A Two-Stage Approach for Generating Topic Models

Yang Gao, Yue Xu, Yuefeng Li, and Bin Liu

School of Electrical Engineering and Computer Science,
Queensland University of Technology, Brisbane, Australia
{y10.gao,b5.liu}@student.qut.edu.au, {yue.xu,y2.li}@qut.edu.au

Abstract. Topic modeling has been widely utilized in the fields of information retrieval, text mining, text classification etc. Most existing statistical topic modeling methods such as LDA and pLSA generate a term based representation to represent a topic by selecting single words from multinomial word distribution over this topic. There are two main shortcomings: firstly, popular or common words occur very often across different topics that bring ambiguity to understand topics; secondly, single words lack coherent semantic meaning to accurately represent topics. In order to overcome these problems, in this paper, we propose a two-stage model that combines text mining and pattern mining with statistical modeling to generate more discriminative and semantic rich topic representations. Experiments show that the optimized topic representations generated by the proposed methods outperform the typical statistical topic modeling method LDA in terms of accuracy and certainty.

Keywords: Topic modeling, Topic representation, Tf-idf, Frequent pattern mining, Entropy.

1 Introduction

The statistical topic modeling technique has attracted big attention due to its more robust and interpretable topic representations and wide applications in the fields of information retrieval, text mining, text classification, scientific publication topic analysis and prediction[1-4] etc. It starts from Latent Semantic Analysis (LSA) [5] that can capture most significant feature of collection based on semantic structure of relevant documents. Probabilistic LSA (pLSA) [6] and Latent Dirichlet Allocation (LDA) [7] are variations to improve the interpretation of results from statistical view of LSA. These techniques are more effective on document modeling and topic extraction, which are represented by topic-document and word-topic distribution, respectively. Many topic models not only automatically extract topics from text, but also detect the evolution of topics over time [8], discover the relationship among the topics [9], supervise the topics [10] with other information (authorship, citations, et al.) for extensional applications, such as recommendation [11] and so on.

Basically, the existing statistical topic modeling approaches generate multinomial distributions over words to represent topics in a given text collection. The word distributions are derived based on word frequency in the collection. Therefore, popular words are very often chosen to represent topics. For instance, Table 1 shows an

J. Pei et al. (Eds.): PAKDD 2013, Part II, LNAI 7819, pp. 221–232, 2013.

example of multinomial word distributions used to represent four topics of a scientific publication collection. It can be seen from Table 1 that word “method” dominantly occurs across all four topics with high probability. It is obvious that “method” is a general word and very popularly used in describing research works in almost all different areas. It actually will not contribute much to uniquely represent distinctive features of any research area or topic. These kind of popular words bring a lot of confusion to the topic representation other than distinctively representing the topics.

Table 1. An example of topic representation using word distributions

Topic 0	Topic 11	Topic 12
method 0.04 , sample 0.04	**method** 0.07 , predict 0.06	classification 0.13, feature 0.08
distribute 0.04, dimension 0.03	linear 0.03, weight 0.03	accuracy 0.04, class 0.04
parameter 0.03	kernel 0.03	**method** 0.04

Except for the ambiguity problem produced by popular words, another fundamental problem is that topics are represented by multinomial distribution of isolated words which lack semantic and interpretable meaning. Although topic models can supply much information and annotate documents with the discovered topics and also supply word distribution for each topic, users still have difficulties to interpret the semantic meanings of the topics only based on the distribution of words, especially for those who are not very familiar with the related area. Mei et al. [12] and Lau et al. [13] developed automatic labeling methods for interpreting the semantics of topics by phrases. But, they heavily depend on candidate resources for labeling topics. If the topics themselves are diverse or novel to the candidate dataset, the systems will mislabel the topics. Although Lau et al. [14] labeled a topic by selecting a single term from the known distribution of words rather than candidate resources, the selected word can hardly represent the whole topic well.

In order to solve the problems of word ambiguity and semantic coherence that exist in almost all topic models, we need new model to update the topic representations. The new method should extract more distinctive representations and discover the hidden associations under multinomial words distributions. In text mining, many methods have been developed to generate text representation for a collection of documents. Most text mining methods are keyword-based approaches which use single words to represent documents. Based on the hypothesis that phrases may carry more semantic meaning than keywords, approaches to use phrases instead of keywords have also been proposed. However, investigations have found that phrase-based methods were not always superior to keyword based methods [15-17]. Recently, data mining based methods have been proposed to generate patterns to represent documents which have achieved promising results [18]. Topic modeling has the advantage of classification from large collections, while text mining is good at extracting interesting features to represent collections. So, it leads us to improve the accuracy and coherence of topic representations by utilizing text mining techniques, especially term weighting and pattern mining methods.

