Computers, Materials & Continua CMC, vol.64, no.1, pp.455-469, 2020

A Phrase Topic Model Based on Distributed Representation

Jialin Ma[1, *], Jieyi Cheng[1], Lin Zhang[1], Lei Zhou[1] and Bolun Chen[1, 2]

Abstract: Traditional topic models have been widely used for analyzing semantic topics from electronic documents. However, the obvious defects of topic words acquired by them are poor in readability and consistency. Only the domain experts are possible to guess their meaning. In fact, phrases are the main unit for people to express semantics. This paper presents a Distributed Representation-Phrase Latent Dirichlet Allocation (DR-Phrase LDA) which is a phrase topic model. Specifically, we reasonably enhance the semantic information of phrases via distributed representation in this model. The experimental results show the topics quality acquired by our model is more readable and consistent than other similar topic models.

Keywords: Phrase, topic model, LDA, distributed representation, Gibbs sampling.

1 Introduction

With the development of information technology, a large number of electronic documents have been accumulating in various fields, which result in information overload. In order to quickly search and find effective information, semantic topic analysis for these documents is one of the hotspots research at present [He (2016); Yu, Johnson and Kavuluru (2013)]. Probabilistic topic models can learn potential semantic information from documents. In recent years, topic models and related theories represented by LDA (Latent Dirichlet Allocation) and PLSA (Probabilistic Latent Semantic Analysis) are extensively studied and applied [Blei, Ng and Jordan (2003)]. However, language expression habits often take the form of multi-word phrase [Hofmann (1999); Xu (2019)]. The traditional topic models are based on 'Bag-of-Words' (BOW), and they model topics in the multinomial distribution of words. Therefore, their topic results have the following defects [Fei, Chen and Liu (2014)]:

(1) Poor readability: topics inferring from traditional topic models like PLSA, LDA, etc., are often difficult to understand and interpret by users. Only the domain experts are possible to guess their meaning. For example, in product reviews, "battery" and "life" are included in the same topic, but they have different probabilities values, which leads to the

[1] Jiangsu Internet of Things and Moblie Internet Technology Engineering Laboratory, Huaiyin Institute of Technology, Huai'an, 223003, China.

[2] University of Fribourg, Fribourg, 1700, Switzerland.

[*] Corresponding Author: Jialin Ma. Email: majl@hyit.edu.cn.

Received: 18 January 2020; Accepted: 30 March 2020.

two words are not close to each other. Therefore, it is hard for users to connect 'batteries life' in the mind. For another example, dismantling the phrase 'white house' in a topic will lose their original meaning, even greatly different.

(2) Prone to semantic association errors in application systems: in the above example, if the retrieval system sees ''white', may be related to 'house'. However, the user only wants to seek the 'while' which is about 'skirt'.

(3) Words co-occurrences were extra increased. Traditional topic models base on 'BOW'. Splitting phrases into independent words for topics analyzing will extra increase the words co-occurrence information in model training. After capturing the extra co-occurrence information of these words in the Gibbs sampling, the topic model will lead to poor quality of the learned topics, even adulterate words that do not belong to this topic.

Hence, there are many defects in the results of traditional topic models due to the BOW. It is easy to think of the way to construct the topic model by replacing BOW with the 'Bag-of-Phrases' (BOP). Many researchers have carried out the research work in this thinking. In summary, there are three kinds of methods to extract topic phrases from documents in related works.

The first one is devoted to building a generation model that combines phrases with topics in the early years. For example, Wallach proposes a Bigram Topic Model (BTM), which combines n-gram with hierarchical binary Dirichlet model [Wallach (2006)]. Their experiments on a small-scale document set showed that the proposed model is superior to the unigram model and the hierarchical binary model in terms of topic quality. Wang et al. [Wang, McCallum and Wei (2007)] propose a Topical N-Gram (TNG). The TNG model is based on BTM， and variable indicators were introduced to indicate whether the generated words are unigram or binary. Nevertheless, n-gram cannot share the same topic in TNG. With further research, Lindsey et al. [Lindsey, Headden and Stipicevic (2012)] propose a Phrase Discovering LDA (PDLDA). The location, length and topic of phrases are inferred at the same time in the PDLDA, and the Pitman-Yor process is used to relax the hypothesis of words bag. Jameel et al. [Jameel and Lam (2013)] also propose a topic model that can generate n-gram topics. Although these kinds of early research works can acquire topic phrases, the combination of phrase segmentation and topic analysis could lead to a sharp increase in model complexity. It is very difficult to deduce the parameters of the generating model with n-gram binary or more. Therefore, it is difficult to infer the parameters of this kind of methods, and the complexity of these models are very high, so they are difficult to apply to practice.

