Neurocomputing xxx (xxxx) xxx

Contents lists available at ScienceDirect

Neurocomputing

journal homepage: www.elsevier.com/locate/neucom

Optimising topic coherence with Weighted Pólya Urn scheme

Rui Wang[a], Deyu Zhou[a,*], Yulan He[b]

[a] School of Computer Science and Engineering, Key Laboratory of Computer Network and Information Integration, Ministry of Education, Southeast University, China
[b] Department of Computer Science, University of Warwick, Coventry CV4 7AL, UK

ARTICLE INFO

Article history:
Received 6 November 2018
Revised 23 July 2019
Accepted 3 December 2019
Available online xxx

Communicated by Dr. Jing Jiang

Keywords:
Pólya urn scheme
Unsupervised learning
Topic model
Sentiment analysis

ABSTRACT

Topic models have been widely used to mine hidden topics from documents. However, one limitation of such topic models is that they are prone to generate incoherent topics. To address this limitation, many approaches have been proposed to incorporate the prior knowledge of word semantic relatedness into the topic inference process. One example is the Generalized Pólya Urn (GPU) scheme. However, GPU-based topic models often require sophisticated algorithms to acquire domain-specific knowledge from data. Moreover, prior knowledge is incorporated into the topic inference process without considering its impact on the intermediate topic sampling results. In this paper, we propose a novel Weighted Pólya Urn scheme and incorporate it into Latent Dirichlet Allocation framework to build the self-enhancement topic model and generate coherent topics. In specific, semantic prior knowledge based on word embedding is employed to measure the semantic coherence of a word to different topics, which is incorporated into the Weighted Pólya Urn scheme. Moreover, semantic coherence is updated dynamically based on the semantic similarity between a word and the representative words in different topics. Experiments have been conducted on seven public corpora from different domains to evaluate the effectiveness of the proposed approach. Experimental results show that compared to the state-of-the-art baselines, the proposed approach can generate more coherent topics.

1. Introduction

Topic models, such as Latent Dirichlet Allocation (LDA) [1], have been powerful approaches for unsupervised topic extraction, particularly within the domain of text processing. Many variants of LDA have been proposed in the natural language processing community to tackle different problems. For example, Lin and He [2] proposed a Joint Sentiment-Topic Model for topic-associated sentiment analysis. Yan et al. [3] proposed a biterm topic model to deal with the data sparsity in short text. Yin and Wang [4] employed a Dirichlet Multinomial Mixture model for short text clustering. However, such unsupervised topic models are prone to generate incoherent topics without using any prior knowledge of word semantic relatedness.

To address this problem, several approaches have been proposed to incorporate the domain knowledge mined from external corpora or supplied by users into the topic inference process. For example, the Dirichlet Forest-LDA model [5] incorporated the knowledge expressed as constraints of must-links and cannot-links. A must-link states that two words should be assigned to the same topic, while a cannot-link states that two words should not appear in the same topic. Chen et al. [6] employed the domain knowledge mined from an external corpus using the Generalized Pólya Urn (GPU) scheme for the extraction of product aspects. In a similar vein, Chen and Liu [7] proposed a lifelong learning framework based on GPU for product aspect mining.

However, GPU-based topic models often require sophisticated algorithms to acquire domain-specific knowledge from data. Moreover, prior knowledge of word semantic relatedness is incorporated into the topic inference process without considering its impact on the intermediate topic sampling results. For example, as illustrated in Fig. 1(b), if the word *'picture'* is sampled from the *'machine learning'* urn (topic), other similar words such as *'image'* and *'vision'* might also be added into the *'machine learning'* urn. However, most existing words in the urn are actually not semantically related to the word *'picture'*. As such, blindly incorporating prior knowledge of word semantic relatedness without considering its impact on the existing topic sampling results may deteriorate the quality (topic coherence) of the *'machine learning'* topic.

To overcome the two weaknesses of GPU, in this paper, we propose a novel self-enhancement topic modeling approach based on the Weighted Pólya Urn scheme. In particular, a novel

* Corresponding author.
E-mail addresses: rui_wang@seu.edu.cn (R. Wang), d.zhou@seu.edu.cn (D. Zhou), Yulan.He@warwick.ac.uk (Y. He).

https://doi.org/10.1016/j.neucom.2019.12.013
0925-2312/

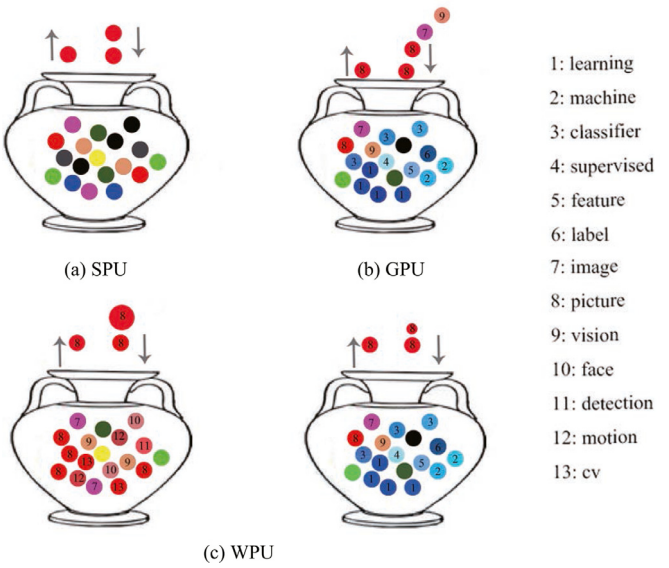


Fig. 1. Comparison of SPU, GPU and WPU. Here, SPU means the Simple Pólya Urn scheme. In WPU, balls in red dominate the left urn (*'computer vision'* topic) and balls in blue dominate the right urn (*'machine learning'* topic). The balls with different numbers represent different words. For example, the ball with digit '8' denotes the word *'picture'*.

measure, semantic coherence, is proposed to describe the semantic similarity between a word and a topic based on their respective word embedding. For example, as illustrated in Fig. 1(c), the semantic coherence between the word *'picture'* and the topic *'computer vision'* (left urn) is higher than that with the topic *'machine learning'*. Such semantic coherence measure is employed in the proposed Weighted Pólya Urn scheme (WPU). Moreover, the semantic coherence value is updated dynamically based on the intermediate topic sampling results during the self-enhancement phase of topic modeling.

The contributions of this paper are summarized as follows.

- To the best of our knowledge, it is the first attempt to employ the semantic coherence measures of words with different topics during the topic inference process. A novel Weighted Pólya Urn scheme is proposed to incorporate such semantic coherence.
- A novel self-enhancement topic modeling approach based on Weighted Pólya Urn scheme is proposed, in which intermediate topic sampling results in each Gibbs sampling iteration are employed to update the semantic coherence values and guide the subsequent sampling iteration.
- Experimental results on seven public corpora show that the proposed approach outperforms the state-of-the-art baselines and generates more coherent topics.

The remainder of the paper is organized as follows: Section 2 briefly introduces the related work; a novel Weighted Pólya Urn (WPU) scheme is dicussed in Section; In Section 4, a WPU-based sampler is incorporated into topic modeling to propose the Self-Enhancement Topic Model (SE-TM); Experiments and result analysis are disscussed in Section 5. Finally, we will conclude the paper in Section 6.

2. Related work

Our work is related to three lines of research, word representation learning, the Simple Pólya Urn (SPU) scheme and the Generalized Pólya Urn (GPU) scheme.

2.1. Word representation learning

Distributional semantic models (i.e. word embeddings) have recently been applied successfully in many NLP tasks [8].

