IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received November 28, 2018, accepted December 10, 2018, date of publication December 18, 2018,
date of current version January 16, 2019.

Digital Object Identifier 10.1109/ACCESS.2018.2887314

Mining Dynamics of Research Topics Based on the Combined LDA and WordNet

CHAO LI[1,2], SEN FENG[1], QINGTIAN ZENG[1], WEIJIAN NI[1], HUA ZHAO[1], AND HUA DUAN[1]

[1]College of Computer Science and Engineering, Shandong Province Key Laboratory of Wisdom Mine Information Technology, Shandong University of Science and Technology, Qingdao 266590, China
[2]Key Laboratory of Embedded System and Service Computing, Tongji University, Ministry of Education, Shanghai 201804, China

Corresponding authors: Qingtian Zeng (qtzeng@sdust.edu.cn) and Hua Duan (hduan@sdust.edu.cn)

This work was supported in part by the National Natural Science Foundation of China under Grant 61702306, Grant 61472229, and Grant 61602278, in part by the Ministry of Education of the China Foundation for Humanities and Social Sciences under Grant 16YJCZH041, Grant 17YJCZH262, and Grant ZR2018BF013, in part by the Shandong Provincial Natural Science Foundation of China under Grant ZR2017BF015, Grant 2016ZDJS02A11, and Grant ZR2017MF027, in part by the Taishan Scholar Climbing Program of Shandong Province, in part by the SDUST Research Fund under Grant 2015TDJH102, in part by the Open Project of the Key Laboratory of Embedded System and Service Computing, Tongji University, under Grant ESSCKF 2016-06, and in part by the Scientific Research Foundation of the Shandong University of Science and Technology for Recruited Talents under Grant 2016RCJJ011.

ABSTRACT A large volume of research documents are available online for us to access and analysis. It is very important to detect and mine the dynamics of the research topics from these large corpora. In this paper, we propose an improved method by introducing WordNet to LDA. This approach is to find latent topics of large corpora, and then we propose many methods to analyze the dynamics of those topics. We apply the methodology to two large document collections: 1940 papers from NIPS 00-13 (1987–2000) and 2074 papers from NIPS 14-23 (2001–2010). Six experiments are conducted on the two corpora and the experimental results show that our method is better than LDA in finding research topics and is feasible in discovering the dynamics of research topics.

INDEX TERMS Research topics mining, dynamics of research topics, latent Dirichlet allocation, WordNet, large corpora.

I. INTRODUCTION

Learning topics or patterns from large corpus has drawn increasing attentions in data mining and related areas as more and more electronic document archives are available on the Internet. Recent researches in machine learning and text mining have developed many classical techniques, e.g., Latent Semantic Analysis (LSA) [1], [2], Probabilistic Latent Semantic Indexing (pLSI) [3], Latent Dirichlet Analysis (LDA) [4], and Topic Word Embedding [5] for finding patterns of words in large document collections. Among all those techniques, hierarchical probabilistic models, also known as ''topic model'', have become a widely used approach for exploratory and predictive analysis of text [6]–[10]. LDA is a three-level hierarchical Bayesian model, in which each item of a collection is modeled as a finite mixture over an underlying set of topics. Each topic is, in turn, modeled as an infinite mixture over an underlying set of topic probabilities.

Although LDA has been applied with remarkable success in different domains, e.g. document modeling, document classification, and collaborative filtering, it has a number of deficits:

First, LDA is a generative model, which simulates the process of human thinking. Imagine that an author tries to write a paper. The author may first design the structure of the paper and determine which topics are included, then select appropriate words of each topic to express his or her idea. But one's knowledge is limited, the author might not select the appropriate words to express a topic, for example, people may continuously use the same word to express their opinions, which could lead to an unsatisfying result trained by LDA.

Second, many synonyms are always found in the same topics. According to the experiments on large document collections using LDA, we find that some of the results are not satisfying. For example, LDA is the abbreviation of Latent Dirichlet Allocation, and ''LDA'' and ''Latent Dirichlet Allocation'' often appears in the same topic.

Therefore, adding pre-existing knowledge, especially lexical knowledge to LDA is necessary. In this paper, we improve LDA by taking WordNet as an external lexical knowledge source. WordNet [11] is a lexical database of English words, which groups nouns, verbs, adjectives and adverbs into sets of synonyms (synsets), to represent a distinct concept.

Synsets are interlinked with conceptual-semantic and lexical relations, including hypernymy, meronymy, causality, etc [12].

Through discovering a set of topics from a large document collection, our approach is capable of analyzing the dynamics of these topics as a means of gaining insight into the dynamics of science. It is meaningful to find which research areas are rising or falling in popularity. With the improved LDA model, we could assign the documents to some of the topics according to the Topic-Document distribution, and each topic keeps a collection of documents. In each topic, the documents are grouped by time (i.e. year). We count the number of documents by time, and are able to find out whether the topics are rising or falling in popularity.

Moreover, since scientific papers contain vast authorship information, our approach is capable of finding topic distributions of authors. With those topic distributions, we can analyze the dynamics of research interests of authors. Considering the sparsity of the scientific papers of specific author, we first build an author network base on the co-author relationship, and then establish the author groups as the ''authors.''

The remainder of the paper is organized as follows: **Section II** briefly reviews related works. **Section III** describes basic knowledge about LDA and Gibbs Sampling. **Section IV** proposes two methods of research topic mining: applying WordNet after LDA and applying WordNet before LDA. The analysis on the dynamics of research topics is showed in **Section V**. The experimental result and analysis are present in **Section VI**. The section includes methods of research topic mining and evolutionary methods of research topic. **Section VII** concludes the paper and discusses some future work.

II. RELATED WORK

Automatic extraction of topics from large corpus has been addressed in prior work using a number of different approaches. An important approach is to cluster the documents into groups containing similar semantic contents, using any of a variety of well-known document clustering techniques [13]–[15]. While clustering can provide useful broad information about topics, clusters are inherently limited by the fact that each document is (typically) only associated with one cluster. This is often at odds with the multi-topic nature of text documents in many contexts. For this reason, other representations (such as LDA discussed below) that allow documents to be composed of multiple topics generally provide better models for sets of documents.

To solve the above problem, Blei *et al.* [4] proposed a more general Bayesian probabilistic topic model called LDA. The LDA is a three-level hierarchical Bayesian model, in which each item of a collection is modeled as a finite mixture over an underlying set of topics. Each topic is, in turn, modeled as an infinite mixture over an underlying set of topic probabilities. Since standard estimation methods (variational EM) of LDA are intractable. Cai *et al.* [16] showed how Gibbs sampling, a Markov chain Monte Carlo technique, could be applied in this model.

WordNet [11] has been used for a number of different purposes in information systems, including word sense disambiguation, information retrieval, automatic text classification, automatic text summarization, and even automatic crossword puzzle generation. An important example of the use of WordNet is to determine the similarity between words. Various algorithms have been proposed [17]–[19], and these include considering the distance between the conceptual categories of words, as well as considering the hierarchical structure of the WordNet ontology.

There are also many scientific papers about the combination of LDA and WordNet. Morris and Hirst [17] used LDA and WordNet for information retrieval (IR). They employed WordNet engine to enrich the user's query with semantic lexical synonymous terms and LDA to extract highly ranked topic from a query's retrieved information. Musat *et al.* [22] proposed the use of WordNet as a post-processing step for detecting and removing outliers (words that are not conceptually similar to the others) from the topic labels learned by the LDA.

Many scholars are also very interested in the learning Dynamic Topic Model (DTM) and Author-Topic Model (ATM). For the dynamic topic model, Wang *et al.* [20] introduced dynamic topic models to analyze the time evolution of topics in large document collections. Under this model, articles are grouped by year, and each year's articles arise from a set of topics that have evolved from the last year's topics. Mooman *et al.* [21] developed the continuous time dynamic topic model, a dynamic topic model that uses Brownian motion to model the latent topics through a sequential collection of documents. More related work are shown in [22]–[28], but what they mainly focused on is to improve the model itself and the variational approximate inference algorithm, they didn't add pr-existing knowledge to the model. For the author-topic model, RosenZvi *et al.* [29] introduced a generative author-topic model for documents that extends the LDA to include authorship information. Moreover, Zhang *et al.* [30] developed the Author-Document Topic Model (ADT) which builds the model for the corpus both at the author level and the document level to figure out the problem of authorship attribution for short texts. More related work are proposed in [31]–[33].