The second research strategy is to obtain topic words based on the unigram topic model, and then reconstructs topic phrases by these topic words. A representative study of this kind of method is Blei et al. [Blei and Lafferty (2009)]. They propose a visual topic model. This method can discover meaningful n-grams related to the topic, and help to understand the meaning of the topic. Danilevsky et al. [Danilevsky, Wang, Desai et al. (2014)] introduce a framework to generate topic key phrases and rank them. This method defines a function to sort the topic phrases to get more phrases that can represent the topic. Due to the limitation of the current natural language development technology, the quality of the topic phrases obtained by these methods, which compulsory combine the result words with the unigram topic model is often poor. These methods tend to produce

some incomprehensible and abnormal topic phrases.

For the kind of third method, some researchers devote to separating phrase segmentation from topic model in recent years. First, they get the phrases, then build topic model based on BOP. For example, El-Kishky et al. [El-Kishky, Song, Wang et al. (2014)] propose a phrase-mining framework called TopMine to generate single words or multi-word phrases of arbitrary length. They limited the component of phrase words to share the same topic in modeling. Li et al. [Li, Wang, Zhou et al. (2016)] propose a CITPM (Cluster-Based Iterative Topical Phrase Mining) framework to construct phrase topic model. The feature of this method is to cluster the corpus into multiple domain clusters, and carry out phrase mining and topic inference by iteration. The CITPM is better at finding phrases in special domain. They improve the accuracy of phrase topic mining.

The third method above can not only get more readable and consistent topic phrases, but also reduce the complexity of the model. Our method also belongs to this way. The closest research to our work is GPU (Generalized Pólya Urn) model that propose by Fei et al. [Fei, Chen and Liu (2014)]. They use a more Generalized Pólya Urn to increase the connection between phrases and its component words. This method improves the semantic contribution of phrases by promoting phrases and their component words directly in Gibbs sampling. They increase the probability of phrases by simply enhancing the count. It will lead to unrealistic over-enhancement problems. Chen et al. [Chen, Mukherjee, Hsu et al. (2013)] argue that a large amount of lexical relational knowledge exists in online dictionaries or other sources and can use to develop more consistent topic models. As a preliminary research, their work only focuses on the three relationships of synonyms, polysemy and adjectives in order to improve the quality of thematic models. Their work has proved that the idea is effective. Therefore, we combine some linguistic laws, and use distributed representation to measure the semantic relationship between the phrase and its component words in order to acquire the higher quality topic phrases.

This paper proposes a phrase topic model based on distributed representation. It named *DR-Phrase LDA*. Our model combines some linguistic laws, and uses distributed representation to measure the semantic relationship between the phrase and its component words. The *DR-Phrase LDA* promotes the semantic contribution of phrases by increasing the statistical information of phrases in Gibbs Sampling. The experimental results show the topics obtained by our model are more readable and consistent than other similar researches.

This paper is organized as follows: Section 2 introduces the principle and our *DR-Phrase LDA* in detail. Section 3 presents the experiments and discussion. Finally, we conclude and discuss further research in Section 4.

2 Our work

Using computer algorithms and models to analyze useful information from large-scale electronic document data has been becoming an important need of big data analysis. Topic models (such as PLDA, LDA, etc.) are important potential semantic analysis models, which can use to acquire semantic topics from a large number of electronic documents. They serve many advanced applications such as information retrieval, recommendation system, and knowledge map, etc. However, the obvious defects of

traditional topic models acquired topics are poor in readability and consistency, only the domain experts are possible to guess their meaning. In fact, phrases are the main unit for people to express semantics. Nonetheless, if we train the topic model for a large number of documents directly in terms of phrases will cause scarce phrase co-occurrence information. It will lead to a very small probability of phrases in topic. In our method, we combine some linguistic laws, and uses distributed representation to improve the co-occurrence information of phrases in the Gibbs sampling in order to acquire the higher quality topic phrases.

To illustrate the principles of our model, this section is beginning with briefly reviewing the Simple Pólya Urn (SPU) model, which is the basic theory of LDA [Blei, Ng and Jordan (2003)]; then we present framework and principles for our *DR-Phrase LDA*; finally, we give Gibbs sampling formula and algorithm.

2.1 SPU in LDA

As mentioned above, whether we can acquire effective topic phrases by directly training LDA on 'BOP'. The answer is obviously no. The reason is that phrase has fewer words frequency, so that a large number of unigrams sink its statistical information. It will lead to a very low probability of phrases in topics. The most fundamental reason is that the traditional LDA model follows the principle of famous SPU model. The SPU is a statistical model proposed by Hungarian mathematician Pólya to describe the dynamic change of probability. In LDA, every word of the corpus takes out at random from the BOW, and then puts it back into the bag. The next word repeats the above process again. There are only two colors of balls in the original SPU model. In the original SPU model, we suppose that an urn contains the number of N balls of two colors (black N_1 and white N_2, $N = N_1 + N_2$). Every time a ball is taken out of the urn at random, and wrote down its color; then puts the same color balls back into the urn with the number of R, and then repeat this process [Fei, Chen and Liu (2014)].