Neural network based approaches have become more efficient, allowing their use in multiple scenarios, thanks to the *skip-gram with negative-sampling training method* (SGNS) [9]. It was widely popularized via *word2vec*, a software to create word embedddings. To model the entailment and the asymmetric relationships between embeddings, Vilnis and McCallum [10] used a multivariate gaussian density to represent a word. Recently, a new word embedding method has been proposed, called *fastText* [11] which treats each word as made of character n-grams. Vector representations are then computed from the sum of their n-gram representations. Furthermore, Athiwaratkun et al. [12] developed a *probabilistic fastText* that can capture multiple word senses, sub-word structure and uncertainty information. More traditional vector representations based on a dimensionality reduction obtained by applying the Singular Value Decomposition (SVD) to the weighted document-term matrix of the corpus; Latent Semantic Analysis (LSA) [13] is a prominent method following this approach.

2.2. Simple Pólya Urn scheme

SPU works on colored balls and urns. When a ball in a particular color is drawn from an urn, the ball is put back to the urn along with a new ball in the same color as illustrated in Fig. 1(a). The generative process of LDA could be interpreted by SPU [14]. In the topic modeling context, a word could be viewed as a ball in a certain color and a topic could be viewed as an urn. The word distribution of a topic is reflected by the color proportions of balls in the urn. The specific process illustrated in Fig. 1(a) corresponds to assigning a word to a specific topic in the Gibbs sampling process in topic modeling. The standard LDA model, following the SPU scheme, could capture the word co-occurrence patterns in a text corpora. However, such simple scheme does not consider the semantic relatedness of words.

2.3. Generalized Pólya Urn scheme

GPU is proposed to overcome the aforementioned limitation of SPU [15]. GPU differs from SPU in that, when a ball in a certain color is drawn, two balls in the same color are put back along with a certain number of balls in some other similar colors, which is illustrated in Fig. 1(b). These added balls in other similar colors are placed into the urn to increase the proportions of other similar colors in the urn, which number is decided by a promotion matrix. In this way, external knowledge can be incorporated into topic modeling. For example, Chen and Liu [7] proposed using frequent pattern mining to extract words relating to the same product aspect from product reviews and incorporated the acquired knowledge into the GPU-based topic model. Li et al. [16] used the GPU scheme in the Dirichlet Multinomial Mixture (DMM) model for short text clustering. But these GPU based topic models rely on the construction of the word promotion matrix and do not consider the impact of promoted other similar words on the existing topic sampling results.

3. Weighted Pólya Urn scheme

In this section, we introduce the proposed Weighted Pólya Urn (WPU) scheme, a new type of Pólya Urn scheme.

Suppose there are K urns (e.g., indexed by [1,2,...,K]), and each urn initially contains the same number of standard balls in one of the V different colors. The weight of each standard ball is initially

Please cite this article as: R. Wang, D. Zhou and Y. He, Optimising topic coherence with Weighted Pólya Urn scheme, Neurocomputing, https://doi.org/10.1016/j.neucom.2019.12.013

1. We repeatedly choose a colored ball from these urns until convergence in three following steps: (1) choose an urn k according to the K-dimensional Multinomial distribution which is proportional to the number of balls in K urns; (2) choose a ball in the vth color from the selected k urn with the probability proportional to the total cumulative weight of all balls with the vth color in urn k; (3) put back the selected ball and a new ball in the vth color, with weight, which is derived based on its similarity to the main color family in the urn k. As illustrated in the left of Fig. 1(c), as the color v (red) is similar to the dominated color (red) of urn k, a new red ball with a higher weight (e.g., 1.2 or 1.5) is added to urn k. Otherwise, a new red ball with lower weight (e.g., 0.5 or 0.8) is added to urn k as illustrated in the right of Fig. 1(c).

When the process stops, a V-dimensional Multinomial distribution proportional to the cumulative weights of different colored balls in each urn is obtained. Obviously, WPU inherits the merit of SPU (*'the rich get richer'*) due to the mechanism of adding same color balls. Furthermore, it also has the strength of *'weigh more if more similar'* by considering the similarity between the sampled color and the dominated color in the urn.

In the context of topic modeling, urns are topics and V different colored balls correspond to V distinct words in the vocabulary. The dominated color of an urn denotes the semantic meaning of a specific topic. The above process could be viewed as a special case of sampling based on WPU. Here, the semantic similarity between the sampled color and the dominated color in an urn is defined as the semantic coherence between the word and the topic, which plays a crucial role in WPU. By using semantic coherence, external knowledge of word semantic relatedness can be incorporated into topic modeling, which will be described in more details in the following section.

4. Self-enhancement topic modeling

In this section, we describe our proposed self-enhancement topic modeling (SE-TM) based on WPU. The training procedure of SE-TM contains two phases, the burn-in phase and the self-enhancement phase. The burn-in phase aims to generate the topic candidates without the incorporation of external knowledge. In the self-enhancement phase, the semantic prior knowledge, stored in a word similarity matrix, and the topic candidates are used to calculate the semantic coherence of each word with different topics. The semantic coherence information is employed by collapsed Gibbs sampling for self-enhancement topic modeling. As the topic candidates evolve in each iteration, the semantic coherence is updated dynamically. In the following, we first describe how we build the semantic coherence matrix from the word similarity matrix, followed by the collapsed Gibbs sampling procedure.

4.1. Word similarity matrix

Recently, word representation learning has gained an increasing interest due to its improved efficiency and effectiveness in mapping one-hot encoding into a low-dimensional space [11]. The learned word embeddings have been proved to encode numerous semantic relations (e.g. similarity and morphology) based on the context of words [8]. Many approaches have been proposed to enhance the topic modeling results using the word embeddings that words with similar semantic and syntactic properties are close to each other in the embeddings space [17,18].

Here, we use the word embeddings to construct the word similarity matrix C with the dimension of $V \times V$ to capture the semantic similarity between words, where V is the vocabulary size. The i, jth entry of the word similarity matrix contains the cosine similarity between two word embeddings $\vec{v_i}$ and $\vec{v_j}$, i.e., $C[i, j] = \cos(\vec{v_i}, \vec{v_j})$.

4.2. Semantic coherence matrix

As discussed before, each word has distinct semantic coherence to different topics which can be calculated based on the intermediate topic sampling results. The semantic coherence matrix, S with the dimension of $K \times V$ is used to store the semantic coherence values of words with topics.

During the self-enhancement phase, the posterior topic-word distribution is calculated at the end of current Gibbs sampling iteration. Then, we get the intermediate representative words in each topic RW^i (ranked by the topic-word probability in the descending order and the superscript i denotes the ith iteration) and their corresponding probabilities P^i. Both RW^i and P^i are $K \times M$ matrix and M is the number of representative words for a topic, which is set to 10 in our experiments. RW^i and P^i capture the information about the prominent words and their composition probabilities respectively of the generated topics.

For each word v, we calculate its correlation value with the representative words of a topic k based on the following equation,

$$CV[k, v] = \sum_{m=1}^{M} P^i[k, m] \cdot C[v, RW^i[k, m]], \tag{1}$$

where $P^i[k, m]$ is the probability of mth representative word in topic k at ith iteration, $RW^i[k, m]$ is the mth representative word in the topic k at ith iteration, $C[i, j]$ denotes the cosine similarity value between two words $\vec{v_i}$ and $\vec{v_j}$.

$CV[\cdot, v]$ captures the semantic coherence of v with different topics. Next, we map the $CV[\cdot, v]$ into an arithmetic progression $\hat{CV}[\cdot, v]$ (normalized correlation value) ranging from -0.5 to 0.5. For example, in the 5-topic case, $CV[\cdot, v]$=[1.2, 0.4, -0.4, -0.05, 0.01] will be mapped to $\hat{CV}[\cdot, v]$ =[0.5, 0.3, -0.3, -0.1, 0.1]. Obviously, $\hat{CV}[\cdot, v]$ keeps the relative ranking of semantic similarity values in the original $CV[\cdot, v]$. Finally, the semantic coherence value of v to the topic k could be calculated based on the following equation,

$$S[k, v] = 1 + tc \cdot \hat{CV}[k, v], \tag{2}$$

where the *trust coefficient*, $tc \in [0, \infty)$, is the hyper-parameter to describe the confidence of using the knowledge provided by word embeddings.

