathbf{z}$), and $n_{DZ}(m, * ; \mathbf{z})$ and $n_{ZW}(k, * ; \mathbf{w}, \mathbf{z})$ are

$$n_{DZ}(m,*;\mathbf{z}) = \sum_{k=1}^{K} n_{DZ}(m,k;\mathbf{z}) = N_m \,,$$
$$n_{ZW}(k,*;\mathbf{w},\mathbf{z}) = \sum_{w=1}^{V} n_{ZW}(k,w;\mathbf{w},\mathbf{z}) \,. \quad (5)$$

In our study, we use the results of topic analysis to estimate representative words for each topic through Gibbs sampling [4].

Blei et al. [5] studied a method for extracting significant multi-word expressions for a topic from the results of topic analysis using the procedure in [1]. Our approach is different from this. We provide representative words of a topic as useful information for understanding the topic. We do not analyze multi-word expressions, but simply treat multi-word expressions in Wikipedia entries as single words in the preprocessing in our experiment, because multi-word expressions help users to understand topics. Our approach of estimating representative words of a topic can be applied to the results of topic analysis

 ISBN: 978-1-61208-329-2

with the procedure of [5].

III. OUR PROPOSED METHOD

LDA works well for a document set that is large and has a variety of contents. It is also necessary to be able to predict topics contained in a document set to some extent so as to identify a topic from a list of words that appear with high probability in the topic. However, for a set of abstracts obtained by a keyword search, it may be difficult to identify a topic with the word list presented by LDA.

TABLE I. REPRESENTATIVE WORD SETS MATCHED TO TOPICS BY LDA

Topic	Representative Word Set Express Topic
1	retrieval model information framework space task theory problem vector process method concept question similarity modeling
2	ir retrieval text paper language issue indexing evaluation xml image processing area application research discussion
3	system information user retrieval study paper process interaction time management knowledge result tag case performance
4	method datum algorithm structure information problem feature data number analysis classification technique music network combination
5	document query collection term retrieval concept approach relevance context result technique feedback performance analysis ir
6	library computer system use access information service storage science index technology labor description resource program
7	search web information user engine approach result need content domain page use ontology interest strategy
8	information retrieval research field development multimedia application technique technology machine researcher type book tool form
9	ir word experiment retrieval work evaluation text function term performance trec measure set system graph
10	database author protocol scheme problem pir record server report privacy communication software requirement file number

This is because such a document set has technical and similar contents.

Table I shows the results of LDA topic analysis for an abstract set consisting of 525 academic papers found by the query "information retrieval" on Cute.Search (the academic search service at Kyushu University). One can see that it is difficult to determine what each topic is.

Hence, we propose a method for estimating the representative words of each topic from the results of LDA topic analysis of an abstract set obtained by a keyword search. Our method consists of the following three components:

a) Improving LDA [4]:
We improve the algorithm so as to calculate the semi-optimum solution z maximizing $p(\mathbf{z} \mid \mathbf{w}, \alpha, \beta)$.

b) Deleting unnecessary words that occur in many topics, and generating a document set for each topic.

c) Estimating representative words from a document set for each topic.

A. *Improving LDA*

Griffiths et al. estimate θ and ϕ using the s-th result of sample $\mathbf{z}$ (where s is large enough) [4]. It does not matter actually if we are only interested in θ and ϕ, and if both the document size and document number are large. In the proposed method, we construct a document set of each topic using $\mathbf{z}$. This makes it a problem using the s-th result of sample $\mathbf{z}$. Therefore, we improve the algorithm so as to get the semi-optimal solution $\mathbf{z}$ that maximizes (4) among s samples. We call the obtained $\mathbf{z}$ the "suboptimal topic assignment."

B. *Deleting Unnecessary Words and Constructing a Document Set of Each Topic*

We calculate the idiosyncrasy of each word for a topic and remove words that have low idiosyncrasy. Specifically, we calculate the entropy of a word. We remove words that seem to be ineffective for topic expression by setting a threshold for entropy.

The entropy of word w is given as

$$E(w) = -\sum_{t=1}^{K} p(t \mid w) \log p(t \mid w) , \quad (6)$$

where $p(t \mid w)$ is the maximum likelihood estimate by suboptimal topic assignment z as follows:

$$p(t \mid w) = \frac{n_{ZW}(t, w; \mathbf{w}, \mathbf{z})}{\sum_{t=1}^{K} n_{ZW}(t, w; \mathbf{w}, \mathbf{z})} . \quad (7)$$

The word that has the lowest idiosyncrasy is the word w that satisfies

 ISBN: 978-1-61208-329-2

$$p(t \mid w) = \frac{1}{K} \quad , \tag{8}$$

for every topic t. The entropy of this word is $\log_2 K$. Then, we consider a word w as unnecessary and remove it if w satisfies

$$E(w) > \kappa \log_2 K \ . \tag{9}$$

Now we set κ to 0.25 for a preliminary experiment.