Traditional LDA topic model follows the principle of SPU model: "BOW" regard as an urn; words are similar to balls with different colors. The process of generation topics in LDA see as repeating to take out balls (words) from the urn. This process embodies in the Gibbs sampling process. Each topic assignment for each word is based on the probability formula: $P(w|t)=(N_{w|t}+\beta)/(N_t+V\beta)$. From this process, we can see taking out w every times will increase the probability of taking w in future. This self-reinforcing nature makes the rich richer [Fei, Chen and Liu (2014)].

2.2 DR-Phrase LDA

In order to get effective topic phrases, Fei et al. [Fei, Chen and Liu (2014)] use the more GPU model to increase the count of phrases and its component words in Gibbs sampling statistics uniformly. They consider that the semantics of a phrase directly relate to its constituent words. In their research, this method has received some success. However, there are obvious shortcomings of this method. In natural language, non-semantic combinatorial phrase, synonym and polysemy are very common in language. This method increases the probability of phrases and its component words by adding counts with simple literal information. It may be inconsistent with the original meaning of most

phrases and cannot get effective results of phrase topic. For example, you can be to increase the count of "white house", but we can't artificially increase the number of other location words "white" and "house" for the document.

In fact, there are many non-semantic combinations of phrases existing as word combination. Like most words, their specific meanings relate to the context. In order to analyze the meaning of phrases, we should not only consider the literal superposition of the component words of phrases, but also examine the meaning of phrases in the context. In recent years, the technology of Distributed Representation (or Word Embedding) based on neural network has made breakthrough progress. Through training corpus, Distributed Representation can map words into N-dimensional space, which makes the semantics of words computable. We use Distributed Representation technology to examine the context of phrases. It can use to measure the relationship between the overall semantics of phrases and the context semantically. For acquiring more comprehensive topic model based on BOP, we improve the position of phrases in the process of Gibbs sampling according to the context. The proposed model named *DR-Phrase LDA*. The specific framework flow of *DR-Phrase LDA* shows in Fig. 1.

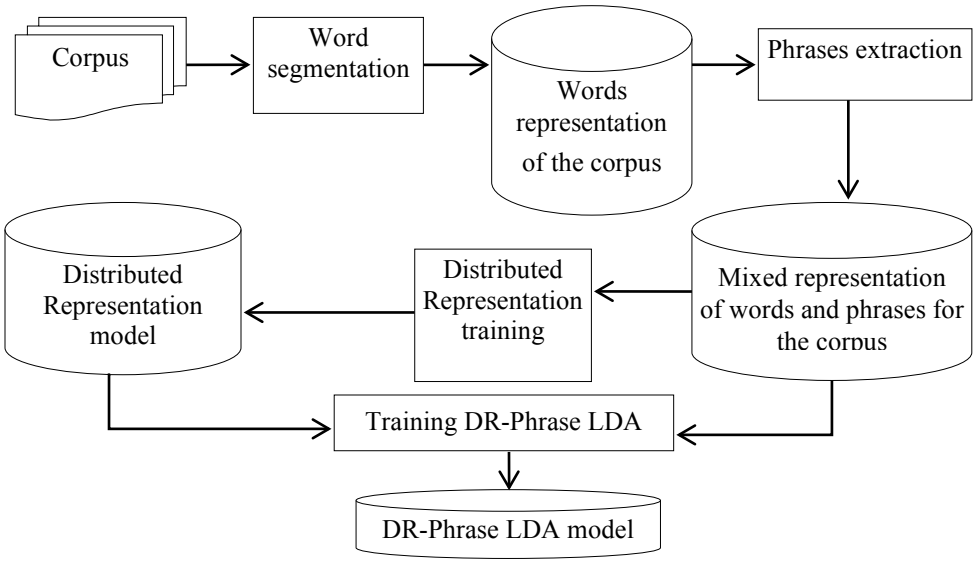

Figure 1: The whole framework flow of the *DR-Phrase LDA*

In the process of phrases extraction, not all words are phrase. Therefore, the next step of the corpus is a mixed representation of words and phrases. For the sake of unified expression, phrases and words all call as term. The *DR-Phrase LDA* focuses on the treatment of low frequency phrases in Gibbs sampling. According to the principle of Pólya Urn model, the counts of phrases increase during the sampling process combining the linguistic characteristics. We mainly consider the following linguistic features for phrases:

(a) The semantic intensity of phrases relate to their length. In general, long phrase expresses stronger semantics.

(b) The overall semantics of a phrase relate to its context. It is not sure the component words of phrase relate to the meaning of its phrase. We should examine the meaning of phrases in its context.

(c) The notional word occupies the main position in the semantic expression. According to the grammatical function and nature, words can divide into two categories: notional words and functional words. Functional words generally do not contain real meanings, and their basic purpose is to express grammatical relations. Functional words mainly refer to adverb, conjunction, preposition, auxiliary, interjection and onomatopoeia. The notional words have real meaning. They mainly refer to noun, verb, adjective, numeral, quantifier, and so on. Notional words are the core of semantic expression. They should play a dominant role in semantic analysis.

