IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received November 30, 2019, accepted January 21, 2020, date of publication January 27, 2020, date of current version February 10, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.2969525

Leveraging Global and Local Topic Popularities for LDA-Based Document Clustering

PENG YANG[1,2,3], YU YAO[1,2], AND HUAJIAN ZHOU[1,2]

[1]Key Laboratory of Computer Network and Information Integration, Ministry of Education, Southeast University, Nanjing 211189, China
[2]School of Computer Science and Engineering, Southeast University, Nanjing 211189, China
[3]School of Cyber Science and Engineering, Southeast University, Nanjing 211189, China

Corresponding author: Peng Yang (pengyang@seu.edu.cn)

This work was supported in part by the Consulting Project of Chinese Academy of Engineering under Grant 2020-XY-5, in part by the National Natural Science Foundation of China under Grant 61472080 and Grant 61672155, and in part by the Collaborative Innovation Center of Novel Software Technology and Industrialization.

ABSTRACT Document clustering is of high importance for many natural language technologies. A wide range of computational traditional topic models, such as LDA (Latent Dirichlet Allocation) and its variants, have made great progress. However, traditional LDA-based clustering algorithms might not give good results due to such probabilistic models require prior distributions which are always difficult to define. In this paper, we propose a probabilistic model named tpLDA, which incorporates different levels of topic popularity information to determine the prior LDA distribution, discover the latent topics and achieve better clustering. Specifically, global topic popularity is introduced to reduce the potential distraction in local cluster popularity and the local cluster popularity draws more attention on certain parts of the global topic popularity. The two popularities contribute complementary information and their integration can dynamically adjust statistical parameters of the model. Experimental evaluations on real data sets show that, compared with state-of-the-art approaches, our proposed framework dramatically improves the accuracy of documents clustering.

INDEX TERMS Document clustering, latent Dirichlet allocation, machine learning, topic modeling.

I. INTRODUCTION

With the unprecedented development of the Internet, web is now overloaded with rapidly growing availability of unstructured textual pages or documents. The large volume of textual data can be utilized to text categorization, document retrieval, public opinion monitoring, decision supporting and emergency management [1]. However, large part of that data generated is unstructured and not annotated, which induces that it is difficult to understand how topic information is diffused among documents [2]. One of the most effective solutions to manage this huge amount of data is to automatically cluster them into meaningful clusters. Clustering is considered an important data mining technique in categorizing, summarizing, organizing and classifying text documents. Topic modeling which belongs to soft clustering of documents, is often taken as a different way of categorizing similar content by extracting meaningful topics from the document collection [3].

In topic modeling and document clustering research, a topic is usually defined as a list of terms that having statistically significant semantic relations. Documents are treated to share the underlying topics with different proportions from the perspective of probability. A document can be an email, a book chapter, a blog posts, a journal article and any kind of unstructured text. Topic modeling is an unsupervised learning technique that aims at extracting a pre-specified number of topics from a set of text documents based on statistical concepts. This process requires no labels or prior knowledge about the text to operate. Clustering is a process of grouping objects that behave in the same manner in uniform clusters [4]. So far, state-of-the-art techniques utilize topic models, such as LSA [5], pLSA [6], LDA [7], VBLDA (Online LDA) [8], Replicated Softmax (RSM) [9] and Document Neural Autoregressive Distribution Estimator (DocNADE) variants [10], [11], to extract topics from documents, predict the probability of each word in a given document belonging to

The associate editor coordinating the review of this manuscript and approving it for publication was Xin-Lin Huang.

each topic, and subsequently use the learned latent document representations to perform document clustering task.

Early significant research in topic modeling and document clustering techniques was initiated by developing a LSA model [5] and a probabilistic LSA model (pLSA) [6]. Owing to inadequate statistical foundation and erroneously assumption of Gaussian noise on term frequencies, LSA has recently been criticized [12]. The extension of LSA, pLSA, assigns a probabilistic mixture model to the words in a document, where the mixture components are viewed as representation of topics. Although pLSA is capable of assigning multiple topics to a document, it cannot generalize to unseen documents due to lack of a well-defined generative model [7]. Moreover, it has been shown that pLSA is prone to overfitting and, empirically, overfitting is indeed a serious problem. Hence, pLSA is not an appealing approach for topic modeling on documents.

