MPTM: A Topic Model for Multi-Part Documents

Zhipeng Xie[1,2(✉)], Liyang Jiang[1,2], Tengju Ye[1,2], and Zhenying He[1,2]

[1] School of Computer Science, Fudan University, Shanghai, China
[2] Shanghai Key Laboratory of Data Science, Fudan University, Shanghai, China
{xiezp,13210240017,13210240039,zhenying}@fudan.edu.cn

Abstract. Topic models have been successfully applied to uncover hidden probabilistic structures in collections of documents, where documents are treated as unstructured texts. However, it is not uncommon that some documents, which we call multi-part documents, are composed of multiple named parts. To exploit the information buried in the document-part relationships in the process of topic modeling, this paper adopts two assumptions: the first is that all parts in a given document should have similar topic distributions, and the second is that the multiple versions (corresponding to multiple named parts) of a given topic should have similar word distributions. Based on these two underlying assumptions, we propose a novel topic model for multi-part documents, called Multi-Part Topic Model (or MPTM in short), and develop its construction and inference method with the aid of the techniques of collapsed Gibbs sampling and maximum likelihood estimation. Experimental results on real datasets demonstrate that our approach has not only achieved significant improvement on the qualities of discovered topics, but also boosted the performance in information retrieval and document classification.

Keywords: Topic models · Gibbs sampling · Maximum likelihood estimation

1 Introduction

In classic topic models, such as probabilistic latent semantic analysis [8] and latent Dirichlet allocation [2], each document is represented as a mixture of topics, and each topic is represented as a probability distribution over words. To generate a document, we first draw a topic distribution independently from a prior Dirichlet distribution, and then for each word in that document, draw a topic randomly from the topic distribution and draw a word from that topic. Once the topic distribution is determined for a document, all the words in it follow the same generative procedure which is not affected by the location where a word appears in the document. In other words, each document is modeled as a whole, which is reflected in the fact that all the content of a document share the same topic distribution.

However, some documents are naturally composed of multiple named-parts, in the form of subdocuments or sections. Such documents are called multi-part documents in this paper. A typical example of multi-part documents is academic research papers, where each document is normally divided into sections such as *Abstract*, *Introduction*,

M. Renz et al. (Eds.): DASFAA 2015, Part II, LNCS 9050, pp. 154–168, 2015.
DOI: 10.1007/978-3-319-18123-3_10

Method, *Experimental Results*, and *Summary*. Logically, each section is self-existent. It is a relatively complete entity that describes the theme of the document from a specific aspect. For example, the section of *Introduction* is normally related to the motivation and related work of the paper, the section of *Method* describes the technical details of paper, while the section of *Experimental Results* may concern the performance measurements, the data used, and the comparison conclusion.

Our primary concern in this current study is taking this document-part structural information into consideration. To do this, we propose a novel topic modeling method for multi-part documents, called *Multi-Part Topic Model* (or MPTM in short). The MPTM model supposes that each topic has multiple versions (called versional topics) where each version corresponds to a specific named-part, while each part of a document is a mixture of the versional topics that corresponds to the part. Two underlying assumptions are also embodied in the model. The first one assumes that *all parts in the same document have similar topic distributions*. To enforce this assumption, we use one single Dirichlet distribution as the prior for all the parts of a document. Each document has its own Dirichlet prior. The mean parameters of the Dirichlet priors are normally different for different documents, but a common concentration parameter (also called the precision of the Dirichlet) is shared by all the Dirichlet priors, which controls how concentrated the distributions of multiple parts in the same document is around its mean. The second assumption is that *all versions of a single topic should have similar word distributions*, which is also enforced in a way similar to the first assumption. All versions of the same topic share a Dirichlet prior distribution, and the Dirichlet priors for different topics normally have different mean parameters.

By modeling document parts and versional topics separately, the proposed MPTM model allows us to judge the qualities of words and topics. A word that occurs in the top-word lists of (almost) all versions of a topic is thought of as a core word. On the other hand, if a word only appears frequently in one version of a topic, but seldom appears in other versions, it is then thought of as a word attached only to the particular version of the topic. Thus, each topic can be represented as a core-attachment structure, which facilitates the topic visualization. Similarly, a topic is thought of as stable and consistent, if it exhibits consistent probabilities across the multi-parts of documents; a topic is unstable or transient if its probabilities across the multi-parts of documents vary acutely. Accordingly, topic quality can be measured as the mean variance across the multi-parts averaged over all documents, which may help to prune unnecessary topics.