In this paper, a two-stage approach is proposed to combine the statistical topic modeling technique with the classical data mining techniques with the hope to improve the accuracy of topic modeling in large document collections. In stage 1, the most recognized topic modeling method Latent Dirichlet Allocation (LDA) is used to generate initial topic models. In stage 2, the most popular used term weighting method tf-idf and the frequent pattern mining method are used to derive more discriminative terms and patterns to represent topics of the collections. Moreover, the frequent patterns reveal structural information about the associations between terms that make topics more understandable, semantically relevant and cover broaden meanings.

2 Stage 1 – Topic Representation Generation

Latent Dirichlet Allocation [7] is a typical statistical topic modeling technique and the most common topic modeling tool currently in use. It can discover the hidden topics in collections of documents with the appearing words. Let $D = \{d_1, \cdots, d_M\}$ be a collection of documents, called documents database. The total number of documents in corpus is *M*. The idea behind LDA is that every document is considered involving multiple topics and each topic can be defined as a distribution over fixed vocabulary of terms that appear in documents. Specifically, LDA models a document as a probabilistic mixture of topics and treats each topic as a probability distribution over words. For the *i*th word in document *d*, denoted as $w_{d,i}$, the probability of $w_{d,i}$, $P(w_{d,i})$ is defined as:

$$P(w_{d,i}) = \sum_{j=1}^{V} P(\, w_{d,i} | z_{d,i} = Z_j) \times P(z_{d,i} = Z_j) \qquad (1)$$

where $z_{d,i}$ is the topic assignment for $w_{d,i}$, $z_{d,i} = Z_j$ means that the word $w_{d,i}$ is assigned to topic *j*, Z_j represents topic j and the *V* represents the total number of topics. Let $\boldsymbol{\phi}_j$ be the multinomial distribution over words for Z_j, $\boldsymbol{\phi}_j = (\varphi_{j,1}, \varphi_{j,2}, \cdots, \varphi_{j,n})$, $\sum_{k=1}^{n} \varphi_{j,k} = 1$. $\varphi_{j,k}$ indicates the proportion of the *k*th word in ic Z_j, that is, $\varphi_{j,i} = P(\, w_{d,i} | z_{d,i} = Z_j)$. $\boldsymbol{\theta}_d$ refers to multinomial distribution over topics in document *d*, which is *P*(*Z*). $\boldsymbol{\theta}_d = (\vartheta_{d,1}, \vartheta_{d,2}, \cdots, \vartheta_{d,V})$, $\sum_{j=1}^{V} \vartheta_{d,j}$ =1. $\vartheta_{d,j}$ indicates the proportion of topic j in document *d*. LDA is generative model that only observed variable is $w_{d,i}$, while $\boldsymbol{\phi}_j$, $\boldsymbol{\theta}_d$, $z_{d,i}$ are all latent variables that need to be estimated. Blei et al. [7] introduce Dirichlet to the posterior probability $\boldsymbol{\phi}_j$ and $\boldsymbol{\theta}_d$, which optimize the topics and documents distributions.

Among many available algorithms for estimating hidden variables, the Gibbs sampling method is a very effective strategy for parameter estimation [19, 20]. The results of LDA are at two levels, corpus level and document level. At corpus level, *D* is represented by a set of topics each of which is represented by a probability distribution over word, $\boldsymbol{\phi}_j$ for topic *j*. Overall, we have $\Phi = \{\boldsymbol{\phi}_1, \boldsymbol{\phi}_2, \cdots, \boldsymbol{\phi}_V\}$ for all topics. For illustrating the results derived by LDA, let's look at a simple example depicted in Table 2 to Table 4. Let $D = \{d_1, d_2, d_3, d_4\}$ be a small set of four documents and there are 12 words appearing in the documents. Assuming the documents in *D* involve 3 topics, Z_1, Z_2, and Z_3. Table 2 illustrates the word distribution for each of the topics. At document level, each document d_i is represented by topic distributions $\boldsymbol{\theta}_{d_i}$. For the simple example mentioned above, the document representation is illustrated in Table 3. Apart from these two level outcomes, LDA also generates word – topic assignment, that is, the word occurrence is considered related to the topics by LDA. Table 4 illustrates an example of the word-topic assignments.