4.3. Gibbs sampling

In this subsection, we introduce the collapsed Gibbs sampling algorithm for the proposed Self-Enhancement Topic modeling approach.

As the semantic coherence matrix S will be updated dynamically according to the correlation between a word and topics in the self-enhancement phase, we use the cumulative semantic coherence value matrix W ($K \times V$) to calculate the sampling distribution. Besides, we also observe that the semantic coherence value of a word with a topic does not always equal to 1. During the derivation process of posterior conditional distribution, the real values in W and S make the derivation more intricate to give a exact Gibbs sampling distribution than SPU based model because that the property of gamma function (e.g. $\Gamma(x) = (x-1)!$) could not be employed to simplify the derivation. Due to the fact that each element in S values around 1, we make an assumption that the semantic coherence value of the current word $w_{d,n}$ (valued v') with each topic equals to 1. We follow Mimno et al. [15] and mimic the true Gibbs sampling distribution with a approximated one. Based on this assumption and the property of gamma function, an analytical posterior conditional distribution is obtained

Please cite this article as: R. Wang, D. Zhou and Y. He, Optimising topic coherence with Weighted Pólya Urn scheme, Neurocomputing, https://doi.org/10.1016/j.neucom.2019.12.013

which is shown in Eq. (3).

$$p(z_j = k' | \vec{z_{-j}}, \vec{w}) \propto \frac{n^{k'}_{d,-j} + \alpha_{k'}}{\sum_{k=1}^{K} n^{k}_{d,-j} + \alpha_k} \cdot \frac{W[k, v']_{-j} + \beta_{v'}}{\sum_{v=1}^{V} W[k, v]_{-j} + \beta_v} \quad (3)$$

where k' is the sampled topic, the subscript j is the current index (d, n), $-j$ means the count or semantic coherence value of the current word is excluded, n^k_d refers to the number of times that topic k has been observed with a word of document d. The $W[k, v]$ corresponds to the cumulative weight of balls with color v in the kth urn. The detailed sampling scheme is shown as Algorithm 1.

Algorithm 1 Training procedure for SE-TM.

Input: $\mathcal{D}$, K, C, α, β, tc, *burn_in*, *max_iter*
Output: the posterior topic word distributions Φ

/* *Initialize semantic coherence matrices* */
Initialize S_{cur}, S_{last} with all entries set to 1.
for $i \leq$ max_*iter* **do**
  if $i \leq$ *burn_in* **then**
    /* *burn-in phase, (similar to LDA)* */
    for $d \in \mathcal{D}$ **do**
      for $w_{d,n} \in \boldsymbol{w}^d$ **do**
        $k = z_{d,n}$, $v' = w_{d,n}$
        $n^k_d = n^k_d - 1$
        $W[k, v'] = W[k, v'] - S_{last}[k, v']$
        sample a new topic k' via Eq. (3)
        $n^{k'}_d = n^{k'}_d + 1$
        $W[k', v'] = W[k', v'] + S_{cur}[k', v']$
      end for
    end for
  else
    /* *self enhancement phase* */
    for $d \in \mathcal{D}$ **do**
      for $w_{d,n} \in \boldsymbol{w}^d$ **do**
        $k = z_{d,n}$, $v' = w_{d,n}$
        $n^k_d = n^k_d - 1$
        $W[k, v'] = W[k, v'] - S_{last}[k, v']$
        sample a new topic k' via Eq. 3
        $n^{k'}_d = n^{k'}_d + 1$
        $W[k', v'] = W[k', v'] + S_{cur}[k', v']$
      end for
    end for
    $\Phi^i \leftarrow getTopicWordDistribution()$
    $RW^i \leftarrow getRepresentWords(\Phi^i)$
    $P^i \leftarrow getProbOfRW(RW^i, \Phi^i)$
    $S_{last} \leftarrow S_{cur}$
    $S_{cur} \leftarrow updateSMatrix(C, RW^i, P^i)$
  end if
end for

5. Experiments

In this section, we first introduce the corpora we used for our experiments, and then describe the baseline approaches, finally present the experimental results.

5.1. Experimental setup

Seven publicly accessible corpora are used for evaluation: NIPS dataset, Grolier dataset, Clinical dataset and four product review datasets. Details are summarized below:

- *NIPS dataset*[1] has been widely used in topic modeling experiments. It is the collection of NIPS articles and most articles are related to neural networks and machine learning.
- *Grolier dataset*[1] is built from Grolier Multimedia Encyclopedia, and its content covers almost all the fields, such as sports, economy, politics and etc.
- *Clincial dataset* [19] has been released for i2b2 Natural Language Processing Challenges. The dataset contains 1249 discharge summaries and has many domain specific words such as medical nomenclature and drug names.
- *Product review datasets* [20] [2] contain Amazon product reviews in different categories. We select four datasets, *mp3, laptop, video* and *mobile* for our experiments.

We choose the following six topic models as the baselines:

- **LDA** [1], is a topic model that generates topics based on word co-occurrence patterns from documents. We implement the LDA model and set the Dirichlet prior of the document-topic distribution $\alpha = 50/K$ and the Dirichlet prior of the topic-word distributions $\beta = 0.01$, following what have been suggested in [21].
- **GPU-LDA** [15], is a topic model based on the Generalized Pólya Urn scheme. We implement the algorithm using Mallet[3].
- **GPU-LDA+embedding**, is a variant version of GPU-LDA [15]. Due to the usage of word embedding in the proposed SE-TM, we employ this variant based on GPU-LDA to make a fair comparison. Unlike the GPU-LDA, the element of the promotion matrix A is obtained based on the cosine similarity of word embeddings of two words. More concretely, $A_{i,j}$ is set to 0.1 if the cosine similarity of i and j is higher than 0.7.) We implement the algorithm using Mallet.
- **LDA-VAE** [22], is a neural topic model based on variational autoencoder. We use the implementation in the original paper [4].
- **ProdLDA** [22], is a variant of LDA-VAE, in which the distribution over individual words is a product of experts rather than the mixture model used in LDA. The original implementation is used.
- **Gaussian-LDA** [23], uses a multivariate gaussian distribution to model the topic-word distribution and takes the word embeddings as input. The original implementation[5] with default configuration is employed.

We obtain the processed NIPS and Grolier dataset online[1]. For clinical dataset, the NLTK stopword list is employed to remove the common words. For the four product review datasets, we follow [6,24] to consider each sentence as a document and perform the following pre-processing steps: (1) convert letters to lowercase; (2) check the spelling using pyenchant[6] and remove mis-spelled words; (3) remove the words with fewer than 3 character; (4) remove the top ten frequent words. The statistics of processed corpora are shown in Table 1.

To incorporate the semantic knowledge into SE-TM and calculate the semantic coherence value between words and topics, three popular word embeddings are employed. Most of our experiments use the 50-dimensional pre-trained Glove word embeddings [25]. To explore whether the performance of the proposed approach is sensitive to the word embeddings used, we use the gensim library[7] to build a 200-dimensional word2vec embeddings [9] with

[1] https://cs.nyu.edu/~roweis/data/.
[2] http://sifaka.cs.uiuc.edu/~wang296/Data/index.html.
[3] http://mallet.cs.umass.edu/.
[4] https://github.com/akashgit/autoencoding_vi_for_topic_models.
[5] https://github.com/rajarshd/Gaussian_LDA.
[6] https://pypi.python.org/pypi/pyenchant/.
[7] https://radimrehurek.com/gensim/apiref.html.