Finally, we evaluate MPTM model empirically on two real datasets. It is shown that the MPTM model not only generates topics of higher coherence than LDA, but also outperforms LDA in the tasks of information retrieval and document classification.

2 Related Work

A lot of existing work has been devoted to the incorporation of additional information into classic topic models, which can be broadly classified into three categories.

The first category of work *explores the correlation between topics*. Classic LDA model fails to model correlation between topics, because of the (nearly) independence assumptions implicit in the Dirichlet distribution on the topic proportions. To model the fact that the presence of one topic is sometimes correlated with the presence of another, [3] replaces the Dirichlet by the more flexible logistic normal distribution that incorporates a covariance structure among the components (or topics), [12] introduces the pachinko allocation model (PAM) that uses a DAG structure to represent and learn arbitrary-arity, nested and possibly sparse topic correlations, and [18] proposes a Latent Dirichlet-Tree Allcoation (LDTA) model that employs a Dirichlet-Tree prior to replace a single Dirichlet prior in LDA.

The second category pays attention to the relationships among words. The DF-LDA model [1] can incorporate the knowledge about words in the form of must-links and cannot-links using a novel Dirichlet Forest prior. Jagarlamudi et al. [9] proposes the Seeded-LDA model, allowing the user to specify some prior seed words in some topics. Chen et al. [6] proposes MC-LDA to deal with the knowledge of m-set (a set of words that should belong to the same topic) and c-set (a set of words that should not be in the same topic).

The third category focuses on the document level, to incorporate certain additional information in the topic modeling. Supervised LDA [4], DiscLDA [10], and Labeled LDA [17] try to predict the label values for input documents, based on labeled documents. TagLDA [19] extends latent Dirichlet allocation model by using a factored representation to combine the text information and tag information. Polylingual topic model [13] deals with polylingual document tuples, where each tuple is a set of documents loosely equivalent to each other, but written in different languages. It assumes that the documents in a tuple share the same tuple-specific distribution over topics, and each "topic" consists of a set of discrete word distributions, one for each language.

Our work falls into the third category, in that it makes an attempt to incorporate the information of document-part relationships into topic modeling. To the best of our knowledge, no previous work has attempted to incorporate the document-part structural information into the topic extraction problem. Our work is thus orthogonal to the previous work and complements them.

3 Multi-Part Topic Model

3.1 Generative Process

We now introduce the multi-part topic model (MPTM), an extension of latent Dirichlet allocation (LDA). Assume that there are D documents containing T topics expressed over W unique words, where each document contains P named-parts. Each document is represented as a set of P multinomial distributions over topics, where each part p of document d corresponds to one multinomial distribution over topics,

denoted as $\psi_d^p = P(z|d,p)$. Each topic has multiple versions, and each versional topic is a multinomial distribution over words. For a given topic t, its versional topic corresponding to named-part p is denoted as $\varphi_t^p = P(w|t,p)$.

We first assume that *all parts within a document d should be similar in their topic distributions*, since they normally concern a common theme, and describe the theme from different aspects. In MPTM model, we enforce this assumption by requiring that all parts within a document d have their topic distributions drawn from a common prior Dirichlet distribution. The mean parameter θ_d of the Dirichlet distribution is exactly the mean of the Dirichlet distribution, which is specific to document d; while the concentration parameter s (also call precision parameter) controls how concentrated the Dirichlet distribution is around its mean θ_d, which is a hyperparameter in MPTM model.

Furthermore, we also assumed that *all versions of a topic should be similar in their word distributions*. It is enforced in MPTM model by requiring that all versions of a topic t have their word distributions drawn from a common prior Dirichlet distribution. The mean parameter ϕ_t of the common Dirichlet distribution is specific to topic t, while the concentration parameter c is also a hyperparameter that controls how concentrated the Dirichlet distribution is around its mean ϕ_t.

Table 1. Notations used

Notation	Meaning
D	the number of documents
T	the number of topics
W	the number of words in the vocabulary
P	the number of named-parts
d	a document
t	a topic
w	a word
p	a named part
ψ_d^p	the topic distribution of the part p in document d
φ_t^p	the word distribution of the version p for topic t
ϕ_t	the mean parameter of the prior Dirichlet distribution for the word distributions of versions of topic t
θ_d	the mean parameter of the prior Dirichlet distribution for the topic distributions of all parts in document d
c	the concentration hyperparameter of the prior Dirichlet distribution for the word distributions of versions of any topic
s	the concentration hyperparameter of the prior Dirichlet distribution for the topic distributions of all parts in any document

The values of s and c play an important role in our model. As we increase the value of s, all parts of a document have increasing concentration, which tends to generate similar topic distributions for those parts. As we increase the value of c, all versions of a topic have increasing concentration, which tends to get similar word distributions of those versions. When s and c go to infinity, the topic distributions of all the parts within a same topic will be constrained to be the same one, and the multi-part topic modeling method reduces to the classic topic modeling method applied on the documents. On the other hand, when s and c go to zero, there will be no constraints on the topic distributions, and the multi-part topic modeling method degenerates to the classic topic modeling methods applied on all the subdocuments where each subdocument is treated as an independent document.