LDA is another typical unsupervised probabilistic learning methodology and exhibits a powerful ability in mining the semantic information from the text data [7]. The LDA approach could reveal underlying content and classify large-scale unstructured online texts into a mixture of hidden topics. LDA obtains the final outcomes of topic–word and document–topic distributions through a posterior maximization with Gibbs sampling. Yet, traditional LDA methods suffers from drawbacks of sensitivity to initialization and are often incapable of measuring the correlation between the results of topic detection.

Recently, deep learning techniques have remedied above problems and have shown remarkable success in exploring semantic representation of words and documents [13]. With layer-wise pre-training [14], neural networks are built to automatically initialize their weight values and tackle topic modeling and document clustering tasks. However, the main problem of deep learning is that there are too many hyper-parameters.

Based on the analysis above, we propose a tpLDA method for document clustering in real-time online streams. First, we segment the dataset by each time period, and then use the LDA method to derive keywords and topic distribution. Second, named entity is introduced to characterize the detected topics. Third, a clustering algorithm called ddCRP is used to cluster entities into different clusters. Next, the global popularity of a topic as well as the local popularity of internal clusters within a topic are calculated. Last, the integrated popularity is utilized to adjust the LDA hyper-parameters. The superiorities of our approach are summarized as follows:

- Integrate local and global topic popularity information into LDA-based methods for the calculation of priori distribution. Coarse-grained global document-level topic popularity is introduced to reduce the potential distraction in finer-grained cluster popularity and the aggregated local cluster popularity draws more attention on certain parts of the topic popularity.
- Named entities are utilized to characterize the detected topics and their internal clusters.
- We illustrate its advantages by comparing our proposed LDA-based methods with state-of-the-art approaches on real datasets in terms of both intrinsic and extrinsic measures.

II. RELATED WORK

A. TOPIC MODELING

Topic modeling techniques are dedicated to detect latent topics from text corpus automatically, where each topic is defined as a distribution over a group of words. Several methods have been proposed from different perspectives for topic modeling.

Topic modeling has attracted much attention in machine learning, information retrieval and social media modeling. Wartena and Brussee [15] proposed a topic model which uses the induced k-bisecting clustering algorithm to extract and cluster keywords based on different similarity measures. Hou *et al.* [16] represented news as a link-centric heterogeneous network and introduce a unified probabilistic model for topic extraction and inner relationship discovery within events. Lim *et al.* [17] used community detection approaches on a network graph with multiple definitions of vertices and edges for automated topic modeling on Twitter. The algorithm performed better than various baselines in terms of topic coherence, pointwise mutual information, precision, recall and F-score.

Among the above models, probabilistic topic models are unsupervised generative models which model document content as a two-step generation process, that is, each document is observed as a mixture of latent concepts or topics, while each topic is set as probability distributions over vocabulary words [18]. Latent Dirichlet Allocation (LDA) and variants, as typical probabilistic topic models, have been applied to information retrieval [19], text mining [20] and recommendation systems [21]. Interesting applications of LDA have also been reported as a powerful technique to create knowledge and discover useful structure in a stream of literature [22].

LDA is essentially a three-layer (document, topic, word) Bayesian probabilistic model, which has good generalization ability by treating parameters as random variables. Recent studies introduce time information into LDA models to many natural language issues [23]. Hoffman *et al*. develop an online variation Bayes (VB) algorithm for Latent Dirichlet Allocation (LDA), based on online stochastic optimization with a natural gradient step, which can handily analyze massive document collections, including those arriving in a stream [8]. However, online LDA cannot make full use of the differences between topics or topics in different time periods, so it lacks in rationality and accuracy. Lu *et al.* [24] adopted a self-adaptively LDA-based method and experiments results showed that the proposed method can reach the appropriate number of social topics. Balikas *et al.* [25] proposed sentenceLDA, an extension of LDA which incorporated the structure of the text in the generative and inference processes. Compared it with LDA, it gets fast convergence and good classification and perplexity performance. Sutton *et al.* [26]

presented a topic model named PRODLDA to replace the mixture assumption at the word-level in LDA with a weighted product of experts. ProdLDA can tackle the problems caused by the Dirichlet prior as well as component collapsing. The experimental study was carried on both the 20 Newsgroups and RCV1 Volume 2 datasets with topic coherence and perplexity as performance metrics. Finally, they concluded that ProdLDA can yield much more interpretable topics than standard LDA with collapsed Gibbs.