Please cite this article as: R. Wang, D. Zhou and Y. He, Optimising topic coherence with Weighted Pólya Urn scheme, Neurocomputing, https://doi.org/10.1016/j.neucom.2019.12.013

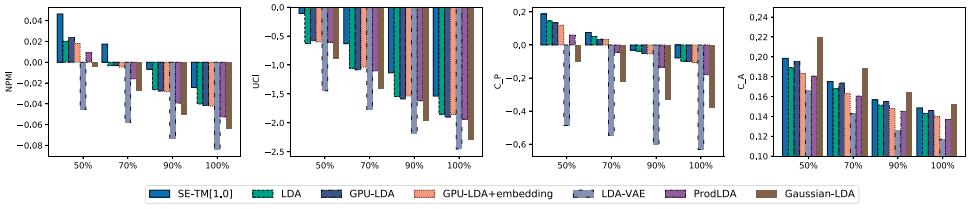


Fig. 2. Average topic coherence vs. different topic proportion on Clinical dataset.

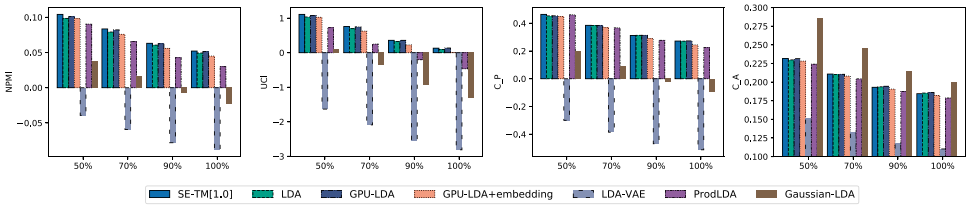


Fig. 3. Average topic coherence vs. different topic proportion on NIPS dataset.

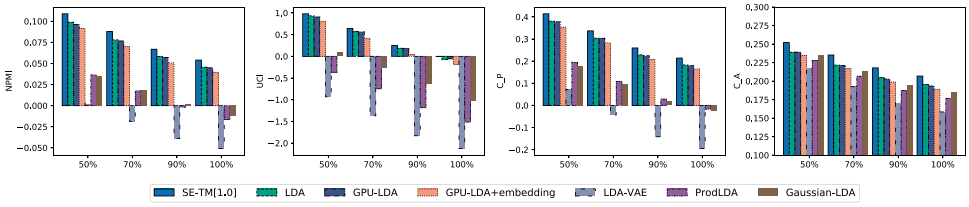


Fig. 4. Average topic coherence vs. different topic proportion on Grolier dataset..

Table 1
The statistics of 7 corpora, each sentence is viewed as a document in Laptops, MP3, Video and Mobile corpora.

Domain	#Document	#Words
NIPS	1500	12,375
Grolier	29,762	15,276
Clinical	1249	7832
Laptops	394,548	6907
MP3	324,436	5725
Video	429,351	6744
Mobile	1,236,465	10,243

default configuration. Also, we use the fastText library[8] to build a 100-dimensional fastText embeddings [11]. The two embeddings are trained on the latest Wikipedia corpus[9].

The trust coefficient *tc* of the proposed approach is set to [0.6, 1.0, 1.5] in our experiments. Also, the burn-in and the Gibbs sampling iterations are set to 200 and 1000, respectively. Typically topic models are evaluated based on the likelihood of held-out documents. However, as pointed out in [26], higher likelihood of held-out documents doesn't necessarily correspond to human judgement of topic coherence. Therefore, in this paper, we follow [27] and select four common coherence metrics including UCI (a coherence measure based on a sliding window and the pointwise mutual information of all word pairs of the given topics), NPMI (an enhanced version of the UCI coherence using the normalized pointwise mutual information), C_P (a coherence measure based on a sliding window, a one-preceding segmentation of the given words and the confirmation measure of Fitelson's coherence) and C_A (a coherence measure based on a context window, a pairwise comparison of the given words and an indirect confirmation measure that uses normalized pointwise mutual information and the cosine similarity) to evaluate the topics generated by models. A higher value implies more coherent topics. In our evaluation, we take the top 10 words sorted by their topic-word probabilities to represent each topic and compute the topic coherence using the Palmetto library[10].

5.2. Topic coherence vs. different topic proportion

To compare the performance of the proposed approach with the baselines, we firstly make a comparison of topic coherence vs. different topic proportions. Experiments are conducted on all the seven corpora with six topic number settings [10, 20, 30, 50, 75, 100]. To compare the comprehensive performance, we calculate the average topic coherence among topics whose coherence values are ranked at the top 50% (or 70%, 80%, 90%, 100%) positions. For example, to calculate the average NPMI coherence of LDA @ 50%, we only select topics whose NPMI are ranked at the top 50% positions for each dataset, and then average the NPMI values of those topics obtained using LDA under different topic number setting for each datasets. Experimental results on the average topic coherence vs. different topic proportions are shown in Figs. 2–8. In Figs. 2–8, the values are obtained by SE-TM with *tc* is set to 1.0, and the Glove embeddings are employed in this subsection.

It can be observed from Figs. 2–8 that our proposed SE-TM model outperforms LDA, GPU-LDA, GPU-LDA+embedding, LDA-VAE, ProdLDA and Gaussian-LDA on first three topic coherence measures (NPMI, UCI and C_P). It should be pointed out that the improvement of SE-TM is slight on NIPS dataset, and this may because that it has too many words have domain specific meaning which are not similar to their common sense stored in word embedding. For C_A measure, the Gaussian-LDA performs the best, and our SE-TM performs the second best across all the datasets. According to the observation in [27] that NPMI and C_P are better than UCI, C_A in terms of their correlations to human judgment,

[8] https://github.com/facebookresearch/fastText.
[9] https://dumps.wikimedia.org/enwiki/latest/enwiki-latest-pages-articles.xml.bz2.
[10] https://github.com/dice-group/Palmetto.

Please cite this article as: R. Wang, D. Zhou and Y. He, Optimising topic coherence with Weighted Pólya Urn scheme, Neurocomputing, https://doi.org/10.1016/j.neucom.2019.12.013

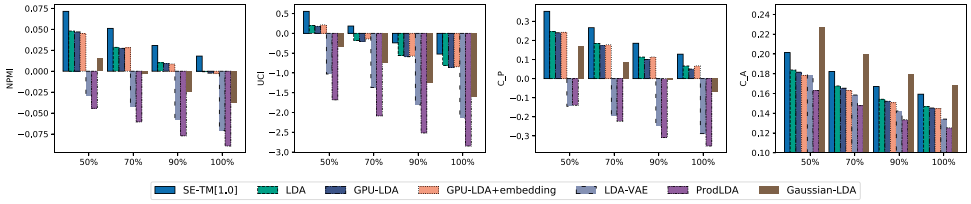


Fig. 5. Average topic coherence vs. different topic proportion on Laptops dataset.

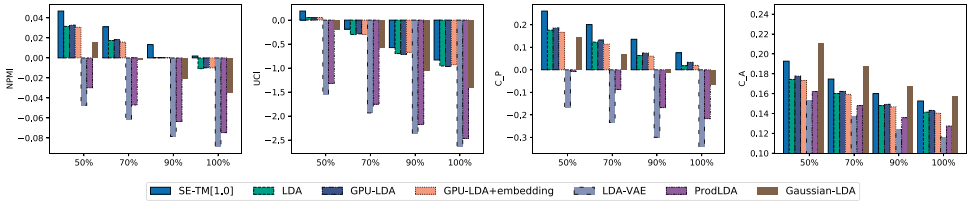


Fig. 6. Average topic coherence vs. different topic proportion on MP3 dataset.

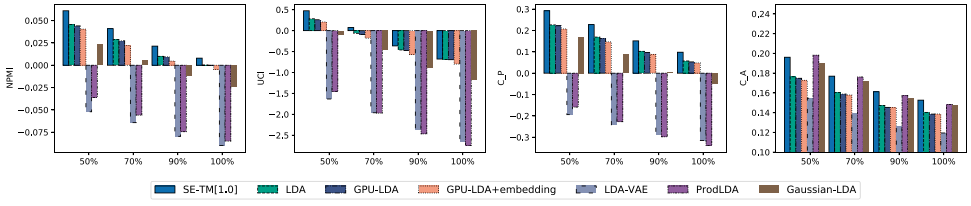


Fig. 7. Average topic coherence vs. different topic proportion on Video dataset.