The notations used in this paper are summarized in Table 1. The generative process for MPTM is given as follows:

1	For each topic $t \in \{1, \dots, T\}$:
2	For each part $p \in \{1, \dots, P\}$:
3	Draw $\varphi_t^p \sim Dirchlet(\phi_t, c)$
4	For each document $d \in \{1, \dots, D\}$:
5	For each part $p \in \{1, \dots, P\}$:
6	Draw $\psi_d^p \sim \text{Dirichlet}(\theta_d, s)$
7	For each word $w_{d,p,n}$ in part p of document d
8	Draw $z_{d,p,n} \sim Multinomial(\psi_d^p)$
9	Draw $w_{d,p,n} \sim Multinomial(\varphi_{z_{d,p,n}}^p)$

In MPTM model, the parameters include the mean vectors $\phi_t (1 \le t \le T)$ and the mean vectors θ_d $(1 \le d \le D)$, which we treat for now as fixed quantities and are to be estimated. When the parameters are fixed, for each versional topic (t, p) (line 1), lines 2-3 draw a multinomial distribution over words φ_t^p (a versional topic) for each named part p. For each part p in each document d (lines 4-5), we first draw its multinomial distribution over topics ψ_d^p (line 6), and then generate all the words in part p of document d (lines 7-9) in the following way: for each word, a topic $z_{d,p,n}$ is randomly drawn from ψ_d^p, and then a word $w_{d,p,n}$ is chosen randomly from $\varphi_{z_{d,p,n}}^p$.

The plate notation for MPTM is given in Fig. 1. As we will see in Section 3, this model is quite powerful in improving the quality of discovered topics and boosting performance of information retrieval and document classification.

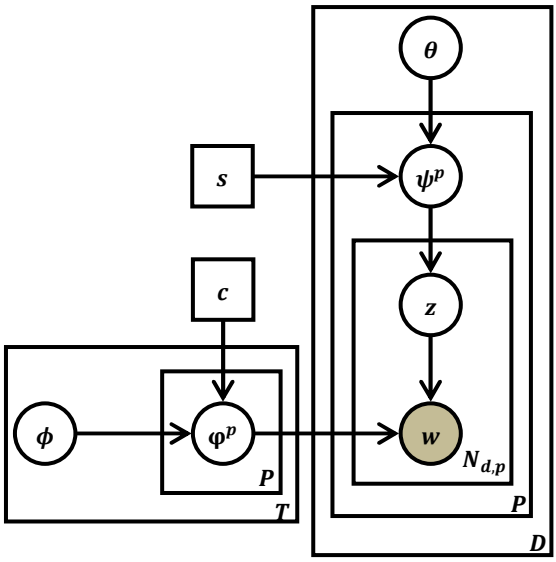

Fig. 1. Plate notation of MPTM model

3.2 Inference and Parameter Estimation

As we have described the motivation behind MPTM and its generative process, we now turn our attention to the detailed procedures for inference and parameter estimation under MPTM. In MPTM, the main parameters of interest to be estimated are the mean vectors $\theta_d\ (1 \le d \le D)$ and $\phi_t\ (1 \le t \le T)$ of the Dirichlet distributions. Other variables of interest include the word distributions φ_t^p of the multiple versions of a topic t, and the topic distribution ψ_d^p of the parts of a document d. Instead of directly estimating the variables φ_t^p and ψ_d^p, we estimate the posterior distribution over topics for the given observed words w, using Gibbs sampling, and then approximate φ_t^p and ψ_d^p using posterior estimates of topics for the observed words. Once φ_t^p and ψ_d^p are approximated, the parameters θ_d and ϕ_t can be estimated with a maximum likelihood procedure for Dirichlet distributions. The algorithmic skeleton for the parameter estimation in MPTM is briefly listed in Table 2.

Table 2. The framework of inference and parameter estimation for MPTM model

Step 1.	Initialize the parameters θ and ϕ
Step 2.	Sampling the hidden variables z with a collapsed Gibbs sampler
Step 3.	Update the parameters
Step 4.	Repeat the steps 2 and 3 for a fixed number of times

Next, we examine the details of the framework step by step, as follows.

