A Comparative Study of Topic Models for Topic Clustering of Chinese Web News

Yonghui Wu\textsuperscript{1,2}, Yuxin Ding\textsuperscript{2}, Xiaolong Wang\textsuperscript{1,2}, Jun Xu\textsuperscript{2}
\textsuperscript{1}Harbin Institute of Technology, Harbin, PRC
\textsuperscript{2}Key Laboratory of Network Oriented Intelligent Computation, Harbin Institute of Technology, Shenzhen Graduate School, Shenzhen, PRC
{yhwu,wangxl}@insun.hit.edu.cn, yxding@hitsz.edu.cn, hit.xujun@gmail.com

Abstract—Topic model is an increasing useful tool to analyze the semantic level meanings and capture the topical features. However, there is few research about the comparative study of the topic models. In this paper, we describe our comparative study of three topic models in the extrinsic application of topic clustering. The topic model distance is defined on the converged parameters of topic models, which is used in the topic clustering. Then, the topic models are compared using the clustering result of the corresponding topic distance matrix. A series of comparative experiments are carried on a corpus containing 5033 web news from 30 topics using the cosine distance as the base-line. Web page collections with different number of topics and documents are used in experiments. The experiment results show that topic clustering using topic distance achieves a better precision and recall in the data set containing related topics. The topic clustering using topic distance benefits from the topic features captured by topic models. The complex topic model does provide further help than the simple topic model in topic clustering.

Index Terms—comparative study, topic model, clustering, distance measure

I. INTRODUCTION

The developing of Internet brings us plenty of text information. Users can get the related information by the help of search engines. However, many users are not satisfied by the information only related by keywords. They need more detailed meanings behind the text, which is another challenging task: to capture the meaning or semantic level information from the web pages, and categorize the web pages into different topical classes. Clustering methods are widely used in this task.

Distance measure is the key point of clustering. The Vector Space (VS) model is widely used in the representation of text documents. News documents are represented as bag of words in the VS model. The similarity between documents is generated from the VS model using different similarity measure methods, such as the cosine-method. However, the definition of a topic is more complex than the definition of a category in text classification. When the clustering method is used in topic detection more topical rich features has to be extracted to achieve a better performance. This distance measure method also suffers a decrease when the number of topics and documents grows.

Topic model is an increasing useful tool to analyze the semantic level meanings and capture the topical features of the text.

By assuming the distribution of word is different in topics, the word distribution in each topic can be treated as a topic model (TM). Then, each document can be treated as a sample of the mixture of \( k \) topic models. Different Topic Models can be derived by specifying different probabilistic generating procedures of the documents. Using Expectation Maximization method, the parameters of the TM can be estimated from the collection of documents. Topic Models has been used in many kinds of applications, such as text segmentation \([1]\), summarization \([2]\), part of speech tagging \([3]\), word disambiguation \([4]\), etc.

However, there are few research about the comparative study of the topic models. As an unsupervised method, the selection of topic models is difficult which makes the comparative study of topic models very important. Generally speaking, the topic models can be compared in two methods.

First, as a generative probabilistic model, the topic models can be evaluated according to their generalization capability on a collection of corpus. Wallach et al. \([5]\) gives a evaluation method of this generative ability using Latent Dirichlet Allocation (LDA).

Second, the topic models can be compared in extrinsic tasks, such as document classification, information retrieval and topic detection. Georgescul et al. \([1]\) make a comparative study of Probabilistic Latent Semantic Analysis (PLSA) and LDA in automatic topic segmentation of dialogues under different latent space dimensions. The accuracy under different latent dimensions are compared, which is a comparison of topic models by feature space reduction. Xing and James make a comparative study of topic models in information retrieval \([6]\).

In this paper, we present our comparative study of three topic models in topic detection using clustering method. We focus on the clustering result over the collection of web pages with different topics and documents instead of the feature space reduction. A distance measure matrix is derived from the topic models, which is then used to cluster the web pages into topic classes. F-measure is used to evaluate the clustering result using different topic models.

The rest of the paper is organized as follows. Section II gives the review of related research about topic models. Section III contains the definition of topic model distance and the evaluation method used in this comparative study. In Section IV we present the experiment result. Conclusions and future works are described in Section V.
II. TOPIC MODELS

Topic model is a probabilistic model for uncovering the underlying semantic structure of a document collection using different probabilistic generating procedures. There are two important topic models widely used in related research: the PLSA model and the LDA model. Both of the two models are based on the work of Latent Semantic Analysis (LSA) [7]. Some variations of these two models are proposed in related research.