According to the above three linguistic features and combined with the principle of Pólya Urn model, our *DR-Phrase LDA* reforms the Gibbs sampling of the traditional LDA by designing appropriate strategies. Specifically, we adopt strategies as the following:

● **Phrase to word:** When the iteration process reaches a term in Gibbs sampling, and if it is a phrase, our *DR-Phrase LDA* increase the count of the phrase and its semantic relate words together. The *DR-Phrase LDA* is to select the first top γ term with semantic similarity to w^p as the context. We obtain them via calculating Distributed Representation trained in the corpus. When the phrase w^p draws from the topic k, the count of topic k about w^p can calculat by the following formula:

$$C(w^p)=\mu len(w^p) \tag{1}$$

where μ>=1, and μ is the adjustment parameter. $\boldsymbol{len}(w^p)$ is the length of the phrase w^p. In the meantime, the count of semantic related words of the phrase w^p in the topic k are increased by the following formula:

$$C(w_i)=\begin{cases} 0 & If\ w_i \text{ is the notional words.} \\ Int(1+(C(w^p)-1)\times Sim(\mathbf{w_i},\mathbf{w^p})) & If\ w_i \text{ is the function words.} \end{cases} \tag{2}$$

In formula (2), the Distributed Representation of w_i is $\mathbf{w_i}$ =($w_{i,0}$, $w_{i,1}$, ……, $w_{i,n}$), and the Distributed Representation of phrase w^p is $\mathbf{w^p}$=($w^p{}_0$, $w^p{}_1$....., $w^p{}_n$); Int() means to take an integer; the $Sim(\mathbf{w_i}\ \mathbf{w^p})$ means Cosine Similarity of $\mathbf{w_i}$ and $\mathbf{w^p}$, which can be calculated by:

$$Sim(\boldsymbol{w_i},\boldsymbol{w^p})=\frac{\sum_0^n(w_i\times w_i^p)}{\sqrt{\sum_0^n w_i^2}\times\sqrt{\sum_0^n w^{b2}_i}} \tag{3}$$

●**Word to Phrase:** When the iteration process reaches a term in Gibbs sampling, if it is a word and has the semantic related phrases, our *DR-Phrase LDA* increases the count of the word w and its top γ semantic related phrases together. When the word w draws for from topic k, the count of topic k about w can calculate by the following formula:

$$C(w)=\begin{cases} Int(1+\sum_{j=0}^{m}Sim(\boldsymbol{w},\boldsymbol{w_j^p})) & If\ w \text{ is a notional word and the context of } w_j^p. \\ 1 & \text{Others} \end{cases} \tag{4}$$

In formula (4), w has the number of m semantic related phrases. Meanwhile, the count of these phrases also should increase by:

$$C(w_j^p)=1+\mu len(w^p)*Sim(\boldsymbol{w}, \boldsymbol{w}_j^p) \quad (5)$$

The formula (1) and (5) are all to enhance the count of phrase. Nevertheless, we consider that the enhancing caused by phrases themselves should be different from that caused by context words. Therefore, the former contributes more directly for topic. We make sure the increased count of phrases relate with the semantic relativity between word and the phrases. The semantic relativity can calculate by the formula (3) (Cosine Similarity).

2.3 Gibbs sampling formula and algorithm

The new strategies for SPU model demonstrates in the Section 2.2 in detail. The *DR-Phrase LDA* not only can make to enhance the probability of 'ball' which is taken again, but also can enhance the probability of its related 'balls' that are taken again. The formula of Gibbs sampling for our *DR-Phrase LDA* can calculate by:

$$P(z=k/\boldsymbol{z}_{\neg}, \boldsymbol{w}, \boldsymbol{\alpha}, \boldsymbol{\beta}) \propto \frac{n_{k/d}+\alpha}{\sum_{i=1}^{K} n_{i/d}+K\boldsymbol{\alpha}} \times \frac{n_{t/k}+\sum_{r=1}^{n_r} n_{t_r/k}+\beta}{\sum_{j=1}^{N_t} n_{t_j/k}+N_t\boldsymbol{\beta}} \quad (6)$$

where t is the current instance of Gibbs sampling. k denotes a topic number; K is the total number of topics; N_t is the total count of terms in the whole corpus; $n_{k/d}$ denotes the count of topic k in document d; $n_{t/k}$ denotes the count of t in topic k; n_r denotes the total number of term which is related with t of semantic. a and β are the hyper-parameters of Dirichlet. Our *DR-Phrase LDA* improves traditional LDA by the different strategies in Gibbs sampling. These strategies are embodied in the formulas of (1)-(5), and the actual process of Gibbs sampling reflects in the formula (7). In order to express simply, the four count promoting strategies in the *DR-Phrase LDA* are simplistically reflects in the latter part of the formula (7). For program designing, the count promoting strategies should realize according to the specific situation. The algorithm of Collapsed Gibbs Sampling for the *DR-Phrase LDA* is as the following:

Algorithm: The Collapsed Gibbs Sampling for the DR- Phrase LDA.

Input: Mixed representation of words and phrases for the corpus; Distributed Representation training on the corpus; parameters: α, β, K, IterNum (iterations), γ, μ.