III. BASIC KNOWLEDGE ABOUT LDA AND GIBBS SAMPLING

In this section, we first introduce the basic knowledge used in the paper, which mainly includes LDA and Gibbs Sampling.

A. LDA

LDA [3] is a generative probabilistic model of a corpus. The basic idea is that documents are represented as random mixtures over latent topics, where each topic is characterized by a distribution over words. LDA can be described as:

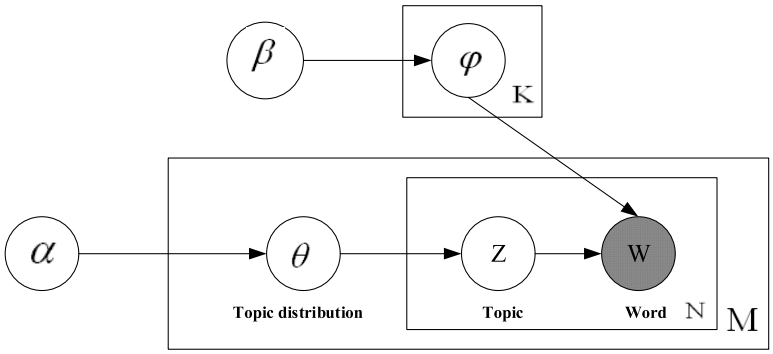


FIGURE 1. Graphical model representation of LDA.

(1) For each topic z $= 1, ..., k$ choose W dimensional $\varphi \sim Dir(\beta)$.
(2) For each document $d = 1, ..., D$.
(3) Choose the length of d:$N_d \sim$ Poisson(ξ).
(4) Choose θd $\sim$ Dir(α).
(5) For each of the N_d words w_n:
(5.1) Choose a topic $Z_n \sim Multinomial(\theta d)$: p($zdn|\theta$).
(5.2) Choose a word $W_n \sim Multinomial(\varphi)$: p($w$dn|$z$dn;β).

The LDA model is represented as a probabilistic graphical model in Figure 1, where there are three levels to the LDA representation. The parameters α and β are corpus level parameters, assumed to be sampled once in the process of generating a corpus. The variables θ_d are document-level variables, sampled once per document. Finally, the variables z_{dn} and w_{dn} are word-level variables and are sampled once for each word in each document.

Given the parameters α and β, the joint probability of observing a corpus can be calculated is:

$$p(w, z, \varphi, \theta|\alpha, \beta) = \prod_{m=1}^{M} p(\theta m|\alpha) \prod_{n=1}^{N} p(wm, n|zm, n, \varphi)p(zm, n|\theta m)p(\varphi|\beta) \tag{1}$$

We should notice that the distribution of θ and φ is intractable to compute in general, so some parameter estimation methods (e.g., Variational inference EM [4], Gibbs sampling [25], [26]) are used to estimate parameters.

B. APPROXIMATE INFERENCE BY GIBBS SAMPLING

The Gibbs sampler [25] is a special case of Metropolis-Hastings sampling wherein the random value is always accepted (i.e. $\alpha = 1$). The task remains to specify how to construct a Markov Chain whose values converge to the target distribution. In this section, our main objective is as follows:

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

$$\theta_{m,k} = \frac{n_{m,k} + \alpha_k}{\sum_{k=1}^{K} n_{m,k} + \alpha_k} \tag{3}$$

where $n_{k,v}$ is the number of times word v has been assigned topic k and $n_{m,k}$ is the number of times topic k appear in document m. But $n_{k,v}$ and $n_{m,k}$ are unknown to us now.

We can derive the chain updates,

$$\begin{aligned} p(zi = k|z - i, w, \alpha, \beta) &= \frac{p(z, w|\alpha, \beta)}{p(z - i, w|\alpha, \beta)} \\ &= \frac{p(w|z, \beta)p(z|\alpha)}{p(z - i, w - i|\alpha, \beta)p(z, wi|\alpha, \beta)} \\ &= \frac{p(w|z, \beta)p(z|\alpha)}{p(w - i|z - i, \beta)p(z - i|\alpha)p(wi|\alpha, \beta)} \\ &\propto \frac{p(w|z, \beta)p(z|\alpha)}{p(w - i|z - i, \beta)p(z - i|\alpha)} \\ &\propto \frac{\beta_v + n_{k,v}}{\sum_{v=1}^{V} \beta_v + n_{k,v}} \frac{\alpha_k + n_{m,k}}{\sum_{k=1}^{K} \alpha_k + n_{m,k}} \end{aligned} \tag{4}$$

where $n_{k,v}$ and $n_{m,k}$ update with the chain updating. With a predefined number of iterations (the so-called burn-in time of the Gibbs sampler), the $\varphi_{k,v}$ and $\theta_{m,k}$ can be calculated.

IV. FRAMEWORK OF RESEARCH TOPIC MINING

Although LDA has been applied with remarkable success in different domains, it has a number of deficits. (1) Documents in the corpora trained by the LDA model always lack semantic knowledge, which could lead to an unsatisfying result. (2) Many synonyms are always found in the same topics. For example, LDA is the abbreviation of Latent Dirichlet Allocation, and "LDA" and "Latent Dirichlet Allocation" often appears in the same topic.

Thus, adding WordNet to LDA is necessary. To evaluate the performance of the combination of LDA and WordNet, we design the following two alternatives: (1) Applying WordNet before LDA, as shown in Figure 2(1). That means, the synonyms of words, extended by WordNet are added to the vocabulary list first. The LDA is adopted to mine the topics. (2) Applying WordNet after LDA, as shown

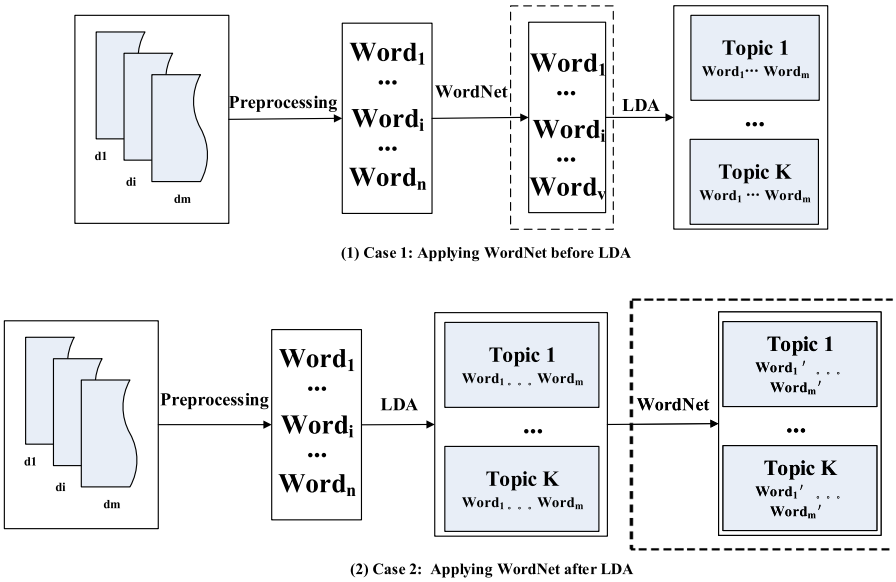


FIGURE 2. The process of our methodology.

in Figure 2(2). That means, the synonyms of words are merged to each topic

A. APPLYING WordNet BEFORE LDA

The method consists of the following three steps:

- **Step1**. For each document, by removing stop words and punctuations, we can get a keyword collection.
- **Step2**. For each word, we use WordNet to add synonyms of the word to the keyword collection, and remove the words which appear on the collection less than d (threshold) times. After this step, we can get a vocabulary list.
- **Step3**. By applying LDA on the vocabulary list, we can get a number of topics.

B. APPLYING WordNet AFTER LDA

This methodology aims to merge synonyms of words in the topic in order that the words in the topic can better describe the topic. The steps are as follows:

- **Step1**. We can get the Topic-Word probabilities with LDA
- **Step2**. Every topic is represented by the top m words with the highest probabilities on the topic. The words are ranked in a descending order according to value of probability.
- **Step3**. As for each word, morphology reduction.
- **Step4**. For the m words in the topic, the WordNet is adopted to judge if one word is synonymous with another. If two words are synonyms, then the word with the smaller probability is removed. The probability of another word is changed as the sum of the two probabilities. The word which ranked as the top is added to $m+1$ the topic.
- **Step5**. Return to **Step3**, until there are no synonyms for the m words.
- **Step6**. Sort the m words in a descending order according the probabilities

V. DYNAMICS OF RESEARCH TOPIC

Because our method discovers a set of topics from a large document collection, it is straightforward to analyze the dynamics of these topics as a means of gaining insight into the dynamics of science. It is meaningful to find which research areas are rising or falling in popularity. At the same time, scientific papers contain vast authorship information, so topic distributions of authors could be learned with our model. Then we can analyze the research interests of authors.