In the following part of this paper, we use the following notations:

- \( D = \{d_1, d_2, \ldots, d_n\} \) is a collection of documents. \( n = |D| \) is the number of documents in \( D \).
- \( W \) is the word set of this documents collection.
- \( T \) is the set of \( k \) topics exists in this documents collection, and \( \{\theta_1, \ldots, \theta_k\} \) are the corresponding hidden topic models. \( K = |T| \) is the number of topics in \( T \).
- \( D_i \) is the documents categorized into topic \( \theta_i \).

A. Latent Semantic Analysis

LSA [7] assumes that the latent semantic can be captured from the linear space from the co-occurrence matrix derived from the collection of documents. Using the Singular Value Decomposition (SVD) algorithm, the co-occurrence matrix can be mapped into a latent semantic space, which is a lower dimension feature space. SVD is a classic linear algebra method widely used in the linear systems. Thus, the LSA is trying to capture the latent semantic in the linear space.

B. Probabilistic Latent Semantic Analysis

PLSA [8] is a latent variable model proposed by Hofmann in 1999, which has been called aspect model. PLSA tries to capture the latent semantic in the probabilistic space instead of the linear space. A hidden variable \( z \) is introduced to denote the latent semantic classes (topics). A topic is a distribution over words with different probabilities. A document can be treated as a sample of the mixture of \( k \) topics (see Fig. 1). In PLSA model, a document \( d_i \) may contain multiple \( k \) topics \( \{\theta_1, \ldots, \theta_k\} \). The probabilistic generating procedure in PLSA is as follows [8]:

- Select a mixture weight of the topics \( p(\theta_i | d_i) \)
- Generate a word \( w_j \) according to the probability of \( w_j \) in topic \( \theta_k \): \( p(w_j | \theta_k) \)

According to the above probabilistic generating procedure the probability of a word \( w_j \in W \) occurs in \( D \) can be defined as:

\[
p(w_j | d_i) = \sum_{k \in |T|} p(w_j | \theta_k)p(\theta_k | d_i). \tag{1}
\]

Here \( p(w_j | \theta_k) \) is the topic-specific word probability, which denotes the occurrence probability of word \( w_j \) in topic \( \theta_k \). \( p(\theta_k | d_i) \) is the mixture parameter of different topic \( \theta_k \). We denote this original PLSA model as: \( PLSA1 \).

Considering the noise word and non-topic specific words, Zhai et al. proposed an improved mixture model [9], which introduced a back ground distribution into the model: the background model \( \theta_B \). A mixture weight \( \lambda_B \) between background model \( \theta_B \) and other topic models \( \theta_i \) are used. Mei and Zhai proved its efficient in the discovering evolutionary theme patterns [10]. This model can be described by the following equation, which is denoted as \( PLSA2 \) in our research.

\[
p(w_j | d_i) = \lambda_B p(w_j | \theta_B) + (1 - \lambda_B) \sum_{\theta_j \in T} p(w_j | \theta_j)p(\theta_j | d_i). \tag{2}
\]

C. Latent Dirichlet Allocation

LDA [11]–[13] is originally introduced by Blei et al. in 2003. The PLSA model provides no probabilistic model at the level of documents [11]. LDA further considers a probabilistic model at this document level. Thus, it’s a generative three level hierarchical Bayesian probabilistic model for collections of discrete data such as text documents. LDA defines Dirichlet distribution on both \( P(w | z) \) and \( P(z | d) \). Usually, the graphic model is used to describe the LDA model in Fig. 2.

Let \( \vec{\alpha} \) be the a positive \( K \)-vector and \( Dir_{|W|}(\vec{\alpha}) \) denote a \( |W| \)-dimensional Dirichlet distribution with parameter \( \vec{\alpha} \). \( \vec{\eta} \) is a scalar and \( Dir_{|K|}(\vec{\eta}) \) denote a \( K \)-dimensional symmetric Dirichlet distribution with parameter \( \vec{\eta} \). Then, the probabilistic generating procedure of LDA is as follows:

- For each topic \( \theta_j \in T \), draw a distribution over words \( \vec{\eta}_j \) from Dirichlet distribution with parameter \( \vec{\beta} \): \( \vec{\eta}_j \sim Dir_{|W|}(\vec{\beta}) \).
- For a document \( d_i \in D \), draw a distribution over \( K = |T| \) topics \( \vec{\varphi}_i \) from Dirichlet distribution with parameter \( \vec{\alpha} \): \( \vec{\varphi}_i \sim Dir_{|T|}(\vec{\alpha}) \).
- For the hidden variable \( Z_{i,j} \), which is a topic indicator, draw a proportion of probability according to the topic assignment \( \vec{\varphi} \): \( Z_{i,j} \sim Multi(\vec{\varphi}_i) \).
- A word \( w_{i,j} \) in document \( d_i \) is generated from the topic denoted by the hidden variable \( Z_{i,j} \): \( w_{i,j} \sim Multi(\vec{\eta}_Z_{i,j}) \).