Table 2. Example results of LDA: Topic representation – probability distribution over words

Topic	Φ
$\boldsymbol{\phi_1}$	w_2: $\frac{1}{3}$, w_1: $\frac{1}{5}$, w_4: $\frac{2}{15}$, w_7: $\frac{2}{15}$, w_3: $\frac{1}{15}$, w_5: $\frac{1}{15}$, w_6: $\frac{1}{15}$
$\boldsymbol{\phi_2}$	w_8: $\frac{1}{3}$, w_1: $\frac{4}{15}$, w_7: $\frac{2}{15}$, w_9: $\frac{2}{15}$, w_2: $\frac{1}{15}$
$\boldsymbol{\phi_3}$	w_{10}: $\frac{4}{13}$, w_{11}: $\frac{3}{13}$, w_1: $\frac{2}{13}$, w_7: $\frac{2}{13}$, w_4: $\frac{1}{13}$, w_{12}: $\frac{1}{13}$

Table 3. Example results of LDA: Document representation – probability distribution over topics

Document	Z_1 $(\vartheta_{d_i,1})$	Z_2 $(\vartheta_{d_i,2})$	Z_3 $(\vartheta_{d_i,3})$
d_1	0.6	0.2	0.2
d_2	0.2	0.5	0.3
d_3	0.3	0.3	0.4
d_4	0.3	0.4	0.3

Table 4. Example results of LDA: word – topic assignments

Docu-	Z_1		Z_2		Z_3	
ment	$\vartheta_{d,1}$	words	$\vartheta_{d,2}$	words	$\vartheta_{d,3}$	words
d_1	0.6	w_1,w_2,w_3,w_2,w_1	0.2	w_1,w_9,w_8	0.2	w_7,w_{10},w_{10}
d_2	0.2	w_2,w_4,w_4	0.5	w_7, w_8,w_1,w_8, w_8	0.3	w_1,w_{11},w_{12}
d_3	0.3	w_2,w_1,w_7,w_5	0.3	w_7,w_1,w_3,w_2	0.4	w_4,w_7,w_{10},w_{11}
d_4	0.3	w_2,w_7,w_6	0.4	w_9,w_8,w_1	0.3	w_1,w_{11},w_{10}

The topic representation using word distribution and the document representation using topic distribution are the most important contributions provided by LDA. The topic representation indicates which words are important to which topic and the document representation indicates which topics are important for a particular document. These representations have been widely used in various application domains such as information retrieval, document classification, text mining etc. On the other hand, the word-topic assignments also indicate which words are important to which topics, which is similar to the topic representation. However, the topic representation is at corpus level, while the word-topic assignments are at document level, which implicate more detailed or more specific association between topics and words. In this paper, we propose to mine word-topic assignments generated by LDA for more accurate or more discriminative topic representations for a given collection of documents.

3 Stage 2 – Topic Representation Optimization

For most LDA based applications, the words with high probabilities in topics' word distributions are usually chosen to represent topics. For example, the top 4 words for the 3 topics, as showed in Table 2, are: w_2, w_1, w_4, w_7 for topic 1, w_8, w_1, w_7, w_9 for topic 2 and w_{10}, w_{11}, w_1, w_7 for topic 3. From the simply example we can see that words w_1 and w_7 have relatively high probabilities for all the three topics. That means, they most likely represent general concepts or common concepts of the three topics and cannot distinctively represent the three topics. Moreover, the words in topic representations generated by LDA are individual single words. These single words provide too limited information about the relationships between the words and too limited semantic meaning to make the topics understandable. In this section, we propose two methods based on text mining and pattern mining techniques, which are detailed in the following sub sections, aiming at alleviating the mentioned problems.

3.1 Tf-idf Weighting Based Topic Modeling

The first method is based on the well-known term weighting method tf-idf (term frequency – inverse document frequency). The distinct feature of the tf-idf method is that it chooses discriminative terms to represent a document or a topic rather than popular terms. As we illustrated in the above example, there exist general or common terms in the topics' word distributions generated by LDA. We propose to utilize the tf-idf technique to process the topics' word distributions in order to generate more discriminative words to represent topics. As illustrated in Table 4, LDA generates word-topic assignments for each document, which reveal word importance to topics for that document. The basic idea of the proposed tf-idf based method is to find the discriminative words from the words which are assigned to a topic by LDA to represent that topic. There are two steps in the proposed method. The first step is to construct a collection called *topical document collection*, denoted as D_{topic}. Each document in the collection consists of all the word-topic assignments to a topic in the original document collection D. The second step is to generate a set of words for representing each document in D_{topic} by applying the tf-idf method to the collection.