B. DOCUMENT CLUSTERING

In document clustering objects are documents and the aim is to group clusters when given a document collection in such a way that each document cluster shares more similarity than others. Many researchers have applied document clustering algorithms to text [5], [9]. For example, Nassif presented a document clustering system [27] and carried out the partitioning and the hierarchical approaches with six well-known clustering algorithms on five real-world datasets. The hierarchical clustering algorithms (the average and complete link methods) produced more accurate clustering results than the partitioning clustering algorithms, while the complexity is higher. Kumar and Ravi [28] utilized LDA, term variance, term significance and document frequency methods to select the discriminated features from document term matrix and evaluated the clustering models in terms of precision, recall and F-score. With the combination of LDA with the K-medoids algorithm, the proposed text document clustering model achieved the best F-score values on 20NG and WebKB datasets. Abualigah *et al.* [29] introduced a new feature weighting scheme called Length Feature Weight (LFW) for feature selection and a new Dynamic Dimension Reduction (DDR) method to reduce the number of features used in clustering, followed by K-Means algorithm for document clustering. Experimental results on seven text mining benchmark text datasets showed that the best F-score and accuracy compare to the existing model.

Additionally, the ddCRP (Distance dependent CRPs) algorithm opens the door to a number of further developments in infinite clustering models [30]. Song *et al.* [31] extended ddCRP to a new nested process that was able to simultaneously model the dependencies among data and the relationship between clusters. Li *et al.* [32] introduced side information into the ddCRP using a robust decay function to handle noisy side information. In the field of text topic mining, compared with topic mining algorithms such as LDA, ddCRP algorithm can not only model multiple data dependencies such as time, space and semantics, and does not need to specify the number of clusters [25] in advance.

In summary, Topic models typically involve methods to group both documents as well as words. As for document clustering, topic modeling can be used as a fundamental and enabling tool for efficient document organization by giving a probability distribution over a range of topics for each document. In the following, we integrate topic modeling techniques with clustering algorithms and incorporates popularity information to the calculation of priori distribution in LDA on document clustering. The results of analysis are presented and discussed subsequently.

III. PROPOSED ARCHITECTURE

A. LATENT DIRICHLET ALLOCATION

LDA was first introduced by Blei *et al.* [7] and has been widely used due to its interpretability and flexibility. LDA aims to reveal the hidden semantic structures based on the observed data from a collection of textual documents. A document within the collection is represented as a probability distribution over latent topics and topic distribution share a common Dirichlet prior. A latent topic is characterized by a probability distribution over the words in the vocabulary and word distributions share another common Dirichlet prior [8].

A statistical topic model represents the words in a collection of documents as mixtures of K "topics", words within documents $w_{d,n}\,(\forall n = 1, \ldots, N_d, \forall d = 1, \ldots, M)$ are observed variables while the probabilistic distribution over words of each latent topic $\varphi_k\,(\forall k = 1, \ldots, K)$ with hyper parameter β, the topic distribution per document $\theta_d\,(\forall d = 1, \ldots, M)$ with hyper parameter α and the per-word topic assignment $z_{d,n}$ are hidden variables, K denotes the number of topics, and M denotes the total numbers of documents D in collection, N is the size of the vocabulary.

For each document in the corpus, the words are generated in a two-staged procedure. In the first stage, a distribution over topics is randomly chosen. Based on this distribution, a topic from the distribution over topics is randomly chosen for each word of the document [33]. In the second stage, the hidden unobserved random variables $\varphi_k\,(\forall k = 1, \ldots, K)$ and $\theta_d\,(\forall d = 1, \ldots, M)$ could be learned through Gibbs sampling and variational EM algorithm via maximizing the probability $p\,(D\,|\alpha, \beta\,)$ [34]. Thus, words generative process of LDA model can be described by the joint probability distribution, the likelihood of generating a whole collection is:

$$p(D\,|\alpha, \beta\,) = \prod_{d=1}^{M} \int p(\theta_d\,|\alpha\,) \left(\prod_{n=1}^{N_d} \sum_{z_{d,n}} p\left(z_{d,n}\,|\theta_{\,d}\right) p\left(w_{d,n}\,\middle|z_{d,n}, \beta\,\right) \right) d\theta_d \tag{1}$$

LDA provides two main outputs, namely, the word distribution per topic $\varphi_k\,(\forall k = 1, \ldots, K)$ and the topic distribution per document $\theta_d\,(\forall d = 1, \ldots, M)$.