A. RESEARCH ON DYNAMICS OF RESEARCH TOPIC

In this section, we propose two methods to analyze the dynamics of these research topics: the first one is based on probability model and the second one is based on clustering.

1) LEARNING DYNAMICS OF RESEARCH TOPIC BASED ON PROBABILITY MODEL

From the discussion above, we could get the value of θ, and the meaning of θ is stated below:

$$\theta_{m\times k}=\begin{matrix}\begin{matrix}Topic\,1 & Topic\,2 & \cdots & Topic\,k\end{matrix}\\ \begin{Bmatrix}\theta_{0,0} & \theta_{0,1} & \cdots & \theta_{0,k}\\ \theta_{1,0} & \theta_{1,1} & \cdots & \theta_{1,k}\\ \cdots & \cdots & \cdots & \cdots\\ \theta_{m,0} & \theta_{m,1} & \cdots & \theta_{m,k}\end{Bmatrix}\begin{matrix}Document\,1\\ Document\,2\\ \cdots\\ Document\,\mathrm{m}\end{matrix}\end{matrix}$$

FIGURE 3. The meaning of θ.

From Figure 3 above, we can see that the value of $\theta_{i,j}$ means the probability of document i assigned to topic j. Thus we could define the Topic Strength below:

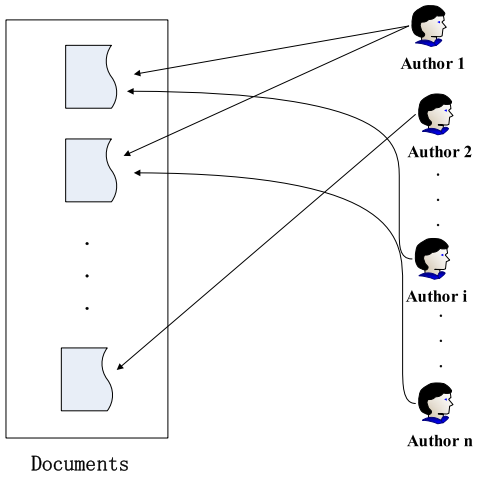


FIGURE 4. The relationship between documents and authors.

$$\theta_{m\times k}=\begin{matrix}\begin{matrix}Topic\,1 & Topic\,2 & \cdots & Topic\,k\end{matrix}\\ \begin{Bmatrix}\theta_{0,0} & \theta_{0,1} & \cdots & \theta_{0,k}\\ \theta_{1,0} & \theta_{1,1} & \cdots & \theta_{1,k}\\ \cdots & \cdots & \cdots & \cdots\\ \theta_{m,0} & \theta_{m,1} & \cdots & \theta_{m,k}\end{Bmatrix}\end{matrix}\begin{matrix}Doc\,1\\ Doc\,2\\ \cdots\\ Doc\,\mathrm{m}\end{matrix}$$

$$\theta^*_{A\times k}=\begin{matrix}\begin{matrix}Topic\,1 & Topic\,2 & \cdots & Topic\,k\end{matrix}\\ \begin{Bmatrix}\theta^*_{0,0} & \theta^*_{0,1} & \cdots & \theta^*_{0,k}\\ \theta^*_{1,0} & \theta^*_{1,1} & \cdots & \theta^*_{1,k}\\ \cdots & \cdots & \cdots & \cdots\\ \theta^*_{A,0} & \theta^*_{A,1} & \cdots & \theta^*_{A,k}\end{Bmatrix}\end{matrix}\begin{matrix}Author\,1\\ Author\,2\\ \cdots\\ Author\,A\end{matrix}$$

$$\theta_{m\times k}=\begin{matrix}\begin{matrix}Topic\,1 & Topic\,2 & \cdots & Topic\,k\end{matrix}\\ \begin{Bmatrix}\theta_{0,0} & \theta_{0,1} & \cdots & \theta_{0,k}\\ \theta_{1,0} & \theta_{1,1} & \cdots & \theta_{1,k}\\ \cdots & \cdots & \cdots & \cdots\\ \theta_{m,0} & \theta_{m,1} & \cdots & \theta_{m,k}\end{Bmatrix}\end{matrix}\begin{matrix}Doc\,1\\ Doc\,2\\ \cdots\\ Doc\,\mathrm{m}\end{matrix}$$

FIGURE 5. Get θ^* from θ.

Let

$$\hat{\theta}k = Ex = \frac{1}{\mathrm{m}}\sum_{\mathrm{i=1}}^{\mathrm{m}}\theta i,k \tag{5}$$

Be the Topic Strength of topic k. Where denotes the number of documents in the corpus.

Thus we could get the following procedure to learn the dynamics of research topics:

- Group documents by time. In this paper, we use year as the time unit. Suppose the document collection for year Ti is Ci.
- We can get the probability θ_k^{di} of each document assigned to topic k according to θ.
- Set the $\hat{\theta}k$ as following,

$$\hat{\theta}k = Ex = \frac{1}{|Ci|}\sum\nolimits_{di\in Ci}\theta kdi \tag{6}$$

As the topic strength of topic k in year Ti.

- With the graphical representation of the topic strength of topic k, we are able to see the dynamics of topic k.

2) LEARNING DYNAMICS OF RESEARCH TOPIC BASED ON CLUSTERING

Our methodology is as follows:

- If the probability of a document assigned to topic k $(\theta., k)$ is more than a threshold, then the document is classified into topic k.
- Suppose there are m documents assigned to topic k, group the m documents by year.
- Count the number of the documents assigned to topic k every year, we are able to see the dynamics of topic k with the graphical representation of the documents.

B. DYNAMICS OF RESEARCH TOPIC OF AUTHORS

In this section, we discuss three ways to learn knowledge on authors: the first one is learning research topics of authors, the second one is learning dynamics of research topic of authors, and the last one is learning author groups as academic teams and finding research topics of them.

1) LEARNING RESEARCH TOPIC OF AUTHORS

We know that documents in the corpus contain vast author information, and the relationship between documents and authors is stated below:

We could see from above figure that the relationship between documents and authors is many-to-many. That is: each document contains at least one author information, and each author wrote at least one document.

Meanwhile, we get the value of θ, then we could easily derive the value of θ.

$\theta^*_{i,j}$ in the figure stands for the probability of author i assigned to j. The procedure to get θ^* is as follows:

- Let θ^* be the $m \times n$ matrix, and m is the number of authors, n is the number of topics. Each value in θ^* is initialized to 0.
- For each document i, suppose document i is written by m authors, and these m authors are Auk, $Auk + 1, \cdots, Auk + m - 1$, and let the values of row k to $k + m - 1$ in θ^* be the value of row i in θ.
- Process each value in $\theta^*_{i,j}$ in θ^* below:

$$\theta i, j* = \frac{\theta i, j*}{\sum_{m=1}^{k}\theta i, m*} \tag{7}$$

After getting θ^*, we get the probability of each author assigned to each topic.

- Learning dynamics of research topic of authors.

Our methodology is to add time dimension to (1). Then we can group the $\mathrm{D}_{An,k}$ papers by year, analysis the dynamics of research topic of authors.

C. LEARNING RESEARCH TOPIC OF AUTHOR GROUPS

The number of papers written by a given author is generally small, so the results of the experiments are probably less than ideal. Considering the sparsity of the scientific papers of specific author, we first get author groups by clustering on an author network with the co-author relationship, and then analyze the research topic of author groups. The method for obtaining author group is as follows:

- Extract all the authors of each paper. For each paper, authors, as nodes, are interconnected. There are several authors connected due to the coauthor relationship. Then we can get the initial author network shown in Figure 6.