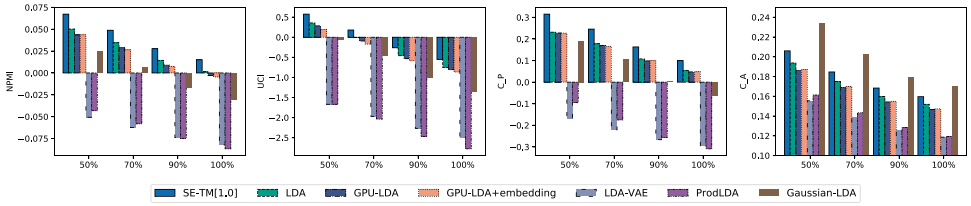


Fig. 8. Average topic coherence vs. different topic proportion on Mobile dataset.

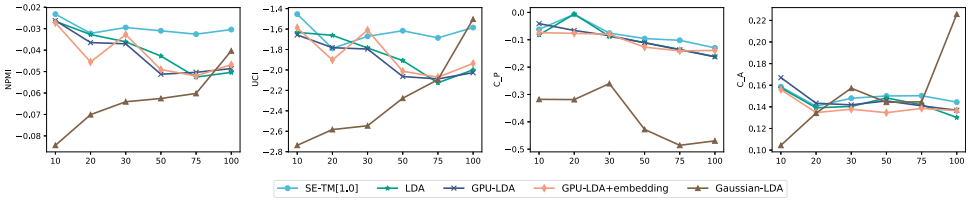


Fig. 9. Average topic coherence (100%) of Clinical dataset vs. different topic number.

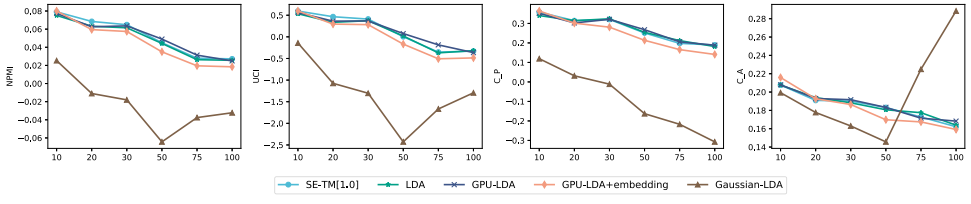


Fig. 10. Average topic coherence (100%) of NIPS dataset vs. different topic number.

we could conclude that SE-TM generally obtains more coherent topics. Meanwhile, it also shows that explicitly taking into account semantic coherence of a word with different topics during the Gibbs sampling process could indeed help to achieve better topic coherence compared to topic models following the SPU and GPU scheme.

5.3. Topic coherence vs. different topic number

Furthermore, to explore how topic coherence results vary with different topic numbers, we show in Figs. 9–15 the average topic coherence of each corpus (with all topics taken into account) vs. different topic number settings. As shown in Figs. 2–8 that

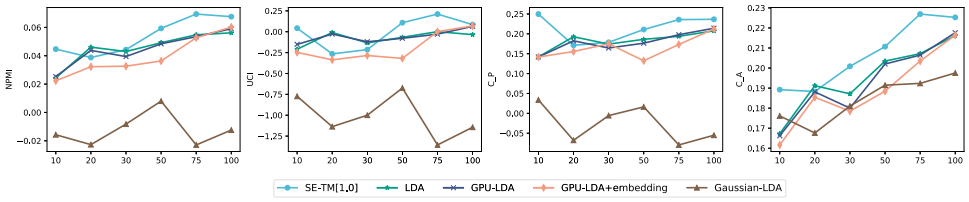


Fig. 11. Average topic coherence (100%) of Grolier dataset vs. different topic number.

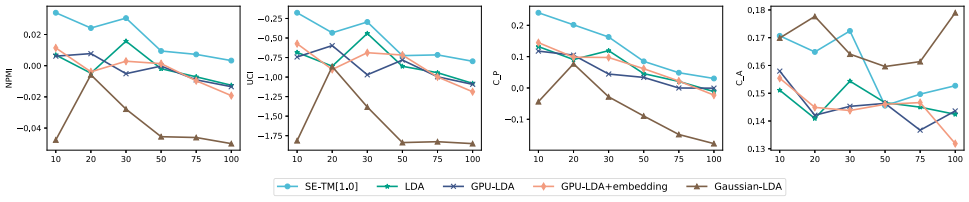


Fig. 12. Average topic coherence (100%) of Laptops dataset vs. different topic number.

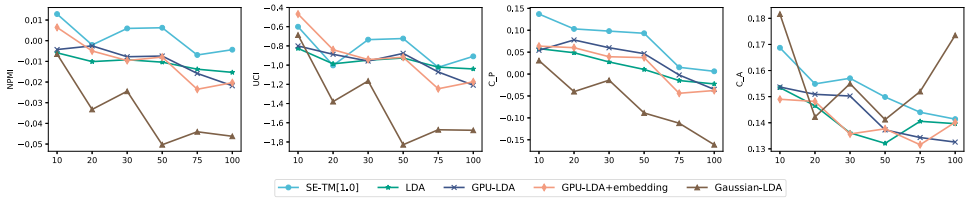


Fig. 13. Average topic coherence (100%) of MP3 dataset vs. different topic number.

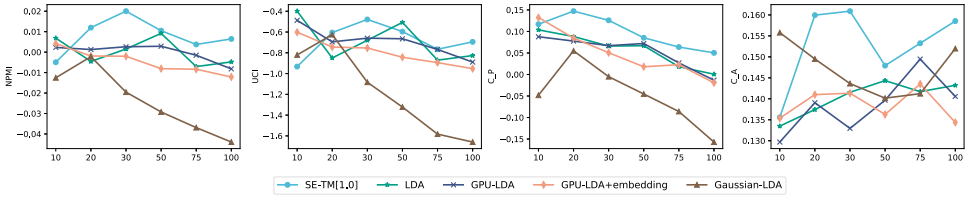


Fig. 14. Average topic coherence (100%) of Video dataset vs. different topic number.

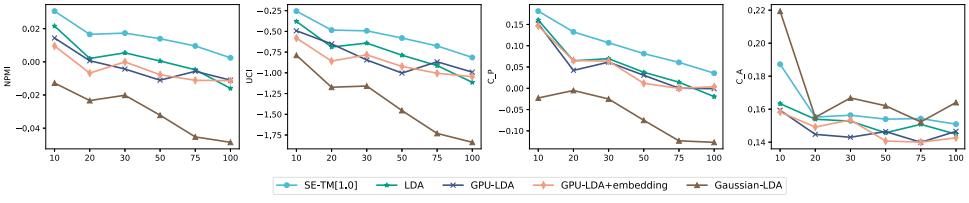


Fig. 15. Average topic coherence (100%) of Mobile dataset vs. different topic number.

LDA-VAE and ProdLDA often performs the worst compared with other approaches, to simplify the figure and make a more clearly comparison, we remove the curves obtained by these two models and only show the partial strong baselines in Figs. 9–15. In this subsection, SE-TM[1.0] with Glove embeddings is also employed in the experiments.

It can be observed that SE-TM outperforms the baselines in most cases in terms of NPMI, UCI, C_P measures. For C_A measure, Gaussian-LDA obtains the best average coherence values except for Grolier and Video datasets. Also, SE-TM performs at least the second best on this measure. Generally, we could find that the results shown in Figs. 9–15 are consistent with the observations in Section 5.1. Besides, with the increasing number of topics, the topic coherence of SE-TM drops slightly.