Output: $\boldsymbol{z}$ (two-dimensional matrix $\boldsymbol{z}$: the rows of $\boldsymbol{z}$ are all documents in the corpus; the columns of $\boldsymbol{z}$ are terms in each document; the elements of $\boldsymbol{z}$ are the number of topics that assign in Gibbs sampling equilibrium.)

Algorithm processing:

Initializing: assigning topic numbers from 0 to K-1for all terms in the documents at random.

For inter=0 to IterNum // IterNum: total iterations

{

For each d in documents // d is a document.

```
  {
    For each t in d  // t is a term in d.
    {
      If t is a phrase Then
       {
          Increase the count of topic k about t by formula (1).
          For each t_r in R  // R is the top γ semantic related word set for t.
            Increase the count of topic k about t_r by formula (2).
       }
      ElseIf t is notional word and it is existing semantically related phrases for t, then
                    Increase the count of topic k about t by formula (3),
                    and increase other statistics accordingly.
                    Increase the count of topic k about semantic related phrases of t by formula (4).
                    and increase other statistics accordingly.

    Else
      t is the other, then the statistics are reduced by 1 according to the normal way.
    P[ ]=double P[K];  // store the probability values of t belong to topic k.
    For i=0 to K-1
      P[i]= Calculating the probability value of t belong to topic k.
    New_k_t=Cumulative (P) // *The New topic number New_k_t of t is obtained by Cumulative
                                  Method. *//
    z[d][t]= New_k_t;   // Update the topic number of t.
     }
   }
}
```

3 Experiment

In order to evaluate the proposed *DR-Phrase LDA* model, we compare it with the three baselines: 1) standard LDA that base on 'BOW' (here in after referred to as LDA). 2) LDA that base on 'BOP' directly (here in after referred to as Phrase-LDA). 3) The GPU proposes by Fei et al. [Fei, Chen and Liu (2014)] (here in after referred to as GPU-LDA).

3.1 Experimental corpus

We used user comments from different sources as experimental corpus for our evaluation. The first part is OpinRank Review-Dataset, which published by Ganesan et al. [Ganesan and Zhai (2012)]. Their data set contains users reviews for cars and hotels collected from Tripadvisor (259,000 reviews) and Edmunds (42,230 reviews). Besides, we also used product reviews from 10 sub-categories (types of product) reviews from jd.com, including Cellphone, MacBook, Camera, Mouse, Keyboard, Printer, Displayer,

Television, Clothing and Air conditioning. Each of them has 5,000 reviews.

3.2 Evaluation methods

Perplexity and Kullback Leibler (KL) distance are the statistical methods use to evaluate the performance of topic models commonly. However, many researchers had found that they often acquired differences results between the complexity evaluation methods and the manual judgment when evaluate the prediction topics of documents [Fei, Chen and Liu (2014); Mimno, Wallach, Talley et al. (2011)]. A new method to evaluate the coherence of topics proposes in above work. This method is agreement with the quality of the manual subject identification. The higher score in this method means the better quality of topics. The specific calculation formula is as follows:

$$C(t;V^{(t)}) = \sum_{m=2}^{M}\sum_{l=1}^{m-1}\log\frac{D(v_m^{(t)},v_l^{(t)})+1}{D(v_l^{(t)})} \quad (7)$$

In the formula (7), $D(v)$ denotes the frequency of documents in which the word v has appeared. D (v, v') denotes the count of documents in which the word v and v' co-occurrence. M denotes the first M words with the highest probability of topic t. $v^{(t)}$ represents the top M words with the highest probability value in topic t. '1' is a smooth count in order to avoid a calculation result of zero. The similar research works [Fei, Chen, and Liu (2014); Mimno, Wallach, Talley et al. (2011)] to us also used this method to evaluate the consistency of their topic model. In order to avoid subjective factors, our experiments also use this objective method to evaluate the *DR-Phrase LDA* with other similar topic models.

3.3 Experimental preparation

(1) Phrases extraction

According to relevant linguistic theories, people use a large number of finished phrases or semi-finished phrases constantly in the process of language communication. The collocation of these phrases consider as a large number of rules stored in the human brain [Blei and Lafferty (2007)]. The task of phrase automatic extraction has become one of the hot issues in the field of natural language research. Many scholars have proposed various techniques for phrase extraction, but phrase extraction is not the focus of our work. In order to reduce the workload, we used the common method based on word frequency co-occurrence to extract phrases in our experiments. The specific method is to calculate the score of binary phrase candidate set *score* (w_i, w_j), and select the first m with high score to form a formal binary phrases, then add them to the phrases set. The *score* (w_i, w_j) of bigram phrase is calculated by the following formula [Mikolov, Sutskever, Chen et al. (2013)]:

$$score(w_i,w_j) = \frac{count(w_i,w_j)-\eta}{count(w_i)*count(w_j)} \quad (8)$$

In the formula (8), w_i and w_j are binary phrase word pairs captured from corpus. *Count* (w_i, w_j) is the co-occurrence frequency of w_i and w_j. *count* (w_i) is the single word frequency of w_i, and *count* (w_j) is the word frequency of w_j. Parameter η use to filter low-

frequency phrases. For example, η=5 can filter out word pairs whose co-occurrence frequency is less than five.