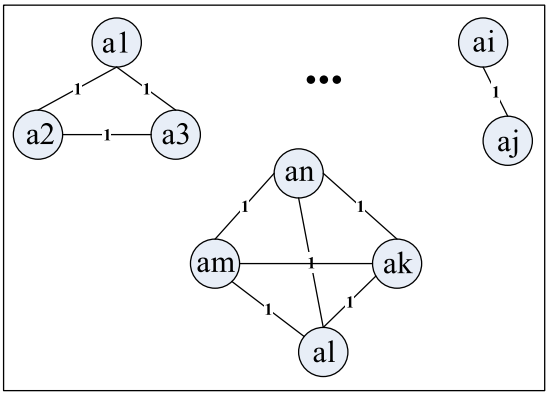


FIGURE 6. The initial author network.

- Merge the repeated nodes and sum up the weight of edges between the repeated edges as is shown in Figure 7. The merged results are that we can get.

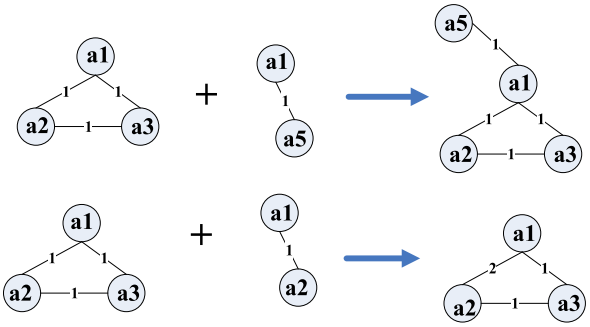


FIGURE 7. The process of merging network.

- Clustering the authors by setting a threshold.
- Then we can analyze the research topics of author groups like learning research topic of authors.

VI. EXPERIMENTS

In this section, we analyzed a subset of 4014 articles from NIPS, available on http://books.nips.cc. All the articles on http://books.nips.cc are pdf format. We downloaded a copy of txt format corpus from NIPS00-13(1987-2000, 1940 articles), then used Pdfbox to convert the pdf format of articles from NIPS 14-23 to the txt format. So we actually have two data sets: NIPS 00-13(1987-2000, 1940 articles) and NIPS 14-23(2001-2010, 2074 articles).

A. THE TOPICS DISCOVERED BY LDA AND OUR METHOD

1) EXAMPLE TOPICS DISCOVERED BY LDA

We trained the LDA model on NIPS 00-13, a set of papers from 14 years (1987 to 2000) of the Neural Information Processing (NIPS) Conference. This data set contains D = 1940 papers, and a vocabulary size of W = 74706 unique words. For a 100-Topic solution, 1000 iterations of the Gibbs sampler took about 6 hours of wall-clock time on a standard 1 GHz PC (19 seconds per iteration). The hyper parameters α and β are fixed at 1.

TABLE 1. Examples of topics 55 (out of 100 topic) to nips papers from 1987 to 2000.

Topic 55	
Word	**Pro.**
network	0.0133
learning	0.0110
neural	0.0075
training	0.0068
time	0.0067
state	0.0063
networks	0.0063
input	0.0063
units	0.0054
system	0.0052
Papers	**Pro.**
Generalization Performance in PARSEC: A structured Connectionist Parsing Arctiteture	0.907
Performance of synthetic neural network classification of noise	0.567

TABLE 2. Examples of topics 67 (out of 100 topics) to nips papers from 1987 to 2000.

Topic 67	
Word	**Pro.**
echoes	0.00078
echo	0.00066
dolphin	0.00056
bat	0.00054
integrator	0.00037
simmons	0.00028
sonar	0.00027
glint	0.00024
gateway	0.00024
echolocation	0.00023
Papers	**Pro.**
Acoustic-Imaging Computations by Echolocating Bats: Unification of Diversely-Represented Stimulus Features into Whole Images.	0.783
Natural Dolphin Echo Recognition Using an Integrator Gateway Network	0.565

TABLE 1, TABLE 2, TABLE 3 and TABLE 4 illustrates examples of 4 topics (out of 100) as learned by the LDA model for the NIPS 00-13 corpus. Each topic is illustrated

TABLE 3. Examples of topics 74 (out of 100 topics) to nips papers from 1987 to 2000.

Topic 74	
Word	**Pro.**
image	0.0279
images	0.0197
object	0.0113
face	0.0067
objects	0.0062
pixel	0.0051
recognition	0.0042
video	0.0041
vision	0.0041
pixel	0.0037
Papers	**Pro.**
Learning to Estimate Scenes from Images	0.219
Invariant object recognition using a distributed associative memory	0.146

TABLE 4. Examples of topics 84 (out of 100 topics) to nips papers from 1987 to 2000.

Topic 84	
Word	**Pro.**
model	0.0279
neurons	0.0197
input	0.0113
figure	0.0067
time	0.0062
neuron	0.0051
cells	0.0042
visual	0.0041
neural	0.0041
cell	0.0037
Papers	**Pro.**
Simulations suggest information processing roles for the diverse currents in hippocampal neurons	0.924
Morphogenesis of the Lateral Geniculate Nucleus: How Singularities Affect Global Structure	0.857

with the top 10 words most likely to be generated conditioned on the topic, and we also show 2 representative papers likely to be generated conditioned on the topic. We could see from TABLE 1 that topic 55 is very likely related to neural network, TABLE 2 that topic 67 is related to echo, TABLE 3 that topic 74 is related to image and video and TABLE 4 that topic 84 is related to neurons. As for the article, for example, ''Simulations suggest information processing roles for the diverse currents in hippocampal neurons'', it is likely related to topic 84 (neurons), for the probability of the article conditioned on the topic is 0.924.

2) EXAMPLE TOPICS DISCOVERED BY OUR METHOD

The data set in this section is the same as NIPS 00-13. As mentioned in Section IV, there are two ways applying WordNet on LDA:

Applying WordNet before LDA: This data set contains $D = 1940$ papers, after applying WordNet, the vocabulary size of W is extended to 89164 unique words or phrases. We also set a 100-topic solution, and the hyper parameters α and β are is also fixed at 1.

TABLE 5. The topic 31(out of 100 topics) extracted from nips papers published in 1987 to 2000.

Topic 31	
Word	**Pro.**
nap	0.00069
sleep	0.00067
eternal rest	0.00066
spoor	0.00065
slumber	0.00063
dream	0.00050
dreaming	0.00048
hobson	0.00042
rapid eye movement	0.00030
paradoxical sleep	0.00029
Papers	**Pro.**
Does the Wake-sleep Algorithm Produce Good Density Estimators	0.673
Models Wanted: Must Fit Dimensions of Sleep and Dreaming	0.188

TABLE 6. The topic 58(out of 100 topics) extracted from nips papers published in 1987 to 2000.

Topic 58	
Word	**Pro.**
language	0.0159
speech communication	0.0128
spoken language	0.0127
speech	0.0126
address	0.0097
realization	0.0071
credit	0.0070
identification	0.0070
acknowledgment	0.0069
recognition	0.0068
Papers	**Pro.**
Noise Suppression Based on Neurophysiologically-motivated SNR Estimation for Robust Speech Recognition	0.253
Speech Denoising and Dereverberation Using Probabilistic Models	0.192

TABLE 5, TABLE 6, TABLE 7 and TABLE 8 illustrate examples of 4 topics (out of 100) as learned by our model for the NIPS 00-13 corpus. Each topic is illustrated with the top 10 words most likely to be generated conditioned on the topic. We also show 2 representative papers which are likely to be generated conditioned on the topic. As shown in TABLE 5, the topic 31 is very relevant to sleep and dreaming, while in TABLE 6 the topic 58 is related to language and speech. TABLE 7 shows the topic 85 is related to neural network and in TABLE 8 demonstrates that the topic 90 is related to image.

Applying WordNet after LDA: Actually, this experiment is based on Section VI. With the methodology introduced

TABLE 7. The topic 85(out of 100 topics) extracted from nips papers published in 1987 to 2000.

Topic 85	
Word	**Pro.**
net	0.00786
web	0.00642
meshing	0.00642
meshnetwork	0.00640
network	0.00639
mesh	0.00639
yield	0.00549
figure	0.00472
neuronic	0.00378
neural	0.00376
Papers	**Pro.**
Generalization Performance in PARSEC: A Structured Connectionist Parsing Architecture	0.669
Performance of synthetic neural network classification of noise	0.623

TABLE 8. The topic 90(out of 100 topics) extracted from nips papers published in 1987 to 2000.