Step 1. Initialization of parameters $\boldsymbol{\theta}$ and $\boldsymbol{\phi}$

To initialize the parameters θ and ϕ, we apply standard latent Dirichlet allocation by using collapsed Gibbs sampling algorithm [7]. We use a single sample taken after 300 iterations of Gibbs sampling to initialize the values of parameters θ_d $(1 \le d \le D)$ and parameters ϕ_t $(1 \le t \le T)$, in the MPTM model.

Step 2. Collapsed Gibbs sampler for latent variables z

We represent the collection of documents by a set of word indices w_i, document indices d_i, and part indices p_i, for each word token i. The Gibbs sampling procedure considers each word token in the text collection in turn, and estimates the probability of assigning the current word token to each topic, conditioned on the topic assignments to all other word tokens. The Gibbs sampler is given by:

$$P(z_i = t|\mathbf{z}_{-i}, w_i, d_i, p_i) \propto \frac{n_{-i,t}^{(w_i,p_i)} + c\phi_{tw_i}}{n_{-i,t}^{(\cdot,p_i)} + c} \cdot \frac{n_{-i,t}^{(d_i,p_i)} + s\theta_{d_i t}}{n_{-i,\cdot}^{(d_i,p_i)} + s} \quad (1)$$

where the subscript "$-i$" means the exclusion of the current assignment of z_i, $n_{-i,t}^{(w_i,p_i)}$ denotes the number of times that word w_i from part p_i has been assigned to topic t, $n_{-i,t}^{(d_i,p_i)}$ denotes the number of times that a word from the part p_i of document d_i has been assigned to topic t, $n_{-i,t}^{(\cdot,p_i)} = \sum_w n_{-i,t}^{(w,p_i)}$ denotes the number of times that a word from all the part p_i has been assigned to topic t, and $n_{-i,\cdot}^{(d_i,p_i)} = \sum_j n_{-i,j}^{(d_i,p_i)}$ denotes the length of the part p_i of the document d_i.

To better understand the factors that affect topic assignments for a particular word, we can examine the two parts of Equation 1. The left part is the probability of word w_i under the part p_i version of topic t; whereas the right part is the probability that topic t has under the current topic distribution for part p_i of document d_i. Therefore, words are assigned to topics according to how likely the word in the part is for a topic, as well as how dominant a topic is in a part of a document. Clearly, the information of which part a word does occur plays an important role in determining its topic assignment.

The Gibbs sampling algorithm gives direct estimates of z for every word. Based on these estimates, the word distributions φ_t^p for part p version of topic t can be estimated from the count matrices as:

$$\varphi_{tw}^p = \frac{n_t^{(w,p)} + c\phi_{tw}}{n_t^{(\cdot,p)} + c}; \quad (2)$$

while topic distributions ψ_d^p for the part p of the document d can be estimated as:

$$\psi_{dt}^p = \frac{n_t^{(d,p)} + s\theta_{dt}}{n_\cdot^{(d,p)} + s}. \quad (3)$$

Once the word distributions of all the versions for a topic and the topic distributions for all parts of a document are calculated in Equations (2) and (3), we can then update (or re-estimate) the mean parameters of the prior Dirichlet distributions in Step 3.

Step 3. How to re-estimate the parameters θ and ϕ?

Assume that a random vector, $\mathbf{p} = (p_1, \ldots, p_K)$, whose elements sum to 1, follows from a Dirichlet distribution with mean vector parameter $\mathbf{m} = (m_1, \ldots, m_K)$ that satisfying $\sum_k m_k = 1$ and concentration parameter s. The probability density at $\mathbf{p}$ is

$$p(\mathbf{p}) \sim \text{Dirichlet}(\mathbf{m}, s) = \frac{\Gamma(\sum_k sm_k)}{\prod_k \Gamma(sm_k)} \prod_k (p_k)^{sm_k - 1} \tag{4}$$

where the concentration parameter s, also referred to as the precision of the Dirichlet, controls how concentrated the distribution is around it mean.

In the context of MPTM model, we want to fix the concentration parameter s and only optimize the mean parameter $\mathbf{m}$ in the maximum-likelihood objective from the observed random vectors $\{\mathbf{p}_1, \ldots, \mathbf{p}_N\}$. To perform this problem, we adopt the fixed-point iteration technique to compute the maximum likelihood solution [15], by iterating the following two steps until convergence:

$$\Psi(\alpha_k) = \log \bar{p}_k - \sum_j m_j^{old} \left(\log \bar{p}_k - \Psi\left(sm_j^{old}\right) \right) \tag{5}$$

and

$$m_k^{new} = \frac{\alpha_k}{\sum_j \alpha_j} \tag{6}$$

where $\log \bar{p}_k = \frac{1}{N} \sum_i \log p_{ik}$, and $\Psi(x) = \frac{d \log \Gamma(x)}{dx}$ is known as the digamma function.