(2) Distributed representation

As early as 1986, Hinton propose a new concept of distributed representation, which maps each word into a vector of a specified length through corpus training, and the words represent in space. It can make the words have computable [Rumelhart, Hinton and Williams (1988)]. The Google team represented by Tomas mikolov released the famous open source package word2vec in 2013. We use word2vec to measure the semantic relationship between the phrase and its component words for our *DR-Phrase LDA* in experiments.

3.4 Results and analysis

The purpose of our research is to acquire more readable topic phrases. Parameters optimization is not the focus for us. Therefore, we set α=50/K and β=0.01 according to the previous research experiences [Heinrich (2005)]. In order to find the suitable topic parameter K for the experimental corpus, we used perplexity to detect the appropriate topic parameter K. In our experiments, we randomly selected 90% of reviews for training, and the rest for testing. We set a series of K=20, 25… 50 to train Phrase-LDA, GPU-LDA and *DR- Phrase LDA*, then calculated perplexity of test samples. The iteration was set to 1,000. The experimental results show in Fig. 2.

It can be seen from Fig. 2 that the perplexities of the four models gradually decrease from k=30, and the perplexities all tends to be relatively stable at about K=30. Therefore, we set K=30 for the participating evaluation models. In addition, the results also reflect that the perplexity of phrase LDA is the highest. It is far higher than the other three. The main reason is that it directly trained model on phrases, and did not take enhanced sampling measures, resulting in little or no co-occurrence information of phrases. In contrast, the GPU-LDA and our *DR-Phrase LDA* all take measures to enhance the information of phrases and their component words, so the result of perplexities are lower than the LDA. In order to save efficiency and computation, we only used the above methods to extract

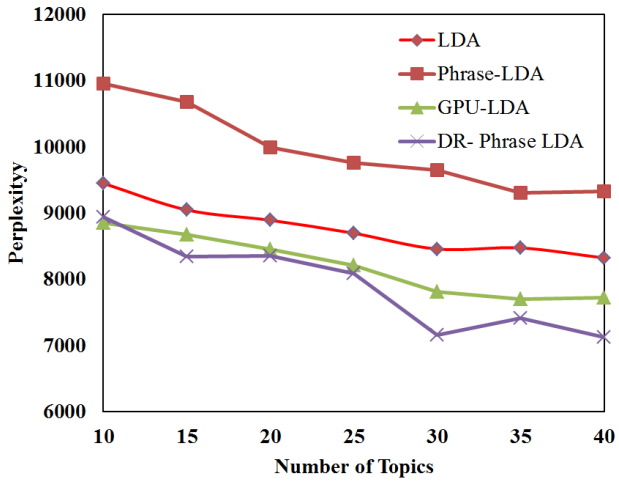

Figure 2: The perplexities for different K

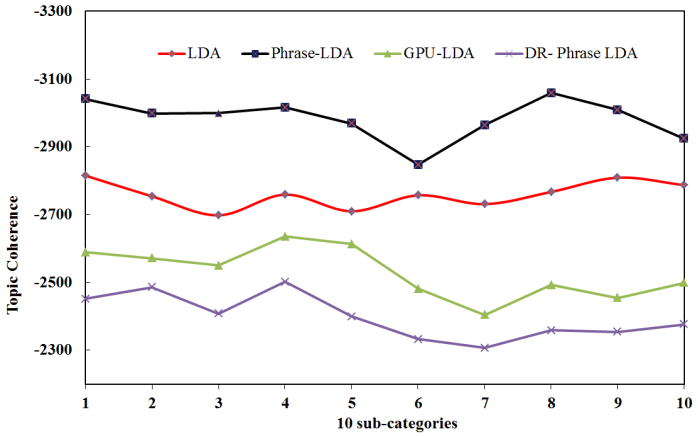

Figure 3: Topics coherence for the four models in 10 sub-categories product reviews

the commonly used binary and ternary phrases in experiments for the Phrase-LDA, GPU-LDA, and the *DR-Phrase LDA*. In addition, we used word2vec package of gensim to train distributed representation for the *DR-Phrase LDA*, and set parameters *min_count*=5, dimension *size*=50. Other parameters are default.