(1) Construct Collection D_{topic}

Let R_{d_i,Z_j} represent the word-topic assignment to topic Z_j in document d_i. $R_{d_i,Z_j}$is a sequence of words assigned to topic Z_j in document d_i. For the example illustrated in Table 4, for topic Z_1 in document d_1, R_{d_1,Z_1} = $<w_1, w_2, w_3, w_2, w_1>$, or simply R_{d_1,Z_1} = $w_1\ w_2\ w_3\ w_2\ w_1$. Each document d'_i in D_{topic} is defined as

$$d'_i = \{R_{d_i,Z_j} | d_i \in D\} \tag{2}$$

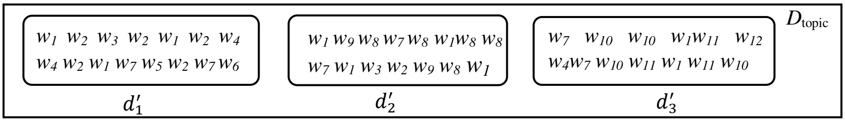

D_{topic}
$w_1\ w_2\ w_3\ w_2\ w_1\ w_2\ w_4$
$w_4\ w_2\ w_1\ w_7\ w_5\ w_2\ w_7 w_6$
d'_1
$w_1 w_9 w_8 w_7 w_8\ w_1 w_8\ w_8$
$w_7\ w_1\ w_3\ w_2\ w_9\ w_8\ w_1$
d'_2
$w_7\ w_{10}\ w_{10}\ w_1 w_{11}\ w_{12}$
$w_4 w_7\ w_{10}\ w_{11}\ w_1\ w_{11}\ w_{10}$
d'_3

Fig. 1. Dtopic with three topical documents

d'_i consists of the word-topic assignments R_{d_i,Z_j} to topic Z_j, each word-topic assignment R_{d_i,Z_j} can be treated as a sentence in the document d'_i. d'_i is called a *topical document* since it consists of the words for a particular topic. Assuming that the original document collection D has V number of topics, the collection D_{topic} is defined as $D_{\text{topic}} = \{d'_1, d'_2, \cdots, d'_V\}$. For the example given in Table 4, a topical document collection can be constructed as showed in Fig.1.

(2) Generate Document Representation for Collection D_{topic}

For the topical document, the word distribution over topic j, denoted as $(\boldsymbol{\phi}_j)_{\text{tf-idf}}$, is generated based on their tf-idf scores, which are calculated by equation (3). $tf(t_{i,j})$ is the frequency of term $t_{i,j}$ in the ith topical document, where $|d'_i|$ is the count of terms in d'_i, $N(t_{i,j})$ is the count of $t_{i,j}$ appearing in d'_i. Inverse document frequency (idf) reflects the popularity of term $t_{i,j}$ across topical documents in D_{topic}, where V is the total number of topical documents and $df(t_{i,j})$ is the document frequency. Thus, high tf-idf term weighting indicates high term frequency but low overall collection frequency.

$$tfidf(t_{i,j}) = tf(t_{i,j}) \times idf(t_{i,j}) = \frac{N(t_{i,j})}{|d'_i|} \times log\frac{V+1}{df(t_{i,j})} \tag{3}$$

Table 5 provides an example of the results which shows that, the tf-idf method weakens the effect of the common words w_1 and w_7, in the meanwhile, increases the weights for the distinctive words in each topic.

Table 5. Example results of tf-idf: Topic representation – probability distribution over words

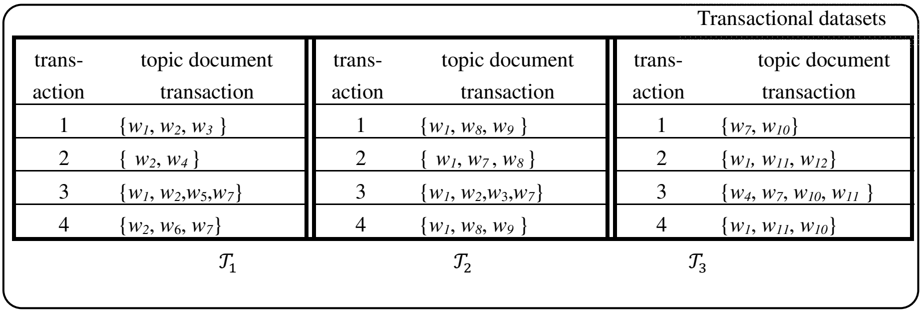
Topic	$\Phi_{\text{tf-idf}}$
$\boldsymbol{\phi}_1$	w_2: 0.1 , w_4: 0.04 , w_5: 0.04, w_6: 0.04, w_1: 0.02 , w_3: 0.02, w_7: 0.017
$\boldsymbol{\phi}_2$	w_8: 0.2 , w_9: 0.08 , w_1: 0.03, w_2: *0.02,* w_7: 0.017
$\boldsymbol{\phi}_3$	w_{10}: 0.19 , w_{11}: 0.14, w_{12}: 0.046 ,$w4$: 0.023 , w_1: 0.019 , w_7: 0.019