The document set of each topic t (=1, 2, …, K),

$$\mathbf{w}[t] = (\mathbf{w}^{(1)}[t], \mathbf{w}^{(2)}[t], \cdots, \mathbf{w}^{(M_t)}[t]) \ ,$$

is constructed from the results (**w**, **z**) of LDA topic analysis as follows:

a) Set m =1 and i =1.

b) Seek the following word set (word sequence):

$$\{w_n^{(m)} \mid z_n^{(m)} = t, \text{and } E(w_n^{(m)}) \le \kappa \log_2 K\} \ .$$

If the number of elements (words) in this set is over L, then set as follows:

$$w^{(i)}[t] = \{w_n^{(m)} \mid z_n^{(m)} = t, \text{and } E(w_n^{(m)}) \le \kappa \log_2 K\}$$

and $i \leftarrow i$+1.

c) If m equals M, the construction process is finished. Otherwise, $m \leftarrow m$ +1 and repeat step (b).

We do not replace pronouns with their antecedents when constructing input data for LDA. Then, a word that appears frequently is not always important for a certain topic. A word that is referred by a pronoun is sometimes important, which is why we delete redundant words in a document. We also delete any document that has less than L words from the document set of a topic, because it seems difficult to estimate a representative word of such a document. There would be noise for estimating a representative word. Now, we set L = 4.

C. *Estimating Representative Words for a topic*

We estimate the representativeness of each word for topic t from the document set of t (document sequence):

$$\mathbf{w} = (\mathbf{w}^{(1)}, \mathbf{w}^{(2)}, \cdots, \mathbf{w}^{(M)}) \ .$$

This is constructed with the method of the preceding paragraph. We omit t from this because the following process is executed for each topic. In addition, the number of documents M, the identification of each word and number of words (by type) V are also set for each topic t. In the same way, parameters introduced in the following model are also set for each topic t.

In our model, the document $\mathbf{w}^{(m)}$ is generated in the following two steps: 1) a word x is generated as a representative word, and 2) x generates the other words in the document. The probability of generating x as a representative word is denoted by η_x, the probability of w occurring in the document whose representative word is x is denoted by $\xi_1^{(x,\,w)}$, and $\xi_0^{(x,\,w)}$ is 1 - $\xi_1^{(x,\,w)}$. Here, the probability of generating x ($\in \mathbf{w}^{(m)}$) as a representative word and generating the other words in $\mathbf{w}^{(m)}$ from x is expressed as follows:

$$\begin{aligned} & p(\mathbf{w}^{(m)}, x \mid \eta, \xi) \\ & = \eta_x \times \left(\prod_{\substack{w=1 \\ w \notin \mathbf{w}^{(m)}, w \neq x}}^{V} \xi_0^{(x,w)} \right) \times \left(\prod_{\substack{w=1 \\ w \in \mathbf{w}^{(m)}, w \neq x}}^{V} \xi_1^{(x,w)} \right). \end{aligned} \tag{10}$$

The prior distribution of $\eta = (\eta_1, \eta_2, \ldots, \eta_V)$ is the dimensionality V-1 of the Dirichlet distribution with parameter γ:

$$p(\eta \mid \gamma) = \frac{\Gamma(V\gamma)}{\Gamma(\gamma)^V} \prod_{x=1}^{V} \{\eta_x\}^{\gamma-1} \ . \tag{11}$$

The prior distribution of ($\xi_0^{(x,\,w)}$, $\xi_1^{(x,\,w)}$) is the beta distribution (one-dimensional Dirichlet distribution) with parameter δ:

$$p(\xi_0^{(x,w)}, \xi_1^{(x,w)} \mid \delta) = \frac{\Gamma(2\delta)}{\Gamma(\delta)^2} \{\xi_0^{(x,w)}\}^{\delta-1} \{\xi_1^{(x,w)}\}^{\delta-1} \ . \tag{12}$$

The representative word of document $w^{(m)}$ is denoted by $x^{(m)}$, and the set of representative words for all documents is denoted by **x**:

$$\mathbf{x} = (x^{(1)}, x^{(2)}, \cdots, x^{(M)}) \ .$$

We define two counters as follows. $n_R(x \ ; \mathbf{x})$ is the number of times x has been selected as a representative word in **x**, and $n_C(x,w \ ; \mathbf{w}, \mathbf{x})$ is the number of elements in the set:

$$\{m \mid x^{(m)} = x, \text{and } w(\neq x) \in \mathbf{w}^{(m)}\} \ .$$

(In other words, n_C is the number of documents that have x as a representative word and contains word w)