Fig. 3 shows the topic coherence of LDA, Phrase LDA, GPU LDA and DR-Phrase LDA according to formula (8). It can see that the topic coherence of phrase LDA is the worst, and the GPU-LDA and *DR-Phrase LDA* is obviously better than the LDA and Phrase-LDA. Although the Phrase-LDA uses 'POW' to replace 'BOW', in the traditional topic model theory, when the whole phrase assigned the same topic for co-occurrence capture, the co-occurrence information of phrase will decline. Therefore, in the phrase LDA, the co-occurrence information of phrases is far lower than single word, which results in a very low probability of many phrases in topics. This lead to the Phrase LDA is worse. On the contrary, the GPU-LDA and *DR-Phrase LDA* take into account the strong semantic expression characteristics of phrases and deliberately. This improves the statistic of phrases and their components words in sampling. The Fig. 3 also reflects the promoting of coherence. Compared with the GPU-LDA, our *DR-Phrase LDA* considers the different parts of speech and the semantics about components words in the phrase. In particular, the component words and the meanings of many phrases are very different from the whole semantics of phrases, which regard as no difference in the GPU-LDA. Therefore, the Fig. 3 also reflects the coherence of our model is better than the GPU-LDA.

Tab. 1 shows the top five topic terms of four sub-categories (Cellphone, MacBook, Television, and Air conditioning) for the four models. We can see the top five terms of LDA and phrase LDA models are single words, the GPU-LDA model has both phrases and words, but its top terms is more words. On the contrary, the *DR-Phrase LDA* has more phrases than words in top five. In fact, the LDA is a complete topic model of BOW. Although the Phrase LDA is based the BOP, due to the same statistical status of phrases and general words, a large number of phrase co-occurrence information is less, so it is

Table 1: Top 5 topic terms of 4 sub-categories for the four models

Cellphone				Air conditioning			
LDA	Phrase -LDA	GPU-LDA	DR- Phrase LDA	LDA	Phrase -LDA	GPU-LDA	DR- Phrase LDA
Huawei	iphone	Huawei	good	Gree	Midea	good	good service
Miui	Huawei	iphone	big screen	Midea	Gree	service	like
Sumsung	Oppo	screen	Feel good	Haier	cooling	compressor	looks beautiful
screen	speed	play game	Huawei	appearance	fast	quietness	service
fluency	photo	Big screen	screen	quality	Haier	Cooling fast	Cooling fast
Television				**MacBook**			
LDA	Phrase -LDA	GPU-LDA	DR- Phrase LDA	LDA	Phrase -LDA	GPU-LDA	DR- Phrase LDA
sony	sharp	big	good service	Apple	like	Apple	Apple
Hisense	sony	good	pure color	performance	Apple	good	good performance
TCL	Hisene	TV	Large size	like	good	battery	long life battery
screen	big	beautiful	service	ultrathin	weight	Screen	battery
good	screen	good service	color	good	screen	life battery	screen

difficult to obtain a higher probability, and fails to reflect the role of strong semantic blocks of phrases. The GPU-LDA not only improves the statistics of phrases, but also improves the statistics of phrase component words. However, the semantics of many phrases do not necessarily relate to their component words in natural language. Therefore, we can see the phrase is lower than its component words in top five terms for the GPU-LDA. For example, in the results of 'Cellphone', the probability of 'Big screen' is lower than 'screen'. In the results of our *DR-Phrase LDA* model, the overall meaning of phrase is stronger than its component words. The probability of 'Big screen' is higher than 'screen'. This is more accord with the law of nature language.

In addition, we recorded the training time of the four models. The experimental data scale and parameter setting are the same as above. For the sake of comparison, we exclude the training time of word2vec for our *DR-Phrase LDA*. The experiments ran on a computer with 16 GB memory and eight cores (Intel i7-9700F). The operating system version is windows 10. The model programs completed with Python 3.5.

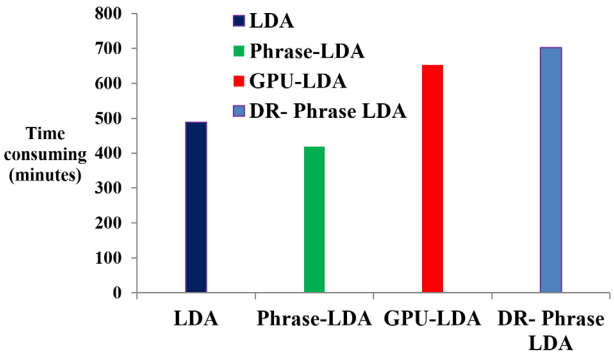

Figure 4: Time consumption comparison of the four models

Fig. 4 shows the models runtime of each method on experimental datasets. Obviously, the LDA and Phrase-LDA take less time than the GPU-LDA and the *DR-Phrase LDA*. The basic reason is that models LDA and Phrase-LDA treat phrases like other common words without any special treatment. It leads to a very low probability of phrases in topics. As a result, they cannot train effective topic phrases, although they spend less time. The GPU-LDA and our *DR-Phrase LDA* need to spend more time dealing with phrases in training according to their own strategies. Compared with the GPU-LDA, the strategy of our *DR-Phrase LDA* is more comprehensive and reasonable, so it takes more time to train. In addition, we did not consider optimization for algorithm program, which is also an important reason for long time consumption.