3.2 Pattern-Based Topic Modeling

A pattern is usually defined as a set of related terms or words. As discussed in Section 1, patterns carry more semantic meaning and are more understandable than isolated words. The idea of the pattern based representations starts from the knowledge of frequent patterns mining. It plays an essential role in many data mining tasks that try to find interesting patterns from datasets. We believe that pattern based representations can be more meaningful and more accurate to represent topics. Moreover, pattern based representations contain structural information which can reveal the association between the terms.

(1) Construct Transactional Dataset

The purpose of the proposed pattern based method is to discover associated words (i.e., patterns) from the words assigned by LDA to topics. With this purpose in mind, we construct a set of words from each word-topic assignment R_{d_i,z_j} instead of using the sequence of words in R_{d_i,z_j}, because for pattern mining, the frequency of a word within a transaction is insignificant. Let I_{ij} be a set of words which occur in R_{d_i,Z_j}, I_{ij} $=\{w|w \in R_{d_i,Z_j}\}$, i.e., I_{ij} contains the words which are in document d_i and assigned to topic Z_j by LDA. I_{ij} is called a *topical document transaction*, is a set of words without any duplicates. From all the word-topic assignments R_{d_i,Z_j} to topic Z_j, we can construct a transactional dataset $\mathcal{T}_j$. Let $D = \{d_1, \cdots, d_M\}$ be the original document collection, the transactional dataset $\mathcal{T}_j$ for topic Z_j is defined as $\mathcal{T}_j = \{I_{1j}, I_{2j}, \dots I_{Mj}\}$. For the topics in D, we can construct V transactional datasets. An example of the transactional datasets is illustrated in Fig.2, which is generated from the example in Table 4.



Transactional datasets

trans-action	topic document transaction	trans-action	topic document transaction	trans-action	topic document transaction
1	$\{w_1, w_2, w_3\}$	1	$\{w_1, w_8, w_9\}$	1	$\{w_7, w_{10}\}$
2	$\{w_2, w_4\}$	2	$\{w_1, w_7, w_8\}$	2	$\{w_1, w_{11}, w_{12}\}$
3	$\{w_1, w_2, w_5, w_7\}$	3	$\{w_1, w_2, w_3, w_7\}$	3	$\{w_4, w_7, w_{10}, w_{11}\}$
4	$\{w_2, w_6, w_7\}$	4	$\{w_1, w_8, w_9\}$	4	$\{w_1, w_{11}, w_{10}\}$

$\mathcal{T}_1$ $\mathcal{T}_2$ $\mathcal{T}_3$

Fig. 2. Transactional datasets generated from Table 4

(2) Generate Pattern-based Representation

Frequent itemsets are the most widely used patterns generated from transactional datasets to represent useful or interesting patterns. The basic idea of the proposed pattern based method is to use the frequent patterns generated from each transactional dataset $\mathcal{T}_j$ to represent topic Z_j. For a given minimal support threshold σ, and itemset p in $\mathcal{T}_j$ is frequent if $supp(p) >= \sigma$, where $supp(p)$ is the support of p which is the number of transactions in $\mathcal{T}_j$ that contain p. Take $\mathcal{T}_2$ as an example, which is the transactional dataset for topic Z_2. For a minimal support threshold $\sigma = 2$, all the frequent patterns generated from $\mathcal{T}_2$ are given in Table 6. $\{w_8\}$ and $\{w_1, w_8\}$ are the dominant patterns for topic 2. Comparing with the term based topic representation, patterns represent the associated words that carry more concrete and identifiable meaning. For instance, "data mining" is more concrete than just one word "mining" or "data".

Table 6. The frequent patterns discovered from the Z_2 topical transaction database. $\sigma = 2$

Patterns	*supp*
$\{w_8\}$,$\{w_1,w_8\}$	3
$\{w_9\}$,$\{w_8,w_9\}$,$\{w_1,w_9\}$,$\{w_1,w_8,w_9\}$,$\{w_1,w_7\}$	2

4 Experiments and Evaluation

We have conducted experiments to evaluate the performance of the proposed two topic modeling methods. In this section, we present the results of the evaluation.