The probability of occurrence (**w**, **x**) is

 ISBN: 978-1-61208-329-2

$$p(\mathbf{w},\mathbf{x}\mid\eta,\xi) = \left(\prod_{x=1}^{V}\{\eta_x\}^{n_R(x;\mathbf{x})}\right) \times\left(\prod_{x=1}^{V}\prod_{\substack{w=1\\ w\neq x}}^{V}\{\xi(x,w)\}_0^{n_R(x;\mathbf{x})-n_C(x,w;\mathbf{w},\mathbf{x})}\{\xi(x,w)\}_1^{n_C(x,w;\mathbf{w},\mathbf{x})}\right). \quad (13)$$

Then, we obtain the conditional probability of $x^{(m)} = x$ given the representative words of $\mathbf{w}$ without $\mathbf{w}^{(m)}$ via

$$p(\mathbf{w},\mathbf{x}/x^{(m)},x^{(m)}=x\mid\gamma,\delta) \propto\left(\frac{n_R(x;\mathbf{x}/x^{(m)})+\gamma}{(M-1)+V*\gamma}\right) \times\left(\prod_{\substack{w\notin\mathbf{w}^{(m)}\\ w\neq x}}^{V}\frac{n_R(x;\mathbf{x}/x^{(m)})-n_C(x,w;\mathbf{w},\mathbf{x}/x^{(m)})+\delta}{n_R(x;\mathbf{x}/x^{(m)})+2\delta}\right) \times\left(\prod_{\substack{w\in\mathbf{w}^{(m)}\\ w\neq x}}^{V}\frac{n_C(x,w;\mathbf{w},\mathbf{x}/x^{(m)})+\delta}{n_R(x;\mathbf{x}/x^{(m)})+2\delta}\right), \quad (14)$$

where $\mathrm{x}/x^{(m)}$ means representative words of $\mathbf{w}$ without $\mathbf{w}^{(m)}$.

We estimate η using Gibbs sampling as follows. $E[\eta_x \mid \mathbf{w}, \mathbf{x}]$, the expectation value of η_x from the posterior distribution of η_x given $\mathbf{w}$ and its representative words $\mathbf{x}$, is calculated according to

$$E[\eta_x \mid \mathbf{w},\mathbf{x}] = \frac{n_R(x;\mathbf{x})+\gamma}{M+V\gamma}. \quad (15)$$

$E[\eta_x \mid \mathbf{w}]$, the expectation value of η_x from the posterior distribution of η_x given the document set $\mathbf{w}$, is found from

$$\begin{aligned} E[\eta_x\mid\mathbf{w}] &= \int \eta_x p(\eta,\xi\mid\mathbf{w})d\eta d\xi \\ &= \int \eta_x \frac{\sum_{\mathbf{x}} p(\mathbf{w},\mathbf{x},\eta,\xi)}{p(\mathbf{w})}d\eta d\xi \\ &= \sum_{\mathbf{x}}\frac{p(\mathbf{w},\mathbf{x})}{p(\mathbf{w})}\int \eta_x \frac{p(\mathbf{w},\mathbf{x},\eta,\xi)}{p(\mathbf{w},\mathbf{x})}d\eta d\xi \\ &= \sum_{\mathbf{x}} p(\mathbf{x}\mid\mathbf{w})E[\eta_x\mid\mathbf{w},\mathbf{x}] \end{aligned}. \quad (16)$$

Let $\mathbf{x}(S_0+1)$, $\mathbf{x}(S_0+2)$, ..., $\mathbf{x}(S_0+S)$ be the sequence of representative words obtained by Gibbs sampling from (S_0+1) to (S_0+S) rounds. Then, $E[\eta_x \mid \mathbf{w}]$ is approximated by (15), (16), and the law of large numbers:

$$E[\eta_x\mid\mathbf{w}] = \frac{1}{S}\sum_{s=S_0+1}^{S_0+S}\frac{n_R(x;\mathbf{x}(s))+\gamma}{M+V\gamma}. \quad (17)$$

Our method presents a list of representative words w with high probability η_w for each topic. Table II shows the results of the topic analysis performed by our method for the same dataset as in Table I.

TABLE II. SAMPLE TOPICS FROM OUR METHOD

Topic	Representative Word Set Express Topic
1	Model space method largesystems findings determine andtajikistan cells researcheshave methodsin efficiency were applied unwanted subjected
2	Processing issue indexing image conference participant format storey forseveral child andperformance ai articolo nostril name
3	System behavior difference interaction medium sinceinformation characterization control recovery management ehrlich gate eigenvector completion agent
4	Datum algorithm method structure value deviation omit market mechanical acceptance complete avenue stemmer between decision
5	document query collection factor temperament preference ohio occupation chicago feedback department finder lsa formalism proposition
6	computer library index access organization storage university control labor science classroom rs skill instruction subscales
7	search web interface dei indexdocuments iv onthe day north request ofwordnet keywordstoindexing print collaboration tapas
8	field part multimedia tool portland researcher roll machine illustration film discovery sidebar facilitarne hypertext diffuse
9	recall sense trec function word effect component weight class investigation iss efficacy combination isss thesaurus
10	database author notice report fax communication general american rule horizon hole analogue correlation radiation the