The problem of finding maximum likelihood solution for mean parameter of Dirichlet distribution (with fixed concentration parameter) exists in two places of MPTM model:

For each part p of a document d, its topic distribution ψ_d^p follows from a prior Dirichlet distribution with mean parameter θ_d and concentration parameter s, given algebraically as:

$$\psi_d^p \sim \text{Dirichlet}(\theta_d, s) = \frac{\Gamma(\sum_t s\theta_{dt})}{\prod_t \Gamma(s\theta_{dt})} \prod_t \left(\psi_{dt}^p\right)^{s\theta_{dt} - 1} \tag{7}$$

For versional topic with respect to part p for a topic t, its word distribution φ_t^p follows from a prior Dirichlet distribution with mean parameter ϕ_t and concentration parameter c, given as:

$$\varphi_t^p \sim \text{Dirichlet}(\phi_t, c) = \frac{\Gamma(\sum_w c\phi_{tw})}{\prod_w \Gamma(c\phi_{tw})} \prod_w \left(\varphi_{tw}^p\right)^{c\phi_{tw} - 1} \tag{8}$$

The above fixed-point iteration technique is used in the MPTM model to estimate θ_d $(1 \le d \le D)$ and ϕ_t $(1 \le t \le T)$, respectively.

Step 4. Repeat the steps 2 and 3 a fixed number of times

The final step is simply to repeat the steps 2 and 3 for a fixed number of times and output the word distributions of all versional topics and the topic distributions of all parts of documents.

4 Core Words and Topic Quality

If a word appears in the top-M word lists of (almost) all versions of a topic, it is called a core word of the topic; otherwise, it is called an attached word. Thus, each version of a topic can be represented as a core-attachment structure, where the attachment represents the part-specific words.

The "core words" embodies the meaning of the topic throughout the text, they can help us understand the name of the topic clearly. While the "part-specific" attached words complement the details of the topic from different aspects, different parts may have different emphasis.

Table 3. Two exemplar core-attachment structures

Core words:	featur word relat label topic translat learn model data method
Abstract	semant paper approach propos text perform task improv languag base extract set
Introduction	approach semant task languag text sentenc work extract system tag document
Method	set term sentenc train document context exampl text tag entiti select
Experiments	tabl set train perform system evalu test baselin term select base
Summary	work approach improv system perform select achiev propos better support

Core words:	network predict social system item rate tag user recommend model method matrix
Abstract	propos effect realworld interest work novel develop review provid
Introduction	product trust work base propos person interest opinion
Method	function time set vector denot product number base group algorithm
Experiments	set data perform dataset review figur number evalu random experi paramet
Summary	represent work base propos reput trust evalu data

Let us take the IJCAI corpus as an example, where each topic has 5 versions (please refer to section 5 for the details of the IJCAI dataset). We set M=20, and define a word to be a core word for a topic if it occurs in the top-20 word lists of at least 4 versions of the topic. Table 3 illustrates the core-attachment structures of two exemplar topics in the MPTM model of IJCAI dataset.

The first example is a topic about "topic model". The core words include "topic", "model", "word", "feature", "label", etc., which well reflect the common characteristics of the topic. The words "term", "sentence", "document" appear as the attached words to the "Method" part, reflecting the technical details of the topic. The words "performance", "evaluation", "baseline", "train", and "set" are listed as the attached words to "Experiments" part. Similar analysis also applies to the second example, which is omitted here.

After analyzing the word distributions of versional topics, let us examine the topic distributions of document parts. A topic is thought of as a stable and consistent topic, if it exhibits consistent probabilities across the multi-parts of documents; otherwise, it is unstable or transient. Here, we measure the quality of a topic simply as the mean variance across the multi-parts averaged over all documents:

$$mVar(t) = \frac{1}{D}\sum_{d=1}^{D}\sum_{p=1}^{P}\left(\psi_{dt}^{p} - \frac{1}{P}\sum_{i=1}^{P}\psi_{dt}^{i}\right)^{2} \tag{9}$$

In the experiment with information retrieval, it will be shown that the pruning of topics with highest mean variance can further improve the performance of MPTM.

5 Experimental Results

In this section, we evaluate the proposed MPTM model against several baseline models on two real datasets.