4 Conclusions and future work

For the shortcomings of traditional topic model, such as poor readability, consistency and visualization, this paper proposes a phrase topic model *DR-Phrase LDA*, which base on distributed representation. We consider the different semantic relations of the whole phrase and its component words in Gibbs sampling in the model. The similar researches works do not pay attention to this. The experimental results show that the topics readability and coherence are higher than the traditional LDA and LDA based on BOP. Besides, the topic quality of the proposed *DR-Phrase LDA* is also further improved than the similar model GUP-LDA.

In addition, we do not consider the time efficiency of algorithm program. This leads to the model training time is longer. In order to improve the efficiency and meet the application requirements, the future work plans to optimize the algorithm.

Funding Statement: This work was supported by the Project of Industry and University Cooperative Research of Jiangsu Province, China (No. BY2019051). Ma, J. would like to thank the Jiangsu Eazytec Information Technology Company (www.eazytec.com) for their financial support.

Conflicts of Interest: The authors declare that they have no conflicts of interest to report regarding the present study.

References

Blei, D. M.; Ng, A. Y.; Jordan, M. I. (2003): Latent Dirichlet allocation. *Journal of Machine Learning Research*, vol. 3, no. 1, pp. 993-1022.

Blei, D. M.; Lafferty, J. D. (2009): Visualizing topics with multi-word expressions. arXiv preprint arXiv: 0907.1013.

Blei, D. M.; Lafferty, J. D. (2007): A correlated topic model of science. *The Annals of Applied Statistics*, vol. 1, no. 1, pp. 17-35.

Chen, Z.; Mukherjee, A.; Liu, B.; Hsu, M.; Castellanos, M. et al. (2013): Discovering coherent topics using general knowledge. *Proceedings of the 22nd ACM International Conference on Information & Knowledge Management*, pp. 209-218.

Danilevsky, M.; Wang, C.; Desai, N.; Ren, X.; Guo, J. et al. (2014): Automatic construction and ranking of topical key phrases on collections of short documents. *Proceedings of the SIAM International Conference on Data Mining. Society for Industrial and Applied Mathematics*, pp. 398-406.

El-Kishky, A.; Song, Y.; Wang, C.; Voss, C. R.; Han, J. (2014): Scalable topical phrase mining from text corpora. *Proceedings of the VLDB Endowment*, vol. 8, no. 2, pp. 305-316.

Fei, G.; Chen, Z.; Liu, B. (2014): Review topic discovery with phrases using the Pólya urn model. *Proceedings of COLING, 25th International Conference on Computational Linguistics: Technical Papers*, pp. 667-676.

Ganesan, K.; Zhai, C. (2012): Opinion-based entity ranking. *Information Retrieval*, vol. 15, no. 2, pp. 116-150.

He, Y. (2016): Extracting topical phrases from clinical documents. *Thirtieth AAAI Conference on Artificial Intelligence*.

Heinrich, G. (2005): Parameter estimation for text analysis. *Technical Report*.

Hofmann, T. (1999): Probabilistic latent semantic indexing. *Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 50-57.

Jameel, S.; Lam, W. (2013): An unsupervised topic segmentation model incorporating word order. *Proceedings of the 36th International ACM SIGIR Conference on Research and Development in Information Retrieval ACM*, pp. 203-212.

Lindsey, R. V.; Headden III, W. P.; Stipicevic, M. J. (2012): a phrase-discovering topic model using hierarchical Pitman-Yor processes. *Proceedings of the Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning Association for Computational Linguistics*, pp. 214-222.

Li, B.; Wang, B.; Zhou, R.; Yang, X.; Liu, C. (2016): CITPM: A cluster-based iterative topical phrase mining framework. *International Conference on Database Systems for Advanced Applications Springer, Cham*, pp. 197-213.

Mimno, D.; Wallach, H. M.; Talley, E.; Leenders, M.; McCallum, A. (2011): Optimizing semantic coherence in topic models. *Proceedings of the Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics*, pp. 262-272.

Mikolov, T.; Sutskever, I.; Chen, K.; Corrado, G. S.; Dean, J. (2013): Distributed representations of words and phrases and their compositionality. *Advances in Neural Information Processing Systems*, pp. 3111-3119.

Rumelhart, D. E.; Hinton, G. E.; Williams, R. J. (1988): Learning representations by back-propagating errors. *Cognitive Modeling*, vol. 5, no. 3, pp. 1.

Wallach, H. M. (2006): Topic modeling: beyond bag-of-words. *Proceedings of the 23rd International Conference on Machine Learning ACM*, pp. 977-984.

Wang, X.; McCallum, A.; Wei, X. (2007): Topical n-grams: phrase and topic discovery, with an application to information retrieval. *Seventh IEEE International Conference on Data Mining*, pp. 697-702.

Xu, F.; Zhang, X.; Xin, Z.; Yang, A. (2019): Investigation on the Chinese text sentiment analysis based on convolutional neural networks in deep learning. *Computers, Materials & Continua*, vol. 58, no. 3, pp. 697-709.

Yu, Z.; Johnson, T. R.; Kavuluru, R. (2013): Phrase based topic modeling for semantic information processing in biomedicine. *12th International Conference on Machine Learning and Applications IEEE*, vol. 1, pp. 440-445.