Topic 90	
Word	**Pro.**
picture	0.0080
ocular	0.0079
image	0.0070
ikon	0.0069
icon	0.0069
mental image	0.0068
persona	0.0068
movement	0.0060
optic	0.0052
position	0.0044
Papers	**Pro.**
Learning to Estimate Scenes from Images	0.416
Invariant object recognition using a distributed associative memory	0.323

in Section IV, we can obtain the final topics. Examples are shown in TABLE 9.

In TABLE 9, taking topic 54 for example, the top 10 words learned by LDA are relevant to bird song, by applying WordNet after LDA on the topic, we could know scientists always conduct experiments on sparrows or finch or even zebra, Hence, we can have a better recognition about the topic. Therefore the topics learned by our model are more representative and imply more knowledge than LDA.

3) COMPARISON AMONG OUR METHODS, LDA AND ATM

From the experimental results, we can see that many topics labeled with the top 10 words are meaningless. Taking in TABLE 10 for example, we have no idea what topic 1 is about. Moreover, the probabilities of the top 10 words are very small value. Thus the top 10 words could not represent the topic precisely. So we can arrive at the conclusion

TABLE 9. The comparison of 2 topics learned by lda and our method.

Topic74			
LDA		**Applying WordNet before LDA**	
Word	Pro.	Word	Pro.
image	0.0279	image	0.0476
images	0.0197	object	0.0175
object	0.0113	face	0.0098
face	0.0067	pixel	0.0088
objects	0.0062	vision	0.0078
pixel	0.0051	recognition	0.0042
recognition	0.0042	video	0.0041
video	0.0041	shape	0.0030
vision	0.0041	feature	0.0029
pixel	0.0037	view	0.0028
Topic 54			
LDA		**Applying WordNet before LDA**	
Word	Pro.	Word	Pro.
song	0.00185	song	0.00250
hvc	0.00041	syllable	0.00067
syllables	0.00040	bird	0.00051
bird	0.00032	hvc	0.00041
birdsong	0.00032	syrinx	0.00019
syllable	0.00027	ra	0.00016
syrinx	0.00019	Iman	0.00015
birds	0.00019	sparrow	0.00015
vocalization	0.00018	zebra	0.00015
ra	0.00016	finch	0.00014

TABLE 10. An example of meaningless topic motivated.

Topic 1	
Word	**Pro.**
uniform	3.97E-05
logz	3.95E-05
genetics	3.95E-05
nominal	3.95E-05
depart	3.95E-05
mnl	3.95E-05
microsoft	3.95E-05
characterize	3.95E-05
cooking	3.95E-05
sic	3.95E-05

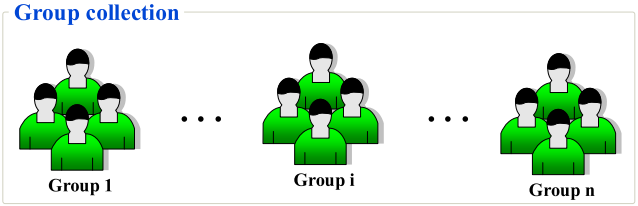


FIGURE 8. Groups learned from the network.

that the number of meaningful topics is far less than our expectations.

Motivated by the above observations, we compare the topics learned by our models (Applying WordNet before

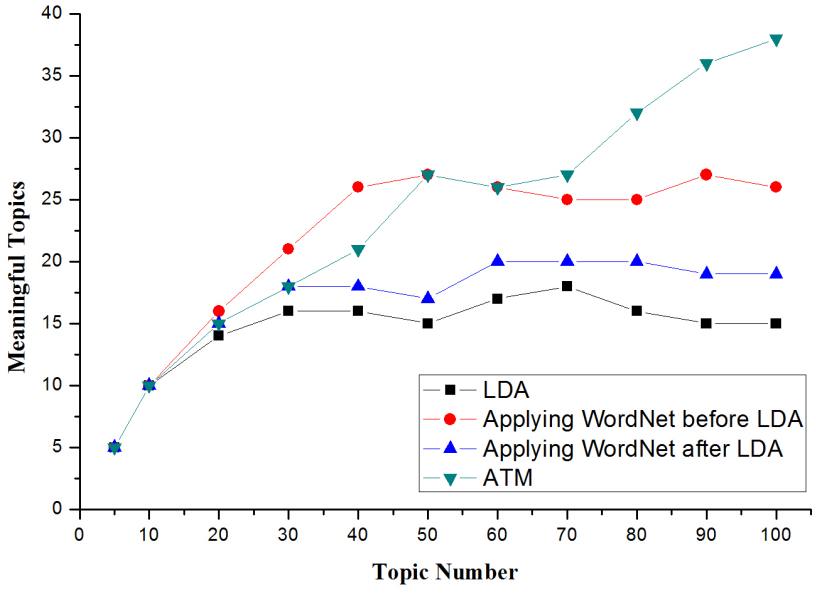
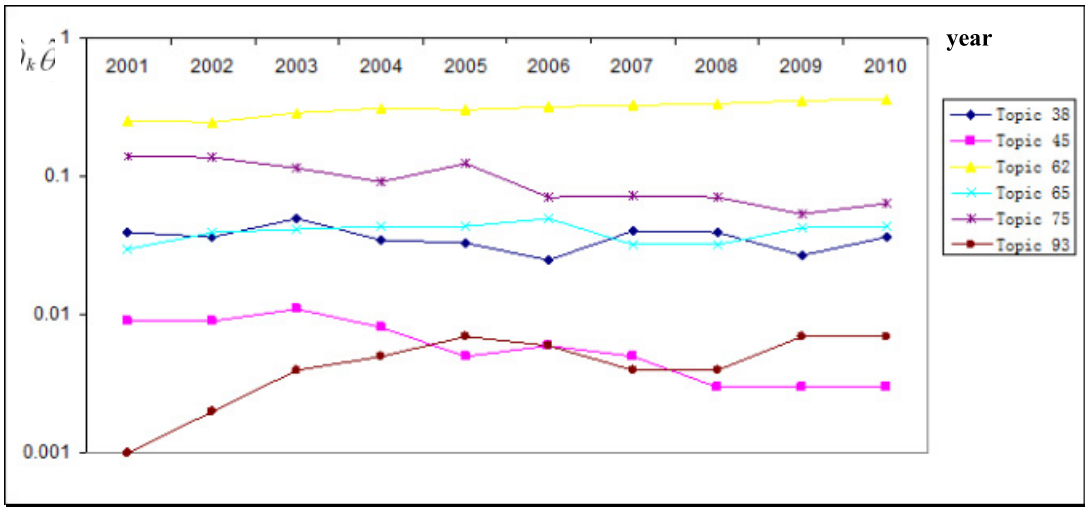


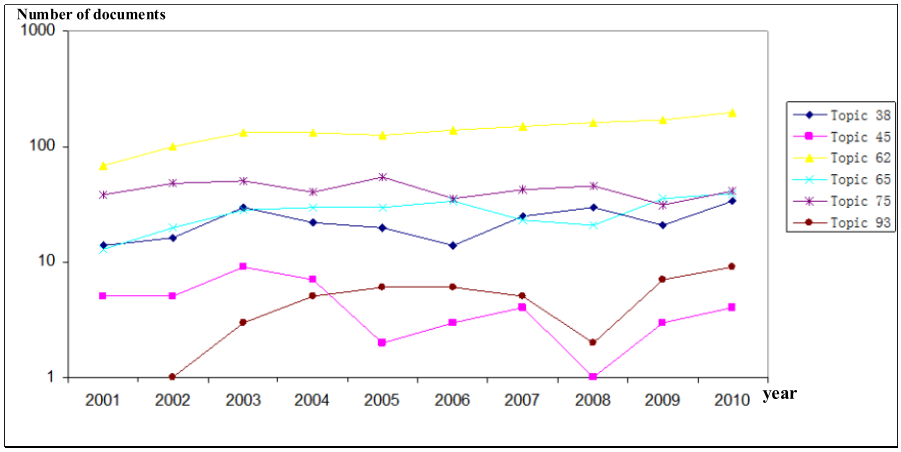
FIGURE 9. The comparison of meaningful topics.