5.1 Data Sets

Two datasets (IJCAI and NIPS) are used in the experiments. The first dataset IJCAI is constructed by ourselves, using papers from the most recent three Proceedings of International Joint Conference on Artificial Intelligence (IJCAI) in years 2009, 2011, and 2013, because the IJCAI conferences in these three years share (almost) the same track organization, and the information of the assignments of papers to tracks can serve as external criterion for measuring the performance in information retrieval and document classification. We extracted 669 papers from 6 common tracks in total, with detailed information listed in Table 4.

The NIPS corpus contains 1740 papers published in the Proceedings of Neural Information Processing Systems (NIPS) Conferences[1] from year 1988 to year 2000.

[1] The dataset is available at the NIPS Online Repository. `http://nips.djvuzone.org/txt.html`.

Table 4. The IJCAI Corpus

Track Name	# papers
Agent-based and Multiagent Systems	165
Constraints, Satisfiability, and Search	107
Knowledge Representation, Reasoning, and Logic	181
Natural-Language Processing	74
Planning and Scheduling	74
Web and Knowledge-based Information Systems	68
Sum	**669**

The IJCAI and NIPS papers have been preprocessed to remove the "*References*" part, to remove stop words, and to do word stemming. Each IJCAI paper is split into parts of "*Abstract*", "*Introduction*", "*Method*", "*Experiments*", and "*Summary*"; while each NIPS paper is simply split into 3 parts of equal length, called "*Head Part*", "*Middle Part*", and "*Tail Part*" respectively. After preprocessing, the IJCAI corpus contains 1,437,916 words with vocabulary size as 32,752, and the NIPS corpus contains 2,014,937 words with vocabulary of size 15,965.

Throughout the experiments, the MPTM models were trained using 1500 Gibbs iterations where the parameters get updated for every 300 iterations. That is, the step 2) in the algorithmic framework executes 300 iterations of collapsed Gibbs sampling, while the outer loop of steps 2) and 3) is repeated 5 times.

5.2 Topic Coherence

As indicated in [5][16][11], the perplexity measure does not reflect the semantic coherence of individual topics and can be contrary to human judges. The topic coherence measure [14] was proposed as a better alternative for assessing topic quality, which only relies upon word co-occurrence statistics within the documents, and does not depend on external resources or human labeling. Given a topic t, if $V^{(t)} = (v_1^{(t)}, \dots, v_1^{(M)})$ is its top-M word list, the topic coherence is defined as:

$$Coherence\left(t; V^{(t)}\right) = \sum_{m=2}^{M} \sum_{l=1}^{m-1} \log \frac{D\left(v_m^{(t)}, v_l^{(t)}\right) + 1}{D\left(v_l^{(t)}\right)}$$

where $D(v)$ denotes the document frequency of word v and $D(v, v')$ denotes the number of documents containing both words v and v'. For a topic in LDA model, its top-M words are the M most probable words in the topic. In our MPTM model, because each topic has multiple versions, we define its top-M words in an intuitive manner as follows.

For a given topic t, let $\tau(\varphi_t^p, w)$ denote the position or rank of word w in the word distribution of versional topic φ_t^p. We use $cnt(t, w) = |\{p: 1 \le p \le P, \tau(\varphi_t^p, w) \le M\}|$ to denote the number of versions of topic t that w occurs in its top M words, and use $sr(t, w) = \sum_p \tau(\varphi_t^p, w)$ to denote the sum of the ranks of

word w in all versions of topic t. A word w is ranked before another word v with respect to a topic *t*, if it satisfies one of the following two conditions:

(1) $cnt(t,w) > cnt(t,v)$
(2) $cnt(t,w) = cnt(t,v)$ and $sr(t,w) < sr(t,v)$.

Accordingly, for each topic, the top-*M* ranked words can be calculated.

Table 5. Average Topic Coherence scores across different numbers of topics

Data Set	# Topics	LDA	MPTM	Improved Percentage
IJCAI	20	−140.7	−123.0	12.6%
	50	−210.5	−187.3	11.0%
	100	−256.6	−225.4	12.2%
NIPS	20	−154.1	−146.4	5.0%
	50	−181.6	−167.5	7.8%
	100	−210.9	−185.6	12.0%

Table 5 shows the topic coherence averaged over all topics. It can be seen that the topic coherences of MPTM models are significantly higher than those of LDA models, indicating higher quality of topics with MPTM. The improvement percentage on NIPS is less significant than IJCAI, which may be caused by the fact that the documents in NIPS are split into three parts of equal length, and it does not reflect the exact document-part relationships.