More concretely, we could have another two observations from Figs. 9–15: (1) SE-TM do not performs very well for NIPS dataset, (2) GPU-LDA+embedding and GPU-LDA do not improve the quality of the extracted topic for most datasets. This maybe caused by the following factors: (i). NIPS dataset has too many words has domain specific meaning which are not similar to their common sense. For example, the nearest neighbor words of 'neural' in this domain maybe 'network', 'input' and 'loss', while in common sense, some medical terms (such as 'neurology', 'cell' and 'synapse') are its most similar words. Thus, the gap between domain knowledge and common knowledge contained in word embeddings lead to the insufficient improvement. (ii). During the inference procedure of GPU based approaches, the corresponding sampler blindly incorporate the pre-obtained semantic relatedness into the model without considering any information intermediate topic representations. And this mechanism may inject the wrong knowledge into the model and further deteriorate the quality of the extracted topics.

Please cite this article as: R. Wang, D. Zhou and Y. He, Optimising topic coherence with Weighted Pólya Urn scheme, Neurocomputing, https://doi.org/10.1016/j.neucom.2019.12.013

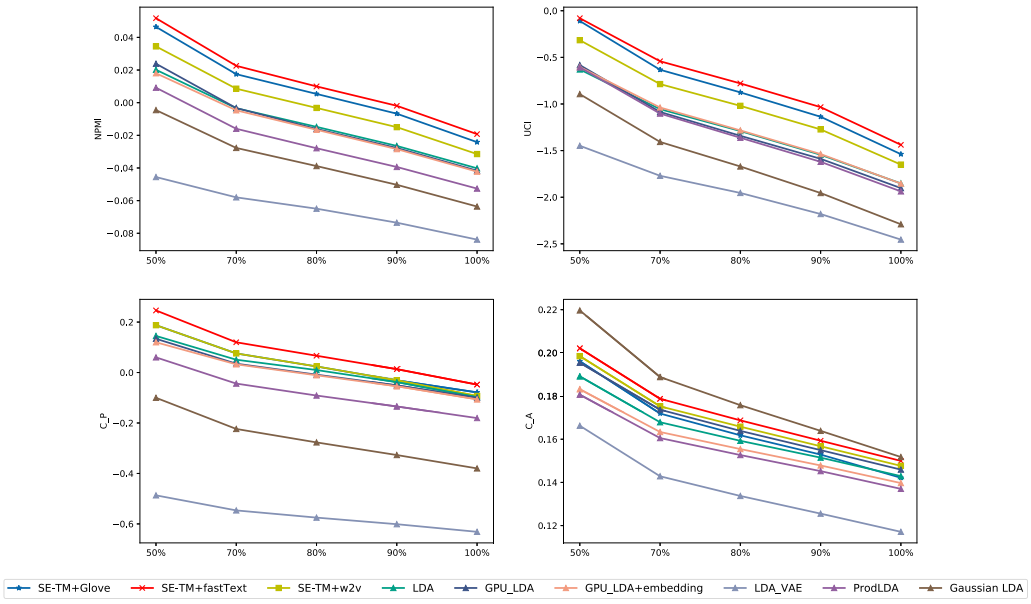


Fig. 16. SE-TM using different word embeddings in comparison with baseline approaches.

Table 2
Word embedding comparison on the clinical dataset, the coherence values are calculated based on 100% topics.

Model	NPMI	UCI	C_P	C_A
SE-TM+Glove	−0.024	−1.538	−0.132	0.142
SE-TM+fastText	**−0.019**	**−1.438**	**−0.048**	0.150
SE-TM+w2v	−0.032	−1.650	−0.093	0.148
LDA	−0.040	−1.852	−0.097	0.143
GPU-LDA	−0.042	−1.902	−0.100	0.146
GPU-LDA+embedding	−0.043	−1.853	−0.106	0.139
LDA-VAE	−0.063	−2.454	−0.631	0.117
ProdLDA	−0.052	−1.937	−0.181	0.137
Gaussian-LDA	−0.064	−2.290	−0.380	**0.152**

Table 3
The comparison of baselines and SE-TM using different *tc* configuration on the Grolier dataset.

Model	NPMI	UCI	C_P	C_A
SE-TM[1.5]	0.0530	−0.0524	0.2077	0.2064
SE-TM[1.0]	0.0540	−0.0053	**0.2137**	**0.2068**
SE-TM[0.6]	**0.0544**	**0.0335**	0.2133	0.2033
LDA	0.0455	−0.0753	0.1826	0.1954
GPU-LDA	0.0449	−0.0570	0.1794	0.1934
GPU-LDA+embedding	0.0394	−0.1861	0.1649	0.1889
LDA-VAE	−0.0507	−2.1231	−0.1947	0.1584
ProdLDA	−0.0164	−1.5172	−0.0177	0.1768
Gaussian-LDA	−0.0123	−1.0141	−0.0263	0.1843

5.4. Word embeddings and parameter analysis

To further study the impact of the different word embeddings and hyper-parameters, we provide a more detailed comparison regard to word embedding, the number of representative words *M* and trust coefficient *tc* in this subsection.

5.4.1. Word embedding comparison

To explore whether the proposed approach is sensitive to the specific type of word embedding, experiments of word embedding comparison have been carried out.

In this subsection, we conduct experiments using three different word embeddings, *Glove, word2vec* and *fastText*. We show in Table 2 the topic coherence results on the clinical dataset. Fig. 16 represents the average topic coherence over top 50%, 70%, 80%, 90% and 100% topics. It can be observed that SE-TM+*fastText* the best. It might attribute to the fact that the clinical dataset contains many medical terms (e.g., drug names or disease names) and fastText based on the composition of character *n*-grams for word embedding learning captures well the semantics of those medical terms.

5.4.2. The impact of the hyper-parameter trust coefficient tc

To explore the impacts of the hyper-parameter *tc* and validate that our proposed SE-TM could perform well with different *tc* settings, we provide a comparison experiment with different trust coefficient on Grolier dataset.

In this subsection, *tc* is set to 1.5, 1.0 and 0.6, respectively, Table. 3 lists the statistics of the average topic coherence values which are computed based 100% topic and best values are highlighted in bold. From the comparison between three SE-TMs in Table. 3, we could observe that the SE-TM[1.0] performs slightly better with two best values (C_P and C_A) and two second best values (NPMI and UCI). It can be explained that a lower *tc* results in the ignorance of word relatedness semantics derived from word embeddings and a higher *tc* may incorporate the semantic knowledge which is not suitable for a specific domain and hence gives slightly worse coherence results. Besides,we also make a comparison between SE-TMs and baselines in Fig. 17. Due to the similar values could be obtained by SE-TM with different *tc*, we only plot the best curve in Fig. 17 to avoid overlapping and give a clear comparison.

5.4.3. The impact of the hyper-parameter M

To explore the impacts of the hyper-parameter *M* and validates the robustness of the proposed approach, experiments have been carried out on Clinical dataset.