4.1 Datasets

Four datasets are used in the experiments, which contain the abstracts of the papers published in the proceedings of KDD, SIGIR, CIKM and HT from 2002 to 2011. The four datasets contain 1227, 1722, 2048 and 483 abstracts, respectively. The abstracts are crawled from the ACM digital library[1], and stemmed by using Porter's stemmer package[2] in the Apache's Lucene Java.

4.2 Experiment Procedure

The whole procedure taken in the experiments is depicted in Fig. 3. The first step is dataset preparation to construct the datasets described in Section 4.1. Then in the step of topic generation, we utilize the sampling-based LDA tool provided in MALLET[3] to generate LDA topic models. The number of topics $V = 20$, the number of iterations of Gibbs sampling is 1000, the hyperparameters of LDA $\alpha = 50/V$=2.5, $\beta = 0.01$ in this experiment [20]. Step 3 is to construct the topical document datasets and the transactional datasets for optimizing topic representations, and the final step is to generate the discriminative terms based and the frequent pattern based topic representations using the pro-posed methods introduced in Section 3. We divide each dataset into training set and testing set, 90% of the documents in each dataset are used as the training set for generating topic models, while the other 10% of the documents in each dataset are left for evaluation.

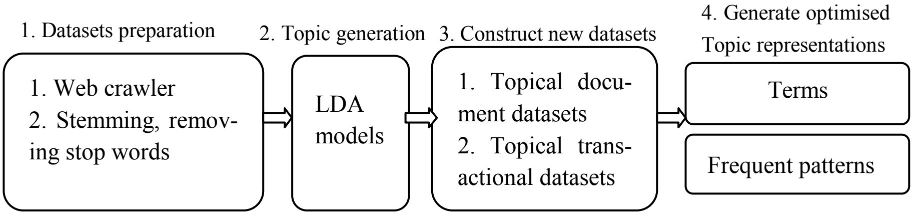

1. Datasets preparation
2. Topic generation
3. Construct new datasets
4. Generate optimised Topic representations
1. Web crawler
2. Stemming, removing stop words
LDA models
1. Topical document datasets
2. Topical transactional datasets
Terms
Frequent patterns

Fig. 3. Four steps taken for optimizing topic representation

4.3 Experiment Result Analysis

LDA is chosen as the baseline model to compare with the two proposed methods in the experiments. Table 7 demonstrates some examples of the topic representations generated by using the three models, i.e., the LDA model, the tf-idf based model, and the pattern based model. The top 12 words or patterns in each of the topic representations generated by the three models are displayed in Table 7 for two topics, topic 4 and topic 0, of dataset KDD.

[1] `http://dl.acm.org/`

[2] `http://tartarus.org/martin/PorterStemmer`

[3] `http://mallet.cs.umass.edu/index.php`. MALLET is a Java-based package for topic modeling and other machine learning applications to text.

Table 7. Examples of topics representations (topic 4 and topic 0 for dataset KDD)

Topic 4			Topic 0		
Baseline	Tf-idf	Patterns	Baseline	Tf-idf	Patterns
large	large	large	method	sample	method
algorithm	scale	algorithm	sample	dimension	distribution
compute	algorithm	compute	distribution	parameter	high
efficient	efficient	efficient	dimension	gene	sample
scale	highly	scale	parameter	distance	dimension
number	fast	number	estimate	outlier	estimate
size	size	size	distance	method	parameter
order	number	large scale	high	low	high dimension
correlate	pair	large algorithm	gene	distribution	number
highly	million	order	paper	high	sample method
local	memory	large compute	random	component	distribution method
fast	faster	large efficient	outlier	random	component

Table 8. Sample patterns in 5 topic representations for dataset KDD

Topic	Patterns
1	Probabilistic model, Information model, Text document, Topic model, Makov model
9	Clustering based algorithm, Result algorithm, Algorithm quality, Hierarchical cluster
10	Data mining, Data set, Data analysis, Data application, Data method, Data set mining
14	Web user, User search, Query search, User query, User recommendation,
18	Pattern mining, Frequent mining, Frequent patterns, Rule mining, Association mining,