5.3 Information Retrieval

For information retrieval applications, the task is to retrieve the most relevant documents to a query document. Here we make use of the cosine similarity to measure the relevance between two documents. Mean Average Precision (MAP), for its especially good discrimination and stability, is adopted as the measure of quality to evaluate the performance of MPTM model in information retrieval.

If the set of relevant documents for a query document q_j is $\{d_1, \ldots, d_{m_j}\}$, and R_{jk} is the set of ranked retrieval results from the top result until you get to document d_k, then

$$\mathrm{MAP(Q)} = \frac{1}{|Q|}\sum_{j=1}^{|Q|}\frac{1}{m_j}\sum_{k=1}^{m_j} Precision(R_{jk}).$$

To check the effects of different configurations of topic number K, parameters s, and parameter c, we have tested the MPTM model on a grid of configurations with $K \in \{5, 10, 20, 30, 50\}$, $s \in \{50, 100, 200, 400\}$, and $c \in \{50, 100, 200, 400\}$. In all the configurations, our model has consistently outperformed the LDA model. However, for different K values, the configuration at which our model obtained the best performance may vary. Without fine-tuning the parameters, we just report the MAP values with the configuration of $s = 200$ and $c = 100$, in Figure 2. Here, five-fold cross validation is conducted, where for each fold, 80% of the documents are used as the training data, and the other 20% are held-out as the query data.

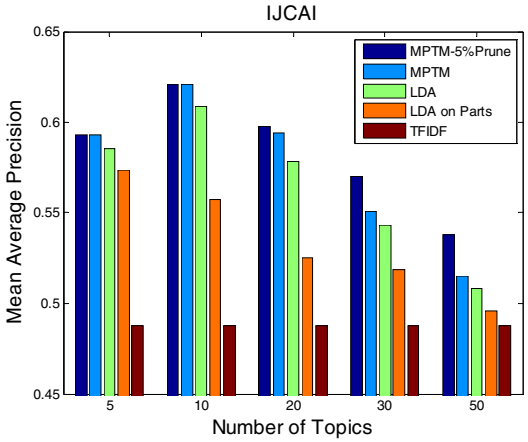

Fig. 2. Average MAP scores across different numbers of topics

In Figure 2, TFIDF method is to represent the documents using a vocabulary of 8000 words with the highest TF-IDF values; LDA on Parts method builds a LDA model by treating each part as an independent document, and then concatenate the topic distributions of all the parts of a document into a $P \times T$-dimensional representation of the document; and MPTM-5%Prune has pruned the 5% topics with highest mean variance for MPTM model.

We can observe that MPTM has achieved higher MAP values than the baseline models, and *MPTM-5%Prune* can further boost the performance of MPTM, with the aid of quality measures of topics.

5.4 Document Classification

The existence of track information associated with each document in the IJCAI corpus has also made it possible to classify a new document into the six tracks. On IJCAI corpus, five-fold cross validation is conducted as follows. At each fold, 80% of the documents are used as the training data, and the other 20% are held-out as the test data. On the training data, we train a MPTM model, with which each training or test document d can be transformed into a vector of length $T \times P$ by concatenating all the $\psi_d^p: 1 \leq p \leq P$. We then train a support vector machine (SVM) on the $(T \times P)$-dimensional representations of training documents provided by MPTM, and use it to classify the test documents. This SVM is compared with an SVM trained on the features provided by LDA. Both SVMs are trained with the libSVM software [5] and get optimized by a grid search with parameter ranges of $10^{-2} \leq C, \gamma \leq 10^4$. The mean accuracy averaged over five folds is reported in Figure 3.

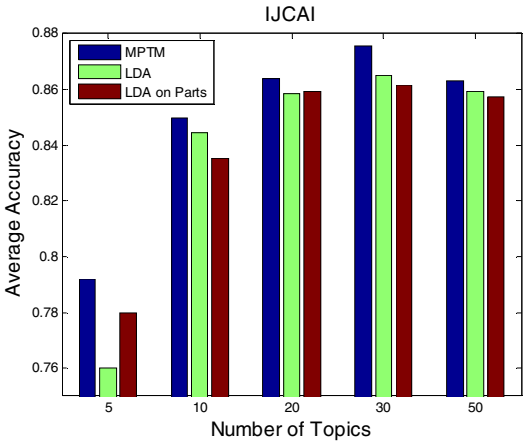

Fig. 3. Average accuracies across different numbers of topics

It can be seen from the results that the accuracy is improved in all cases, which suggests that the features provided by MPTM may be more informative in the task of document classification.