Topic 38	Topic 45	Topic 62	Topic 65	Topic 75	Topic 93
commonwealth	speech communication	function	picture	stimulation	learning ability
nation	spoken language	algorithmic rule	image	stimulus	encephalon
land	speech	permit	icon	input	Mentality
province	spoken communication	map	persona	reply	brainpowermental
sate	sound	algorithm	mental image	neural	capacity
country	language	algorithmic program	images	figure	Brain
policy	address	single-valued function	form	neuronal	electroencephalogram
insurance	rootage	mathematical function	physical object	neuronic	Encephalogram
policy					

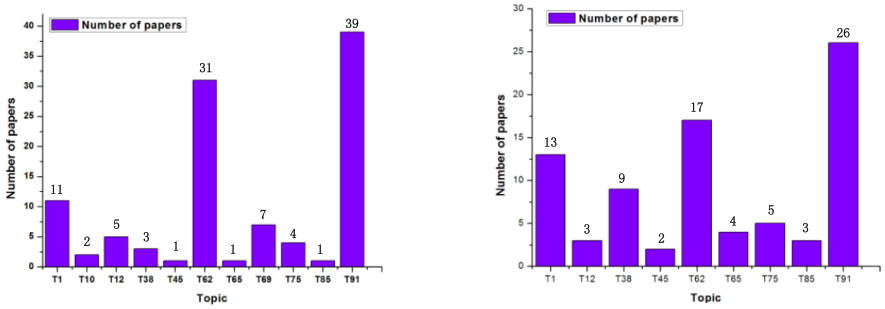
FIGURE 10. The dynamics of 6 topics from 2001 to 2010.

LDA and Applying WordNet after LDA), LDA [4] and ATM [29] (Author Topic Model). However, it is difficult to determine the count of topics for the topic model. There are two main ways to check the number of topics. The first one is to get the count of the topics decided by the minimum value of the perplexity which is the sum of similarity among topics [4]. The second one is to determine the count of topics according to the probability value of each Top-N keywords in each topic [30]. In this section, the goal of our paper is to detect the count of meaningful topics. Therefore, we assumes



Topic 38	Topic 45	Topic 62	Topic 65	Topic 75	Topic 93
commonwealth	speech communication	function	picture	stimulation	learning ability
nation	spoken language	algorithmic rule	image	stimulus	encephalon
land	speech	permit	icon	input	Mentality
province	spoken communication	map	persona	reply	brainpowermental
sate	sound	algorithm	mental image	neural	capacity
country	language	algorithmic program	images	figure	Brain
policy	address	single-valued function	form	neuronal	electroencephalogram
insurance policy	rootage	mathematical function	physical object	neuronic	Encephalogram

FIGURE 11. Dynamics of 6 topics from 2001 to 2010.



Topic 1	Topic 10	Topic 12	Topic 38	Topic 45	Topic 62
learning	cluster	language	commonwealth	speech communication	function
training	bunch	subject	nation	spoken language	algorithmic rule
education	clump	word	land	speech	permit
classification	clustering	tidings	province	spoken communication	map
categorisation	bunch up	intelligence	sate	sound	algorithm
data	constellate	news	country	language	algorithmic program
information	clusters	give-and-talk	policy	address	single-valued function
categorization	flock	lyric	insurance policy	rootage	mathematical function
	phantasmal				

Topic 65	Topic 69	Topic 75	Topic 85	Topic 91
picture	graph	stimulation	Helicopter	model
image	chart	stimulus	whirlybird	framework
icon	tree	input	eggbeater	simulation
persona	edge	reply	chopper	poser
mental image	nodes	neural	flying	information
images	knob	figure	sideway	data
form	trees	neuronal	flight of steps	figure
physical object	node	neuronic	rotor	number

FIGURE 12. The result of research topics of Michael Jordan and Andrew Ng.

that the topic in which the probability of Top 10 words is greater than 0.0001 is meaningful, according to the analysis of the experimental results and the themes presented by the topic keywords. We set a 5-topic, 10-topic, 20-topic, 30-topic, 40-topic, 50-topic, 60-topic, 70-topic, 80-topic, 90-topic and 100-topic, respectively on our method

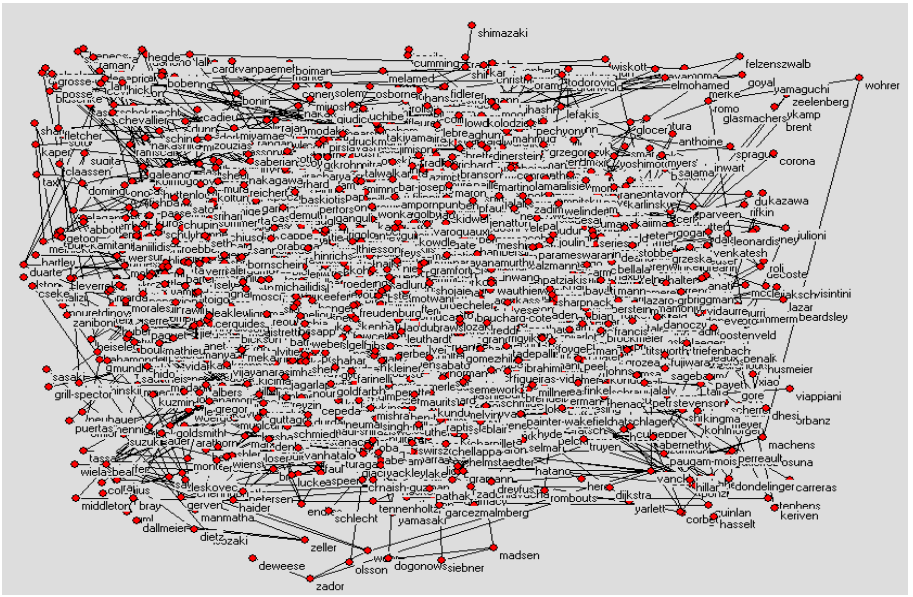


FIGURE 13. The network with of 2516 authors connected by co-author relationship.

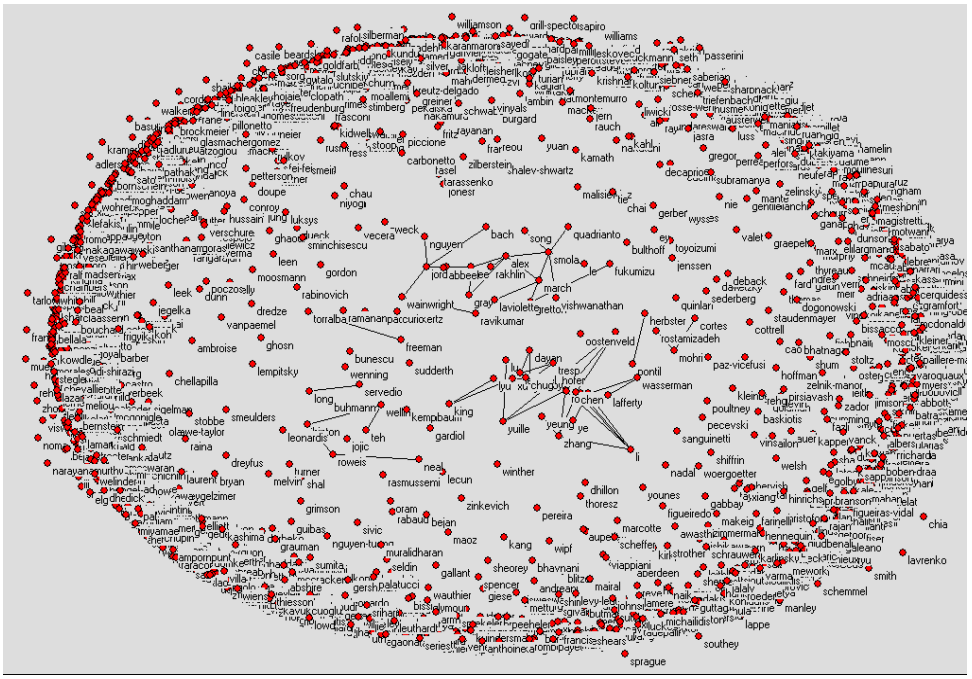
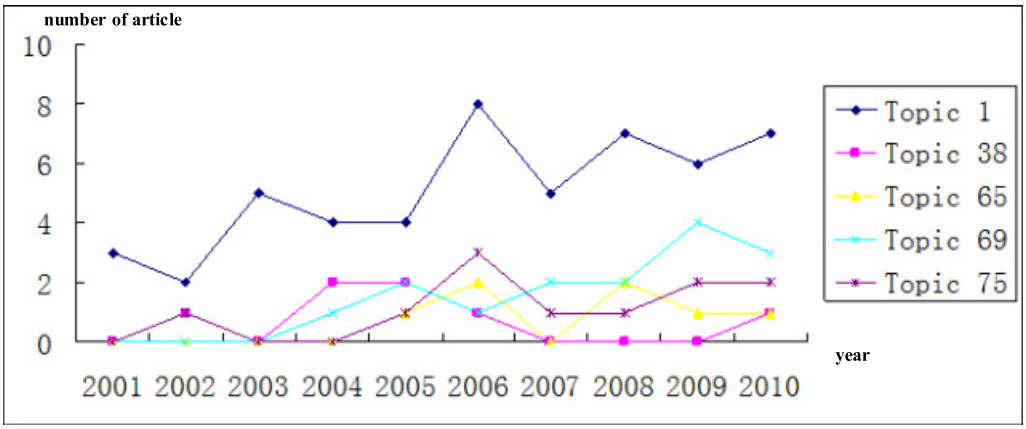


FIGURE 14. Net of 2516 authors with the threshold $\eta = 4$.