IV. EXPERIMENT

We performed an experiment to compare the method of [1] and our method. We prepared 20 queries and collected about 500 to 1500 Japanese abstracts from the article database CiNii for each query. We did topic analysis for the collected abstracts using LDA with 10 topics, and then estimated representative words of each topic using our method. We set α and β of LDA's meta-parameters to 2.0

 ISBN: 978-1-61208-329-2

and 0.1 respectively. We set γ and δ of our model's meta-parameters to 0.05 and 0.1, respectively.

We evaluated the analysis results with each method as follows. 1) Four students studying the areas of electrical engineering and computer science evaluated the results. We divided them into two groups of two students. The four evaluators are denoted by a1, a2, b1 and b2. For each query, we assigned methods to the students so that the method that a1 and a2 evaluated was different from the method that b1 and b2 evaluated. For every query, we replaced the methods being evaluated. As a result, each student evaluated the results for 10 queries using each method. 2) We evaluated the analysis result for method M and query Q. The evaluators were given the word list (15 words) for every topic determined by method M and the 10 abstracts randomly selected from the search results by query Q. For each abstract *a*, the evaluator selected three topics that seemed to be included in *a* using his sense based on the word lists of topics, and we scored the size of the intersection between selected topics and the following set to *a*:

$$\{t \mid \theta_t^{(a)} \geq \text{the third highest value in } \theta_1^{(a)}, \theta_2^{(a)}, \cdots, \theta_{10}^{(a)} \}.$$

As a result, the score for an abstract is from 0 to 3. The score for query Q is the sum of the scores of each evaluator in a group for every abstract in a set retrieved by Q. As a result, a score for query Q (that is, an abstract set retrieved by Q) is from 0 to 60.

The results of the evaluation are in Table III. The scores for our method are higher than those for LDA for most queries (No. 1, 4–7, 10–13, and 16–20), and the average of the scores for all abstract sets for our method is 1.3 points higher than for LDA. However, this is not a very big increase. We assume that users use the topic analysis to narrow down the results of a keyword search about their own field or related fields. The evaluators were unfamiliar with some of the prepared queries. For the abstract sets retrieved by familiar queries (No. 1, 4–10, 13, 16, 19, and 20 in Table III), the average score for our method is 2.52 points higher than for LDA.

TABLE III. RESULTS OF EXPERIMENT

No.	Query	LDA	Our method
1	“Natural language”	23	25
2	“Translation”	25	21
3	“Medical treatment”	26	24
4	“Light, Energy”	25	27
5	“Ion, Electricity”	14	19
6	“Sensor, Measurement”	19	20
7	“Energy, Environment”	20	25
8	“Electric power, Supply”	19	17
9	“Retrieval, Support”	15	17
10	“Radio wave, Transmission”	19	20
11	“Concrete”	20	21
12	“Fluid mechanics”	13	14
13	“Quantum”	13	20
14	“Plasma”	22	22
15	“Nuclear fusion”	20	17
16	“Sensing”	24	25
17	“Project management”	18	19
18	“Aviation, Cosmos”	21	23
19	“Artificial intelligence”	20	23
20	“Communication network”	16	19

V. CONCLUSION AND FUTURE WORK

The representative word lists generated by our method does not contain some unnecessary words that are contained in word lists generated by LDA, but there are many non-content words and general terms in our lists. Our goal is to make LDA analysis more intelligible. We cannot expect a very big improvement in expression of topic contents when LDA analysis is not good. In this work, the meta parameters α and β in the LDA were set to 2 and 0.1, respectively. In future work, we will explore better values of the meta parameters and compare the results for LDA and our method. In addition, we will evaluate the effect of filtering words in LDA word lists using entropy and explore a better entropy threshold.

VI. REFERENCE

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

[2] D. Andrzejewski and D. Buttler, Latent Topic Feedback for Information Retrieval, Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining, August 21-24, 2011, pp. 600-608.

[3] Y. Tu, N. Johri, D. Roth, and J. Hockenmaier, Citation Author Topic Model in Expert Search, Proceedings of the 23rd International Conference on Computational Linguistics: Posters, August 23-27, 2010, pp. 1265-1273.

[4] T. L. Griffiths and M. Steyvers, Finding scientific topics, PNAS, vol. 101, 2004, pp. 5228-5235.

[5] D. M. Blei and J. D. Lafferty, Visualizing Topics with Multi-Word Expressions, Technical Report, arXiv:0907, 1013vl [stat.ML], 2009.

 ISBN: 978-1-61208-329-2