Following the above generating procedure, the joint distri-
bution of a document collection is given by:

\[
p(\varphi, \eta, Z, W | \alpha, \beta) = \prod_{j=1}^{|W|} p(z_{i,j} | \varphi)p(w_{i,j} | z_{i,j}, \eta)
\]

(3)

III. TOPIC MODEL DISTANCE AND EVALUATION

A. Topic Model Distance

Topic models can be evaluated by their intrinsic ability or extrinsic ability. The intrinsic ability of a topic model is the generalization capability which is independent of any other applications. Topic models are used in many applications, such as information retrieval, text classification, topic segmentation, etc. The Comparison of topic models in these applications is an extrinsic way of evaluation. In this paper, topic models are compared in the topic detection task using clustering method, which is an extrinsic way of evaluation.

Usually, the parameters of the topic models are estimated by expectation maximization (EM) method. As can be seen from section II, the topic models we used in this research all defines a mixture weight of documents to different topics \( P(\theta_k | d) \). A simple clustering method is clustering the documents using this mixture weight. As a matter of fact, we can further define a distance measure between documents using this mixture weight.

When the EM iteration converges, we will get the mixture weight of TM \( \theta_k \) over document \( d_i \) \( P(\theta_k | d_i) \). Then using the TM as \( K \) dimensional coordinates each document \( d_i \) can be represented as a \( K \) dimensional vector:

\[
\vec{V}(d_i) = (P(\theta_0 | d_i), \ldots, P(\theta_K | d_i)).
\]

(4)

The distance between two document \( d_i \) and \( d_j \) can be measured by this topic distance \( \vec{V}(d_i) \) and \( \vec{V}(d_j) \). Kullback-Leibler (KL) distance is widely used in the measure between probability density distributions. Thus, the topic model distance between document \( d_i \) and \( d_j \) can be calculated as the KL distance derived from \( \vec{V}(d_i) \) and \( \vec{V}(d_j) \):

\[
D(d_i, d_j) = \sum_{k=1}^{K} P(\theta_k | d_i) \ln \left( \frac{P(\theta_k | d_i)}{P(\theta_k | d_j)} \right).
\]

(5)

Using topic model distance, we can get the similarity matrix of a web news collection. The topic model induced distance considers the word distribution in different topics, which is supposed to be more discriminative in topic detection.

B. Base-line Method

To prove the effect of topic mode clustering for topic detection problem, the cosine distance is used as the baseline method.

The cosine distance used in this research is defined by the follow equations.

\[
cosine(d_i, d_j) = \frac{\sum_{w \in d_i \cap d_j} tf_{w,d_i} \cdot tf_{w,d_j} \cdot (idf_w)^2}{\sqrt{\sum_{w \in d_i} (tf_{w,d_i} \cdot idf_w)^2} \times \sqrt{\sum_{w \in d_j} (tf_{w,d_j} \cdot idf_w)^2}}.
\]

(6)

Usually, the maximum document frequency(DF) \( DF_{max} \) and the minimum document frequency \( DF_{min} \) is widely used in the VS model distance. Words with a high DF above \( DF_{max} \) or a low DF below \( DF_{min} \) will be dropped. From the view of topic representation, words occurred in too many topics contain few topical information. The minimum DF is not used since many topic specific words have a low DF. We use the \( DF_{max} \) for cosine distance , which is set to 0.35 according the comparison experiments using different \( DF_{max} \).

The DF is not used in topic model clustering. In PLSA2, the parameter \( \lambda_B \) is the mixture weight of back ground distribution \( \theta_B \) and topic distribution \( \theta_k \). This parameter is related with the document collections. In our collections, \( \lambda_B \) is set to 0.9.

C. Evaluation Method

There are three topic models compared in our research: the original PLSA model: \( PLSA1 \), the PLSA model with a background distribution: \( PLSA2 \), and the LDA model. The cosine distance is used as the base-line method.