From the results we can see that the top 12 words or patterns have a large overlap between each pair of the three methods, which could indicate that all the three methods can derive similar representations. But, when taking a close look, we can find that the results generated by the pattern based method provide much more concrete and specific meaning. For example, for topic 4, all the three methods rank ‘large’ as the top 1 word which is a general term. However, the pattern based method generates more specific patterns ‘large algorithm’, ‘large scale’, and ‘large compute’ which make the topic representation much easier to understand, while the other two methods cannot. Similar evidence can be seen for topic 0 as well. We have showed an example in Table 1 that the word ‘method’ was chosen by LDA for representing three topics including topic 0. In Table 7, the topic representations for topic 0 generated by the three methods are listed, from which we can see that, the ranking of the word ‘method’ was decreased by the tf-idf based method. This indicates that the word ‘method’ is not a discriminative word for uniquely representing topic 0. Moreover, the pattern based representations enrich the content of the topic representations generated by existing models such as LDA by discovering hidden associations among words, which makes the topics more detailed and comprehensive. Just for illustrating the usefulness of the pattern based method, we display in Table 8 some other patterns contained in the topic representations for dataset KDD. From the results we can see that patterns supply meaningful and semantic topic representations.

4.4 Evaluation

The ultimate goal of the proposed methods as well as other existing topic modeling methods is to represent the topics of a given collection of documents as accurately as possible. For the existing topic modeling methods and the proposed methods, the topic representations are word or pattern distributions with probabilities. The more certain the chosen words or patterns are in the topic representations, the more accurate the topic representations become. By taking this view, in this paper, we use *information entropy*, a well known certainty measurement developed in information theory, as the merit to evaluate the generalization performance of the proposed methods. Using the documents in the testing set, we compute the entropy of the topic models generated from the training set to evaluate the performance of the proposed models. The lower the entropy, the more certain the topic models to represent the topics and therefore the more predictable the documents' topics are. Formally, for a testing set D_{test}, the entropy of the topic models is defined as:

$$\text{entropy}(D_{\text{test}}) = -\sum_{z\in Z}\sum_{\text{d}\in D_{\text{test}}}\sum_{\text{w}\in d} p(w|z)p(z)\log[p(w|z)p(z)] \tag{4}$$

where $p(w|z)$ is the topic representation $\boldsymbol{\phi}_z$for a topic derived by LDA, the tf-idf based, and the pattern based methods. $p(z)$ is the document representation $\boldsymbol{\theta}_d$ generated from LDA. For the evaluation, both the tf-idf weighting and patterns supports have been normalized into probabilities. The evaluation result is presented in Table 9.

Table 9. Evaluation results on 4 datasets

Datasets	Baseline(LDA)	Tf-idf	Patterns
KDD	32.6	31.8	12.4
SIGIR	42.5	40.4	20.1
CIKM	49.7	47.7	26.6
HT	10.9	10.2	4.5

The evaluation clearly indicates that the tf-idf based model fairly achieved lower entropy values than the baseline model, meaning that, it has better performance when interpreting the meaning of the topics. Furthermore, the pattern based method achieved even much lower entropy values than any of the other two. Based on the results, we can conclude that the pattern based method apparently can generate more certain and more accurate representations for the topics of a document collection.

5 Related Work

Topic models have been extended to capture more interesting properties [7-10,19-20], but most of them represent topics by multinomial word distributions. Topic labeling [12-14] is a prevalent method to express semantic meaning of topics as mentioned in Introduction. For another example, Magatti et al. [21] present a method to calculate the similarities between given topics and known hierarchies, then choose the most agreed labels to represent the topics. However, the drawback of the existing methods of topic labeling is that they are heavily restricted to candidate resources and limited on semantic coverage. Topical *n*-gram (TNG) [22] model discovers topically-relevant phrases by Markov dependencies in word sequences based on the structure of LDA, which is relevant to our work. Except for the method of generating topic phrases, Zhao et al. [23] proposed a principled probabilistic phrase ranking algorithm for extracting top keyphrases as topic representations from the candidate phrases. The results provided in [22] and [23] show that the topics represented by the phrases are more interpretable than that of its LDA counterpart. But comparing with the pattern based representations proposed in this paper, the phrases may share low occurrences in documents, which can't achieve effective retrieval performance.