(Applying WordNet before LDA and Applying WordNet after LDA), LDA and ATM. The result of the meaningful topics are shown in Figure 9.

According to Figure 9, we can clearly see that: (1) In general, the number of meaningful topics generated by applying WordNet before LDA is larger than LDA. That is to say, applying WordNet before LDA can learn more meaningful topics compared with LDA. It is not surprisingly because WordNet brings more ''meaningful'' words to the model. (2) Applying WordNet after LDA is brings less meaningful topics than applying WordNet before LDA, It indicates that, WordNet can improve the quality of input text effectively for topic model. (3) Applying WordNet before LDA is better than ATM when the count of the topics is smaller than 50. Otherwise, the ATM is more excellent than Applying WordNet before LDA. The results show that WordNet can provide rich meaningful keywords for the topic model, but the effect on the topic model is limited. With the number of topics increases in



Topic 1	Topic 38	Topic 65	Topic 69	Topic 75
learning	commonwealth	picture	graph	stimulation
training	nation	image	chart	stimulus
education	land	icon	tree	input
classification	province	persona	edge	reply
categorisation	sate	mental image	nodes	neural
data	country	images	knob	figure
information	policy	form	trees	neuronal
categorization	insurance policy	physical object	node	neuronic

FIGURE 15. The dynamics of 5 (out of 11) topics.

ATM model, each the author is playing a very important role in meaningful topic discovery.

B. EXPERIMENTS ON THE DYNAMICS OF RESEARCH TOPIC

As discussed in Section V, there are two methods to analyze the dynamics of these research topics: The first one is based on probability model and the second one is based on clustering.

In this section, for researching on the dynamics of research topics, we choose the data set NIPS 13-22 (a set of papers from 2001-2010), which can better show the popularity of research topics recently. This data set contains 2074 papers, 119661 unique words as the vocabulary. We also set a 100-topic solution, and the hyper parameters α and β are also fixed to be 1.

1) EXPERIMENT ON LEARNING DYNAMICS OF RESEARCH TOPIC BASED ON PROBABILITY MODEL

As discussed in Section V, we applied the method to the data set NIPS 13-22 to analyze the dynamics of topics learned by our method from 2001 to 2010. Figure 10 shows the dynamics of 6 topics (the 8 most probable words in those topics are shown below the plots). As shown in Figure 10, we can clearly see that most of the topics are relatively stable. (For example topic 38(state policy), topic 65 (picture), topic 93(brain)). For the topic 45(speech and language) it implies a decreasing trend. While for the topic 62, it indicates an increasing trend.

2) EXPERIMENT ON LEARNING DYNAMICS OF RESEARCH TOPIC BASED ON CLUSTERING

In this section, we applied the method introduced in Section V on the data set NIPS 13-22. To make a reasonable comparison, we conduct the experiments on topic 38, topic 45, topic 62, topic 65, topic 75 and topic 93, which are the same as in Section VI. In this experiment, the threshold is set to be 0.1. As shown in Figure 11, the similar results are get as the Section VI, which indicates both of the above methods are feasible.

C. EXPERIMENTS ON THE DYNAMICS OF RESEARCH TOPIC OF AUTHORS

1) FINDING RESEARCH TOPIC OF AUTHORS

In this section, we also conduct experiments on NIPS 13-22. The top 2 authors who publish articles most on NIPS 13-22 are: Michael I. Jordan and Andrew Y. Ng. In our experiment, we extracted 48 articles published by Michael Jordan and 29 articles published by Andrew Ng from the corpus. Figure 12 shows the result of this experiment.

From Figure.12, we found 11 topics that Michael Jordan is interested in, and 9 topics for Andrew Ng. Michael Jordan have dedicated in a variety of field, such as classification (topic 1), cluster (topic 10), language (topic 12) and image process (topic 69) and so on. Andrew Ng follows the same pattern.

We also find a strange phenomenon: 39 of 48 articles written by Jordan and 26 of 29 articles written by Ng are all related to topic 91. That is to say, almost all the articles are related to topic 91. The topic 91 actually relevant to data

modeling, which can be found in almost every article. Thus topic 91 is a common topic and could not stand for any field. Topic 62 is also a common topic about function and algorithm.

From the Figure.12, we can clearly see that even the top author (Michael I. Jordan) only publishes a small number of articles for a specific topic, so it would be highly biased to learn the dynamics of research topic of each author. So it is necessary for us to learn author groups of the corpus and then find the dynamics of research topic of author groups.

2) FINDING RESEARCH TOPIC OF AUTHOR GROUPS

For the NIPS 13-22 data set, we extracted 2516 authors and got 6114 co-author relationships. By considering co-author relationship, we can get the network shown in Figure 13.

By setting threshold $\eta = 4$, we can get the network shown in Figure 14. We finally get 39 author groups. The number of groups contains 2 authors is 29, containing 3 authors is 6, containing 4, 5, 16, 27 authors is 1 respectively.

In this section, we choose an author group obtained above, and the author group is:

{Kenji Fukumizu, Arthur Gretton, Alex Smola, Alexander Gray, Dongryeol lee,S.V.N Vishwanathan, Quoc Le, Mario March, Andrew Ng, Michael Jordan, Francis Bach, Martin Wainwright, XuanLong Nguyen, Pieter Abbeel, Pradeep Ravikumar, Novi Quadrianto}

Then we find 129 articles written by them. We also find 11 topics from the 129 articles. Figure 15 shows the dynamics of 5 (out of 11) topics.

As is shown in Figure 15, most topics are on the rising (e.g. topic 1(classification)), meaning that the group is growing actively. Meanwhile, some topics, such as topic 38(national policy) is on the falling, showing that they gradually show no interests on that field.

VII. CONCLUSION

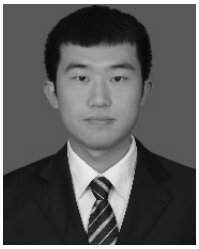
In this paper, we have presented an improved method based on LDA and have shown how our model can be used to gain insight into the dynamics of scientific papers. A number of experimental results have shown that our method has several interesting applications that can make it easier for people to deeply understand the knowledge implied in data. We can clearly explore the topic dynamics and identify the roles that words are playing in the documents. In future research, we intend to add lexical analysis to our method, treating phrases as words in LDA.