Using the topic model distance defined in section III-A, we can get the corresponding distance matrix of each topic model. Then, K-center clustering method is used to clustering the collection of documents into classes according to the topic distance matrix. Since our research is a comparative study of topic models, the number of topics is assumed known to us. To eliminate the random factors of the comparison using K-center, the clustering using different distance shares the same random centers in the initiation of each K-center loop.

In order to compare the topic models in different document collections, we sampled document collections with different topic numbers and document numbers. The topic models are compared by the clustering result using topic models in different document collections.

In the following experiments, we use the balanced test set with topics randomly sampled from the corpus. In each randomly sampled topic, 30 web news pages are randomly selected.
IV. EXPERIMENTS AND RESULTS

A. Testing Corpus

The corpus used in experiments is a collection of web news pages totally 5033 documents from 30 topics. The web news are collected from the special news column of major portal sites in China using our web crawler pooler. News pages in the special news column are grouped by the reported topic which makes it easy to label the web pages with topics.

The web news are preprocessed by our Chinese word segmentation system ELUS [14] before they are used in the experiments.

B. Clustering Using the Topic Distance Matrix

In this comparative study, the three topic models are compared using the clustering result of the corresponding distance matrix. As referred in section III-C, there are two ways of using the converged parameters derived from topic models. Thus, we first device following experiment to compare the clustering result of using topic model distance and directly using the most salient \(P(\theta_i|d_i)\). The data set is collections of web pages with different topic numbers in \(\{6, 7, 8, 9, 10, 20\}\). The result can be seen from Table I.

The comparison of the web page collections with topic numbers in \(\{9, 20\}\) are plotted in bar graph. Fig. 3 is the experiment result on a collection of 9 topic totally 270 documents. Fig. 4 is the compare result on a collection of 20 topic totally 600 web pages.

From Table I, Fig. 4 and Fig. 3, we can see that, the clustering using the topic model distance (denoted as “Kmeans”) is better than directly using the most salient \(P(\theta_k|d_i)\) (denoted as “Direct” in bar graph) in all of the three topic models under different topic size. Thus, we will use the clustering on the topic model distance matrix instead of directly using \(P(\theta_k|d_i)\) in the following comparative study.

C. Comparison of Topic Model Clustering on Different Data Set

In order to compare the performance of using topic models for the task of topic clustering in different web page collections, we sampled collections with different numbers of topics and documents. Fig. 5 presents the comparison of three topic models: PLSA1, PLSA2 and LDA. The topic models are compared using the clustering result of corresponding topic model distance. The best results in 5000 runs of K-center on different web page collections are selected. F-measure is used to evaluate the clustering result.

Since the topics in a collection are randomly selected, some of the topic maybe related to each other in different level. Thus, a web collection with more topics and documents may be not hard to clustered into topic classes than the collection with less but related topics.

From Fig. 5 we can see that, topic clustering using topic model distance is more effective than cosine distance. The cosine distance suffers from the growth of the topic and documents. As the growing of number of topics and documents, the F-measure of cosine distance decreased. The topic model distance clustering, however, suffered a little from the growth of topics.

The complex model PLSA2 and LDA is better than the simple model PLSA1. This can be seen more clearly in the web page collections with more topics and documents. Thus, we can draw the conclusion that the complex topic models, which could capture more latent semantic features, are helpful in the topic clustering using topic model distance than the simple topic models. Both of the two topic models: PLSA2 and LDA are complex than the PLSA1 and both of them get a better topic classes than the PLSA1. In the web page collections with topic numbers in \(\{7, 9, 10, 20\}\), the F-measure of the two topic models is close to each other. Taking the average F-measure over all of the web page collections, the LDA is a little better than the PLSA2.
TABLE I
Comparison of Direct Clustering and Topic Model Distance Clustering

<table>
<thead>
<tr>
<th>Topic Num</th>
<th>Document</th>
<th>Cosine</th>
<th>PLSA1</th>
<th>PLSA2</th>
<th>LDA</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>Direct</td>
<td>Kcenter</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>180</td>
<td>0.686311</td>
<td>0.847569</td>
<td>0.856546</td>
<td>0.717610</td>
</tr>
<tr>
<td>7</td>
<td>219</td>
<td>0.618406</td>
<td>0.653663</td>
<td>0.704855</td>
<td>0.790091</td>
</tr>
<tr>
<td>8</td>
<td>240</td>
<td>0.655064</td>
<td>0.826424</td>
<td>0.855554</td>
<td>0.790091</td>
</tr>
<tr>
<td>9</td>
<td>270</td>
<td>0.695048</td>
<td>0.661855</td>
<td>0.729350</td>
<td>0.768294</td>
</tr>
<tr>
<td>10</td>
<td>300</td>
<td>0.7035353</td>
<td>0.762097</td>
<td>0.781480</td>
<td>0.854705</td>
</tr>
<tr>
<td>20</td>
<td>600</td>
<td>0.576831</td>
<td>0.796309</td>
<td>0.820734</td>
<td>0.846477</td>
</tr>
</tbody>
</table>