6 Conclusion

This paper proposed a two stage model to generate more discriminative and semantic rich representations for modeling the topics in a given collection of documents. The main contribution of this paper is the novel approach of combining data mining techniques and statistical topic modeling techniques to generate pattern based representations and discriminative term based representations for modeling topics. In the first stage of the proposed approach, any topic modeling method, as long as it can generate words distributions over topics, can be used to generate the initial topic representations for documents in the collection. In the second stage, we proposed to mine the initial topic representations generated in the first stage for more accurate topic representations by using the term weighting method tf-idf and the pattern mining method. Our experiment results show that the pattern based representations and the discriminative term based representations generated in the second stage are more accurate and more certain than the representations generated by the typical statistical topic modeling method LDA. Another strength provided by the pattern based representations is the structural information carried within the patterns. In the future, we will further study the structure of the patterns and discover the relationship between words which will represent the topics at a more detailed level.

References

1. Mei, Q., Zhai, C.X.: Discovering Evolutionary Theme Patterns from Text: An Exploration of Temporal Text Mining. In: KDD 2005, pp. 198–207 (2005)
2. Zhai, C.X., Velivelli, A., Yu, B.: A Cross-collection Mixture Model for Comparative Text Mining. In: KDD 2004, pp. 1285-129 (2004)
3. Wei, X., Croft, W.B.: LDA-based Document Models for Ad-hoc Retrieval. In: SIGIR 2006, pp. 178–185 (2006)
4. Wang, X., McCallum, A., Wei, X.: Topical N-grams: Phrase and Topic Discovery, with an Application to Information Retrieval. In: ICDM 2007, pp. 697–702 (2007)
5. Deerwester, S., et al.: Indexing by Latent Semantic Analysis. Journal of the American Society for Information Science 41(6), 391–407 (1990)

6. Hofmann, T.: Unsupervised Learning by Probabilistic Latent Semantic Analysis. Machine Learning 42(1), 177–196 (2001)
7. Blei, D.M., Ng, A.Y., Jordan, M.I.: Latent Dirichlet Allocation. Journal of Machine Learning Research 3, 993–1022 (2003)
8. Blei, D.M., Lafferty, J.D.: Dynamic Topic Models. In: Proceedings of the 23rd International Conference on Machine Learning, pp. 113–120 (2006)
9. Blei, D.M., Lafferty, J.D.: A Correlated Topic Model of Science. Annals of Applied Statistics 1(1), 17–35 (2007)
10. Blei, D.M., McAuliffe, J.D.: Supervised Topic Models. In: Adv. NIPS (2007)
11. Wang, C., Blei, D.M.: Collaborative Topic Modeling for Recommending Scientific articles. In: KDD 2011, pp. 448–456 (2011)
12. Mei, Q., Shen, X., Zhai, C.: Automatic Labeling of Multinomial Topic Models. In: KDD 2007, pp. 490–499 (2007)
13. Lau, J.H., et al.: Automatic Labelling of Topic Models. In: Proceeding of the 49th Annual Meeting of the Association for Computational Linguistics, pp. 1536–1545 (2011)
14. Lau, J.H., et al.: Best Topic Word Selection for Topic Labelling. In: Proceedings of the 23rd International Conference on Computional Linguistics, pp. 605–613 (2010)
15. Lewis, D.D.: An Evaluation of Phrasal and Clustered Representations on a Text Categorization task. In: SIGIR 1992, 15th ACM International Conference on Research and Development in Information Retrieval, pp. 37–50 (1992)
16. Scott, S., Matwin, S.: Feature Engineering for Text Classification. In: The 16th International Conference on Machine Learning, pp. 379–388 (1999)
17. Sebastiani, F.: Machine Learning in Automated Text Categorization. ACM Computing Surveys (CSUR) 34(1), 1–47 (2002)
18. Wu, S.-T., Li, Y., Xu, Y.: Deploying Approaches for Pattern Refinement in Text Mining. In: IEEE International Conference on Data Mining, pp. 1157–1161 (2006)
19. Steyvers, M., Griffiths, T.L.: Finding Scientific Topics. Proceedings of the National Academy of Sciences 101, 5228–5235 (2004)
20. Steyvers, M., Griffiths, T.: Probabilistic Topic Models. Handbook of Latent Semantic Analysis 427(7), 424–440 (2007)
21. Magatti, D., et al.: Automatic Labeling of Topics. In: Ninth International Conference on Intelligent Systems Design and Applications, pp. 1227–1232 (2009)
22. Wang, X., McCallum, A., Wei, X.: Topical N-grams: Phrase and Topic Discovery, with an Application to Information Retrieval. In: ICDM 2007, pp. 697–702 (2007)
23. Zhao, W.X., et al.: Topical keyphrase extraction from Twitter. In: Proceedings of 49th Annual Meeting of the Assocation for Computational Linguistics: Human Language Technologies (HLT 2011) (2011)