6 Conclusions

This paper proposed a novel method to exploit the multi-part composition information of documents for producing better-quality topics. To the best of our knowledge, this has not been done before. To model the multi-part documents, a novel topic model called MPTM is proposed by taking two assumptions such that *all parts within a document d should be similar in their topic distributions* and *all versions of a topic should be similar in their word distributions*. It has been manifested empirically by two datasets that MPTM has successfully produced topics of high quality, and outperformed the baseline methods in information retrieval and document classification tasks.

Finally, it is possible to remove the existence of multiple versions of topics from the MPTM model, in order to widen its applicability. For example, for a corpus where documents are labeled, it is expected to make sense to assume that the topic distributions of the documents with the same class label be drawn from a common Dirichlet prior, that is to say, to assume that the documents with same class label have similar topic distributions. Such a model can make use of the supervised information in topic modeling and may make contribution in solving the document classification task.

Acknowledgements. This work is supported by National High-tech R&D Program of China (863 Program) (No. SS2015AA011809), Science and Technology Commission of Shanghai Municipality (No. 14511106802), and National Natural Science Fundation of China (No. 61170007). We are grateful to the anonymous reviewers for their valuable comments.

References

1. Andrzejewski, D., Zhu, X., Craven, M.: Incorporating domain knowledge into topic modeling via Dirichlet Forest priors. In: ICML, pp. 25–32 (2009)
2. Blei, D.M., Ng, A.Y., Jordan, M.I.: Latent Dirichlet allocation. Journal of Machine Learning Research **3**, 993–1022 (2003)
3. Blei, D., Lafferty, J.: Correlated topic models. Advances in neural information processing systems **18**, 147–154 (2006). MIT Press, Cambridge, MA
4. Blei, D., McAuliffe, J.: Supervised topic models. (2010). arXiv preprint arXiv:1003.0783
5. Chang, C.C., Lin, C.J.: LIBSVM: a library for support vector machines. ACM Transactions on Intelligent Systems and Technology **2**(3), 27 (2011)
6. Chen, Z., Mukherjee, A., Liu, B., Hsu, M., Castellanos, M., Ghosh, R.: Exploiting domain knowledge in aspect extraction. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP 2013) (2013)
7. Griffiths, T.L., Steyvers, M.: Finding scientific topics. Proc. Natl. Acad. Sci. U.S.A. **101**(Suppl 1), 5228–5235 (2004)
8. Hofmann, T.: Probabilistic latent semantic indexing. In: Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval, pp. 50–57 (1999)
9. Jagarlamudi, J., Daumé III, H., and Udupa, R.: Incorporating lexical priors into topic models. In: Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics, pages 204–213 (2012)
10. Lacoste-Julien, S., Sha, F., and Jordan, M.: DiscLDA: discriminative learning for dimensionality reduction and classification. In: Advances in Neural Information Processing Systems, pp. 89–904 (2008)
11. Lau, J.H., Baldwin, T., Newman, D.: On collocations and topic models. ACM Transactions on Speech and Language Processing (TSLP) **10**(3), 10 (2013)
12. Li, W., McCallum, A.: Pachinko allocation: DAG-structured mixture models of topic correlations. In: Proceedings of the 23rd International Conference on Machine Learning, pp. 577–584 (2006)
13. Mimno, D., Wallach, H.M., Naradowsky, J., Smith, D.A., McCallum, A.: Polylingual topic models. In: Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing, pp. 880–889 (2009)
14. Mimno, D., Wallach, H.M., Talley, E., Leenders, M., McCallum, A.: Optimizing semantic coherence in topic models. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing, pp. 262–272 (2011)
15. Minka, T.: Estimating a Dirichlet distribution. Technical Report (2012). http://research.microsoft.com/en-us/um/people/minka/papers/dirichlet/minka-dirichlet.pdf
16. Newman, D., Lau, J.H., Grieser, K., Baldwin, T.: Automatic evaluation of topic coherence. In: Human Language Technologies: Proceedings of the 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics, pp. 100–108 (2010)
17. Ramage, D., Hall, D., Nallapati, R., Manning, C.D.: Labeled LDA: a supervised topic model for credit attribution in multi-labeled corpora. In: Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing, vol. 1, pp. 248–256 (2009)
18. Tam, Y.-C., Schultz, T.: Correlated latent semantic model for unsupervised LM adaptation. In: IEEE International Conference on Acoustics, Speech and Signal Processing, pp. 41–44 (2007)
19. Zhu, X., Blei, D., Lafferty, J.: TagLDA: bringing document structure knowledge into topic models. Technical Report TR-1553, University of Wisconsin (2006)