In this subsection, *M* is set to 6, 7, 8, 9, 10 respectively, and the average topic coherences over top 50%, 70%, 80%, 90%, 100% are represented in Fig. 18. Besides, Table. 4 lists the statistics of the average coherence values which are computed based on 100% topics,

Please cite this article as: R. Wang, D. Zhou and Y. He, Optimising topic coherence with Weighted Pólya Urn scheme, Neurocomputing, https://doi.org/10.1016/j.neucom.2019.12.013

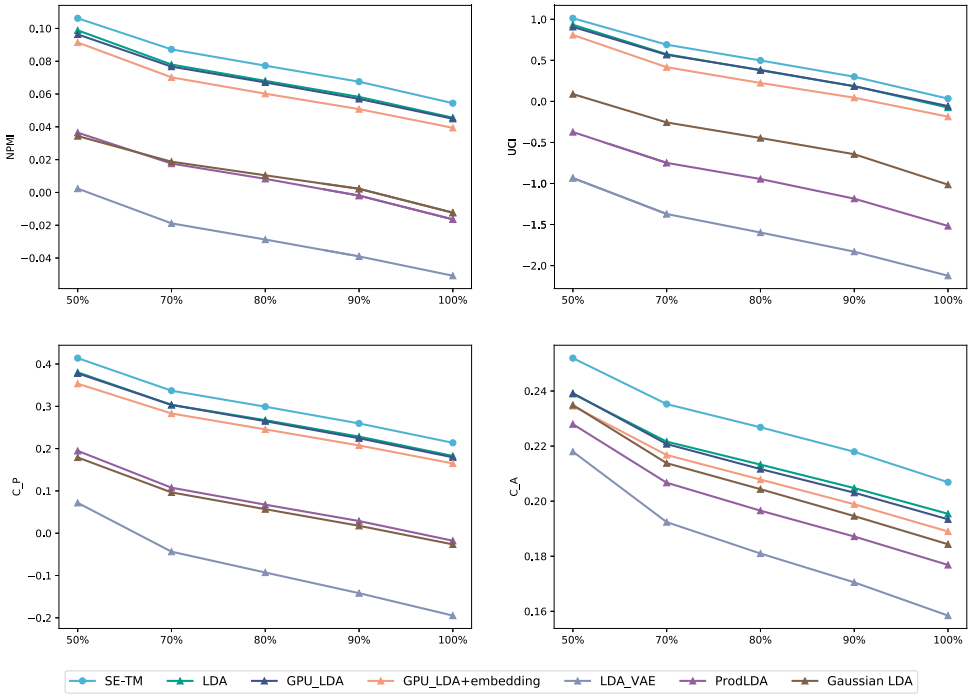


Fig. 17. Comparison between best coherence obtained by SE-TM using various *tc* and baseline approaches.

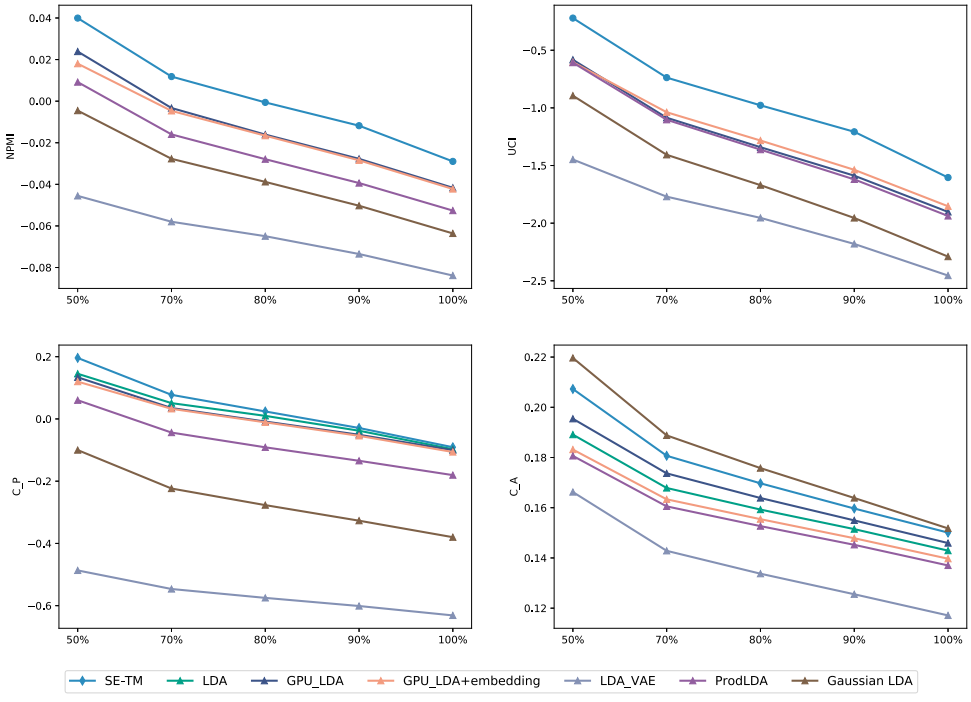


Fig. 18. Comparison between best coherence obtained by SE-TM using various *M* and baseline approaches..

and best values are highlighted in bold. To avoid the overlapping caused by similar values obtained by SE-TM, we only plot the best curves obtained by SE-TM in Fig. 18, and it can be observed that SE-TM outperform the compared baselines on all measures except 'C_A'. Though Gaussian-LDA obtains a high 'C_A' score, it performs the second worst on 'NPMI', 'UCI' and 'C_P' measures. Moreover, it can be also observed that the hyper-parameter *M* do not impact the topic quality of SE-TM. Namely, the proposed SE-TM is not sensitive to the configuration of the hyper-parameter *M*.

5.5. Example topics

To directly compare the topic representations extracted by different approaches, we make a qualitative comparison in this subsection. Table 5 shows the top 10 words of example topics extracted using our proposed SE-TM in comparison with GPU-LDA on the Laptops corpus with the topic number set to 75. And the topics may denote '*macbook*', '*size & weight*', '*keyboard*', '*mouse*' and '*battery life*'. It can be observed that SE-TM gives more coher-

Please cite this article as: R. Wang, D. Zhou and Y. He, Optimising topic coherence with Weighted Pólya Urn scheme, Neurocomputing, https://doi.org/10.1016/j.neucom.2019.12.013

Table 4
The comparison of baselines and SE-TM using different M configurations on the clinical dataset.

Model	NPMI	UCI	C_P	C_A
SE-TM(M=10)	−0.032	−1.650	−0.093	0.148
SE-TM(M=9)	−0.030	−1.646	**−0.090**	0.151
SE-TM(M=8)	−0.030	−1.616	−0.097	0.150
SE-TM(M=7)	−0.031	−1.658	−0.096	0.147
SE-TM(M=6)	**−0.028**	**−1.604**	−0.105	0.145
LDA	−0.040	−1.852	−0.097	0.143
GPU-LDA	−0.042	−1.902	−0.100	0.146
GPU-LDA+embedding	−0.043	−1.853	−0.106	0.139
LDA-VAE	−0.063	−2.454	−0.631	0.117
ProdLDA	−0.052	−1.937	−0.181	0.137
Gaussian-LDA	−0.064	−2.290	−0.380	**0.152**

Table 5
Example topics of SE-TM vs. GPU-LDA on Laptops corpus, italics means out-of-topic.

Topic 1	Topic 2	Topic 3	Topic 4	Topic 5
SE-TM				
mac	size	keyboard	mouse	battery
apple	small	key	pad	life
version	full	type	button	hour
book	light	function	click	charge
linux	perfect	typing	finger	long
compatible	large	backlit	touchpad	last
mini	heavy	style	touch	longer
ultrabooks	carry	layout	cursor	cell
osx	fit	volume	scroll	depending
ipod	smaller	delete	scrolling	charged
GPU-LDA				
mac	*tablet*	keyboard	mouse	battery
book	size	type	touchpad	life
decided	full	typing	click	hour
wanted	perfect	*nice*	button	long
apple	small	*hand*	finger	last
try	*ipad*	*comfortable*	trackpad	charge
mbp	*keyboard*	*rest*	*left*	cell
expensive	*netbook*	*feel*	*move*	longer
display	smaller	trackpad	pad	*advertised*
fact	larger	*edge*	*gesture*	depending

ent topic results compared to GPU-LDA. For example, Topic 2 is about '*size & weight*'. But some words returned by GPU-LDA are less semantically relevant to '*size & weight*', such as '*tablet*', '*ipad*', '*keyboard*' and '*notebook*' highlighted in italics in Table 5.

6. Conclusion

In this paper, we have proposed a Weighted Pólya Urn model. It has two merits that could be described as *'the rich get richer'* and *'weigh more if more similar'*. We also model the semantic coherence of a word with different topics by incorporating the Weighted Pólya Urn scheme into the topic inference process and propose the self-enhancement topic modeling approach. The SE-TM employs the intermediate topic sampling results and the general semantic knowledge provided by word embeddings to guide the topic sampling in subsequent Gibbs sampling iterations. Experimental comparison with the state-of-the-art approaches shows an improved coherence score of generated topics and stable performance across different domains.