TABLE II
Comparison of Precision and Recall in Each Class (8 Topic)

<table>
<thead>
<tr>
<th>ID</th>
<th>Topic</th>
<th>Precision</th>
<th>Recall</th>
<th>F-measure</th>
<th>Precision</th>
<th>Recall</th>
<th>F-measure</th>
<th>Precision</th>
<th>Recall</th>
<th>F-measure</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Bali Blast</td>
<td>0.909601</td>
<td>1.00000</td>
<td>0.952381</td>
<td>0.560000</td>
<td>0.466667</td>
<td>0.509091</td>
<td>0.380000</td>
<td>0.633333</td>
<td>0.475000</td>
</tr>
<tr>
<td>2</td>
<td>Bush Visit Middle East</td>
<td>0.517857</td>
<td>0.966667</td>
<td>0.674419</td>
<td>0.631579</td>
<td>0.800000</td>
<td>0.705882</td>
<td>0.638889</td>
<td>0.766667</td>
<td>0.696970</td>
</tr>
<tr>
<td>3</td>
<td>Chinese Boat Kidnap</td>
<td>0.987742</td>
<td>1.00000</td>
<td>0.983607</td>
<td>0.687500</td>
<td>0.366667</td>
<td>0.478261</td>
<td>0.909091</td>
<td>1.00000</td>
<td>0.925381</td>
</tr>
<tr>
<td>4</td>
<td>Israel Attack Hamas</td>
<td>0.464929</td>
<td>0.500000</td>
<td>0.482790</td>
<td>0.476199</td>
<td>0.300000</td>
<td>0.392157</td>
<td>1.00000</td>
<td>0.153333</td>
<td>0.846154</td>
</tr>
<tr>
<td>5</td>
<td>Bird Flu</td>
<td>1.00000</td>
<td>1.00000</td>
<td>1.00000</td>
<td>1.00000</td>
<td>1.00000</td>
<td>1.00000</td>
<td>1.00000</td>
<td>1.00000</td>
<td>1.00000</td>
</tr>
<tr>
<td>6</td>
<td>Qu XinHua Killing Crime</td>
<td>1.00000</td>
<td>1.00000</td>
<td>1.00000</td>
<td>0.833333</td>
<td>0.833333</td>
<td>0.833333</td>
<td>1.00000</td>
<td>1.00000</td>
<td>1.00000</td>
</tr>
<tr>
<td>7</td>
<td>Unit Election in 2006</td>
<td>0.906250</td>
<td>0.966667</td>
<td>0.935484</td>
<td>0.827586</td>
<td>0.800000</td>
<td>0.813559</td>
<td>0.600000</td>
<td>1.00000</td>
<td>0.700000</td>
</tr>
<tr>
<td>8</td>
<td>Yao Ming Buy Shanghai Shark</td>
<td>1.00000</td>
<td>0.966667</td>
<td>0.983051</td>
<td>0.631579</td>
<td>0.800000</td>
<td>0.705882</td>
<td>1.00000</td>
<td>0.900000</td>
<td>0.947368</td>
</tr>
</tbody>
</table>

V. CONCLUSION

In this paper, a comparative study of the topic models for topic clustering is described. A collection of 5033 web news from 30 topics is gathered from the special news column of Chinese websites. The topical tags are labeled by hand. The topic distance derived from different topic models are compared using the cosine distance as the base-line method.

The topic models are compared using K-center clustering. To eliminate the random factors of the comparison, the clustering based on the two distance shares the same random centers in initiation of each K-center loop. The F-measure method is used to measure the clustering result.

The experiments show that, more topic features can be captured by a complex topic model. The topic model distance will generate more topic classes than the cosine distance taking the advantage of the topic features. Using a complex topic models, such as PLSA2 and LDA, we can get more topical classes than simple topic models.

ACKNOWLEDGMENT

This investigation was supported by the National Natural Science Foundation of China (No. 60703015 and No. 60973076).

REFERENCES