REFERENCES

[1] S. Deerwester, S. T. Dumais, G. W. Furnas, T. K. Landauer, and R. Harshman, "Indexing by latent semantic analysis," *J. Amer. Soc. Inf. Sci. Technol.*, vol. 41, no. 6, pp. 391–407, Sep. 1990.
[2] H. T. Tu, T. T. Phan, and K. P. Nguyen, "An adaptive latent semantic analysis for text mining," in *Proc. Int. Conf. Syst. Sci. Eng.*, Ho Chi Minh City, Vietnam, Jul. 2017, pp. 588–593.
[3] T. Hofmann, "Probabilistic latent semantic indexing," in *Proc. 22nd Annu. Int. Conf. Res. Develop. Inf. Retr.*, 1999, pp. 50–57.
[4] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent Dirichlet allocation," *J. Mach. Learn. Res. Arch.*, vol. 3, pp. 993–1022, Mar. 2003.
[5] Y. Liu, Z. Liu, T.-S. Chua, and M. Sun, "Topical word embeddings," in *Proc. 29th AAAI Conf. Artif. Intell.*, 2015, pp. 2418–2424.
[6] T. L. Griffiths and M. Steyvers, "Finding scientific topics," *Proc. Nat. Acad. Sci. USA*, vol. 101, no. 1, pp. 5228–5235, 2004, doi: 10.1073/pnas.0307752101.
[7] A. G. Schreyer *et al.*, "A review of scientific topics and literature in abdominal radiology in Germany—Part 1: Gastrointestinal tract," *Fortschr Röntgenstr*, vol. 188, no. 02, pp. 134–145, 2016.
[8] D. M. Blei and J. D. Lafferty, "Dynamic topic models," in *Proc. 23rd Int. Conf. Mach. Learn.*, 2006, pp. 113–120.
[9] P. Jähnichen, F. Wenzel, M. Kloft, and S. Mandt, "Scalable generalized dynamic topic models," in *Proc. 21st Int. Conf. Artif. Intell. Statist.*, 2018, pp. 1427–1435.
[10] Z. Zhao, S. Zheng, C. Li, J. Sun, L. Chang, and F. Chiclana, "A comparative study on community detection methods in complex networks," *J. Intell. Fuzzy Syst.*, vol. 35, no. 1, pp. 1077–1086, 2018.
[11] C. Fellbaum, "An electronic lexical database," *Library Quart. Inf. Community Policy*, vol. 25, no. 2, pp. 292–296, 1998.
[12] M. Parchami, B. Akhtar, and M. Dezfoulian, "Persian text classification based on K-NN using wordnet," in *Advanced Research in Applied Artificial Intelligence*, vol. 7345. Berlin, Germany: Springer, 2012, no. 2, pp. 283–291.
[13] Y. Li, C. Luo, and S. M. Chung, "A parallel text document clustering algorithm based on neighbors," *Cluster Comput.*, vol. 18, no. 2, pp. 933–948, 2015.
[14] M. J. Basha and K. P. Kaliyamurthie, "Effective linear-time document clustering in text mining using Web document categorization," *Int. J. Civil Eng. Technol.*, vol. 8, no. 10, pp. 224–234, 2017.
[15] Z. Zhao, C. Li, X. Zhang, F. Chiclana, and E. H. Viedma, "An incremental method to detect communities in dynamic evolving social networks," *Knowl.-Based Syst.*, vol. 163, pp. 404–415, Jan. 2019.
[16] G. Cai, L. Peng, and Y. Wang, *Topic Detection and Evolution Analysis on Microblog*, vol. 432. Berlin, Germany: Springer, 2014, pp. 67–77.
[17] J. Morris and G. Hirst, "Lexical cohesion computed by thesaural relations as an indicator of the structure of text," *Comput. Linguistics*, vol. 17, no. 1, pp. 21–48, 1991.
[18] C. Corley and R. Mihalcea, "Measuring the semantic similarity of texts," in *Proc. ACL Workshop Empirical Modeling Semantic Equivalence Entailment*, Stroudsburg, PA, USA, 2005, pp. 13–18.
[19] Z. Zhan and X. Yang, "Measuring semantic similarity in short texts through complex network," *J. Chin. Inf. Process.*, vol. 30, no. 4, pp. 71–81, 2016.
[20] C. Wang, D. Blei, and D. Heckerman, "Continuous time dynamic topic models," in *Proc. 24th Conf. Uncertainty Artif. Intell.*, 2012, pp. 579–586.
[21] A. Mooman, O. Basir, and A. Younes, "An intelligent model to construct specialized domain ontologies," in *Proc. 3rd Int. Conf. Comput. Sci. Inf. Technol.*, 2010, pp. 696–702.
[22] A. Musat, J. Velcin, M.-A. Rizoiu, and S. Trausan-Matu, "Concept-based topic model improvement," in *Emerging Intelligent Technologies in Industry*, vol. 369. Berlin, Germany: Springer, Jun. 2011, pp. 133–142.
[23] T. Iwata, T. Yamada, Y. Sakurai, and N. Ueda, "Online multiscale dynamic topic models," in *Proc. 16th Int. Conf. Knowl. Discovery Data Mining ACM SIGKDD*, Washington, DC, USA, 2010, pp. 663–672.
[24] L. Alsumait, D. Barbará, and C. Domeniconi, "On-line LDA: Adaptive topic models for mining text streams with applications to topic detection and tracking," in *Proc. 8th IEEE Int. Conf. Data Mining*, Dec. 2008, pp. 3–12.
[25] Y. Papanikolaou, J. R. Foulds, T. N. Rubin, and G. Tsoumakas, "Dense distributions from sparse samples: Improved gibbs sampling parameter estimators for LDA," *Statistics*, vol. 18, no. 62, pp. 1–58, 2017.
[26] S. Geman and D. Geman, "Stochastic relaxation, gibbs distributions, and the Bayesian restoration of images," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. PAMI-6, no. 6, pp. 721–741, Nov. 1984.
[27] Y. He, C. Wang, and C. Jiang, "Mining coherent topics with pre-learned interest knowledge in twitter," *IEEE Access*, vol. 5, pp. 10515–10525, Jun. 2017.
[28] J. Wang, P. Gao, Y. Ma, K. He, and P. C. K. Hung, "A Web service discovery approach based on common topic groups extraction," *IEEE Access*, vol. 5, pp. 10193–10208, Jun. 2017.
[29] M. Rosen-Zvi, T. Griffiths, M. Steyvers, and P. Smyth, "The author-topic model for authors and documents," in *Proc. 20th Conf. Uncertainty Artif. Intell.*, Banff, AB, Canada, 2004, pp. 487–494.
[30] H. Zhang, P. Nie, Y. Wen, and X. Yuan, "Authorship attribution for short texts with author-document topic model," in *Proc. Int. Conf. Knowl. Sci., Eng. Manage.* Changchun, China: Springer, 2018, pp. 29–41.

[31] Z. Zhao, W. Liu, Y. Qian, L. Nie, Y. Yin, and Y. Zhang, "Identifying advisor-advisee relationships from co-author networks via a novel deep model," *Inf. Sci.*, vol. 466, pp. 258–269, Oct. 2018.
[32] M. Yang, D. Zhu, Y. Tang, and J. Wang, "Authorship attribution with topic drift model,"in *Proc. 31st AAAI Conf. Artif. Intell.*, 2017, pp. 5015–5016.
[33] C. Li, W. K. Cheung, Y. Ye, X. Zhang, D. Chu, and X. Li, "The author-topic-community model for author interest profiling and community discovery," *Knowl. Inf. Syst. Arch.*, vol.44, no. 2, pp. 359–383, 2015.

CHAO LI received the B.S. and M.S. degrees from the Shandong University of Science and Technology and the Ph.D. degree from the Chinese Academy of Sciences, in 2014. He was a Visiting Scholar with The Hong Kong University of Science and Technology, from 2014 to 2015. He is currently a Lecturer with the Shandong University of Science and Technology. His research interests include social media, natural language processing, data mining, and network embedding learning. He is a member of CCF.



SEN FENG received the B.S. and M.S. degrees from the Shandong University of Science and Technology. His research interests include natural language processing, semantic network, data mining, and deep learning. He is a member of CCF.

QINGTIAN ZENG received the B.S. and M.Sc. degrees from the Shandong University of Science and Technology and the Ph.D. degree from the Institute of Computing Technology, Chinese Academy of Sciences, in 2005. In 2005 and 2008, he was a Visiting Fellow with the City University of Hong Kong. He is currently a Professor with the Shandong University of Science and Technology. His main research interests include process mining, Petri net, data mining, Web mining, and machine learning. He is a Senior Member of CCF.

WEIJIAN NI received the Ph.D. degree from Nankai University, in 2008. He was with the Microsoft Asia Research Institute and Baidu. He was a Visiting Scholar with the State University of New York, from 2015 to 2016. He is currently with the Shandong University of Science and Technology. His main research interests include personalized recommendation, text mining, and deep learning.

HUA ZHAO received the Ph.D. degree from the School of Computer Science and Technology, Harbin Institute of Technology, in 2008. She is currently with the Shandong University of Science and Technology. Her main research interests include topic model, sentiment compute, and opinion mining.

HUA DUAN received the B.S. and M.S. degrees from the Shandong University of Science and Technology and the Ph.D. degree from Shanghai Jiao Tong University. She is currently with the Shandong University of Science and Technology. Her main research interests include machine learning, privacy protection, and graph theory.