Declaration of Competing Interest

We declare that we have no financial and personal relationships with other people or organizations that can inappropriately influence our work, there is no professional or other personal interest of any nature or kind in any product, service and/or company that could be construed as influencing the position presented in, or the review of, the manuscript entitled.

Acknowledgements

We would like to thank anonymous reviewers for their valuable comments and helpful suggestions. This work was funded by the National Natural Science Foundation of China (61772132) and the National Key Research and Development Program of China (2016YFC1306704).

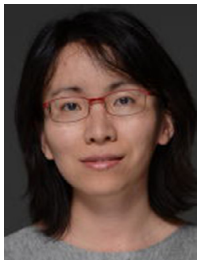
References

[1] D.M. Blei, A.Y. Ng, M.I. Jordan, Latent Dirichlet allocation, J. Mach. Learn. Res. 3 (Jan) (2003) 993–1022.
[2] C. Lin, Y. He, Joint sentiment/topic model for sentiment analysis, in: Proceedings of the ACM Conference on Information and Knowledge Management, ACM, 2009, pp. 375–384.
[3] X. Yan, J. Guo, Y. Lan, X. Cheng, A biterm topic model for short texts, in: Proceedings of the Twenty-second International Conference on World Wide Web, ACM, 2013, pp. 1445–1456.
[4] J. Yin, J. Wang, A Dirichlet multinomial mixture model-based approach for short text clustering, in: Proceedings of the Twentieth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ACM, 2014, pp. 233–242.
[5] D. Andrzejewski, X. Zhu, M. Craven, Incorporating domain knowledge into topic modeling via Dirichlet forest priors, in: Proceedings of the Twenty-sixth Annual International Conference on Machine Learning, ACM, 2009, pp. 25–32.
[6] Z. Chen, A. Mukherjee, B. Liu, M. Hsu, M. Castellanos, R. Ghosh, Leveraging multi-domain prior knowledge in topic models, in: Proceedings of the International Joint Conference on Artificial Intelligence, 2013, pp. 2071–2077.
[7] Z. Chen, B. Liu, Mining topics in documents: standing on the shoulders of big data, in: Proceedings of the Twentieth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ACM, 2014, pp. 1116–1125.
[8] O. Levy, Y. Goldberg, I. Dagan, Improving distributional similarity with lessons learned from word embeddings, Trans. Assoc. Comput. Linguist. 3 (2015) 211–225.
[9] T. Mikolov, K. Chen, G. Corrado, J. Dean, Efficient estimation of word representations in vector space, Comput. Sci. (2013).
[10] L. Vilnis, A. McCallum, Word representations via gaussian embedding, arXiv preprint arXiv:1412.6623 (2014).
[11] P. Bojanowski, E. Grave, A. Joulin, T. Mikolov, Enriching word vectors with subword information, arXiv preprint arXiv:1607.04606 (2016).
[12] B. Athiwaratkun, A.G. Wilson, A. Anandkumar, Probabilistic fasttext for multi-sense word embeddings, arXiv preprint arXiv:1806.02901 (2018).
[13] S. Deerwester, S.T. Dumais, G.W. Furnas, T.K. Landauer, R. Harshman, Indexing by latent semantic analysis, J. Am. Soc. Inf. Sci. 41 (6) (1990) 391–407.
[14] H. Mahmoud, Polya urn models, Crc Texts Stat. Sci. 43 (2) (2008) xii+290.
[15] D. Mimno, H.M. Wallach, E. Talley, M. Leenders, A. Mccallum, Optimizing semantic coherence in topic models, in: Proceedings of the Conference on Empirical Methods in Natural Language Processing, 2010, pp. 262–272.
[16] C. Li, H. Wang, Z. Zhang, A. Sun, Z. Ma, Topic modeling for short texts with auxiliary word embeddings, in: Proceedings of the Thirty-ninth International ACM SIGIR Conference on Research and Development in Information Retrieval, ACM, 2016, pp. 165–174.
[17] D.Q. Nguyen, R. Billingsley, L. Du, M. Johnson, Improving topic models with latent feature word representations, Trans. Assoc. Comput. Linguist. 3 (2015) 299–313.
[18] S. Li, T.S. Chua, J. Zhu, C. Miao, Generative topic embedding: a continuous representation of documents, in: Proceedings of the Meeting of the Association for Computational Linguistics, 2016, pp. 666–675.
[19] Ö. Uzuner, I. Solti, F. Xia, E. Cadag, Community annotation experiment for ground truth generation for the i2b2 medication challenge, J. Am. Med. Inf. Assoc. 17 (5) (2010) 519–523.
[20] H. Wang, Y. Lu, C. Zhai, Latent aspect rating analysis without aspect keyword supervision, in: Proceedings of the Seventeenth ACM SIGKDD international Conference on Knowledge discovery and data mining, ACM, 2011, pp. 618–626.
[21] T.L. Griffiths, M. Steyvers, Finding scientific topics, Proc. Natl. Acad. Sci. 101 (suppl 1) (2004) 5228–5235.
[22] A. Srivastava, C. Sutton, Autoencoding variational inference for topic models, arXiv preprint arXiv:1703.01488 (2017).
[23] R. Das, M. Zaheer, C. Dyer, Gaussian lda for topic models with word embeddings, in: Proceedings of the Meeting of the Association for Computational Linguistics and the International Joint Conference on Natural Language Processing, 2015, pp. 795–804.
[24] I. Titov, R. McDonald, Modeling online reviews with multi-grain topic models, in: Proceedings of the Seventeenth international conference on World Wide Web, ACM, 2008, pp. 111–120.
[25] J. Pennington, R. Socher, C. Manning, Glove: global vectors for word representation, in: Proceedings of the Conference on Empirical Methods in Natural Language Processing, 2014, pp. 1532–1543.
[26] J. Chang, J. Boyd-Graber, S. Gerrish, C. Wang, D.M. Blei, Reading tea leaves: how humans interpret topic models, in: Proceedings of the International Conference on Neural Information Processing Systems, 2009, pp. 288–296.
[27] A. Both, A. Hinneburg, Exploring the space of topic coherence measures, in: Proceedings of the Eighth ACM International Conference on Web Search and Data Mining, 2015, pp. 399–408.

Please cite this article as: R. Wang, D. Zhou and Y. He, Optimising topic coherence with Weighted Pólya Urn scheme, Neurocomputing, https://doi.org/10.1016/j.neucom.2019.12.013

Rui Wang received the M.S degree in computer science and technology from Guangxi University for Nationalities, Nanning, China, in 2016. He is currently pursuing the Ph.D. degree in computer science at Southeast University. His research interest includes natural language processing, sentiment analysis, unsupervised machine learning, neural generative models and evolutionary computation.

Deyu Zhou received the BASc and MEng degrees from Nanjing University, Nanjing, China and the Ph.D. degree in computer science and technology from University of Reading, UK. He is currently a professor in Southeast University, Nanjing, China. His research interest includes natural language processing, opinion mining, event extraction, sentiment analysis and bioinformatics.



Yulan He received the BASc and MEng degrees in Computer Engineering from Nanyang Technological University, Singapore. She obtained her PhD degree in spoken language understanding from University of Cambridge. Currently, she is a professor of Computer Science at the University of Warwick, Coventry CV4 7AL, UK. Her research interest includes sentiment analysis and opinion mining, natural language processing, social media analysis and clinical text mining. She has served as an Area Chair in top natural language processing conferences including ACL, EMNLP and NAACL and a Senior Programme Committee member in AAAI and IJCAI.

Please cite this article as: R. Wang, D. Zhou and Y. He, Optimising topic coherence with Weighted Pólya Urn scheme, Neurocomputing, https://doi.org/10.1016/j.neucom.2019.12.013