B. DISTANCE DEPENDENT CRPs

Distance dependent CRPs (ddCRP) [35] is a clustering algorithm that allows certain dependencies between things to be classified. It is described by considering a Chinese restaurant with an infinite number of tables and a sequential process by which customers enter the restaurant and each customer sits

Algorithm 1 Entity-Topic Correlation

Input: The document collection D, distribution of documents on topics θ, distribution of topic on terms φ_k

Output: The Entity-Topic Relevance ξ_k

```
for d_m ∈ D
    E_m = NER(d_m)// extract named entity
    for e_i ∈ E_m
    tf_i = n_{i,j} / (Σ_k n_{k,j})
    idf_i = log(|D|/(|{j : n_i ∈ d_j}| + 1))
    ξ_{i,k} = tf_i^* idf_i^* θ_m^k
    if e_i in EsEntries()
        Update(e_i, ξ_{i,k}, φ_k)
    else
        Insert (e_i, ξ_{i,k}, φ_k)
    end if
  end for
...end for
return ξ_k
```

down at a randomly chosen table [35]. The probability that the i-th customer chooses to sit on the same table as j is shown as follows:

$$p(ta_i = j \,|Dist, \eta) \infty \begin{cases} f\left(dist_{i,j}\right) & i \neq j \\ \eta & i = j \end{cases} \tag{2}$$

where ta_i is the table assignment of i-th customer, $dist_{i,j}$ is the distance metric between the i-th and j-th customers, $Dist$ is the set of distance metrics between all customers, η is a given scaling parameter, f represents the attenuation function, which can adjust the distance dependence between users.

C. THE FRAMEWORK OF tpLDA

The framework of our proposed approach is shown in Fig. 1 and the iteration process of tpLDA in each time slice is illustrated in Fig. 2. We first use LDA model to detect topics, generate the distribution of topics on terms φ and the distribution of documents on topics θ. Then we use the spacy toolkit (https://spacy.io/api/annotation#named-entities) to extract named entities ('PERSON', 'NORP', 'ORG', 'LOC', 'GPE', 'EVENT', 'FAC', 'PRODUCT') from original documents. After that, topic-entity association algorithm introduces named entity to characterize the topic and link the topics among different time slices. Distance dependent Chinese restaurant processes (ddCRP) is utilized to generate topic internal cluster, which gathered by different named entities. Then topic popularity is calculated by both local and global topic popularity. Lastly, topic distribution parameters are updated adaptively, which can be utilized as the prior distributions of the sequent iteration. In the following sections, we present the detailed description of the algorithm.

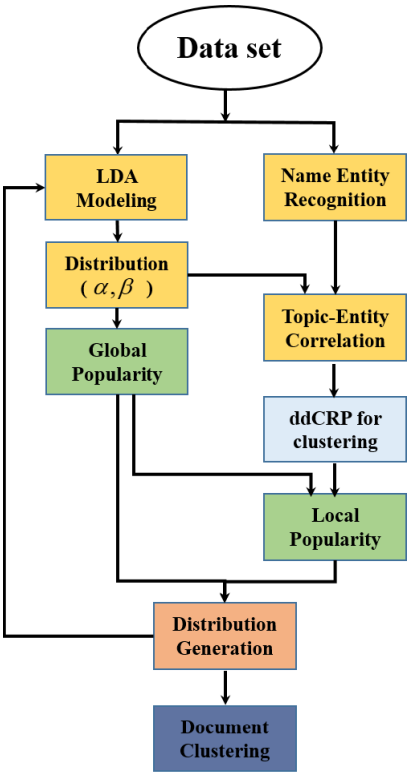

FIGURE 1. The framework of tpLDA.

1) TOPIC AND ENTITY ASSOCIATION

Since the topic distribution is constantly changing while named entities in the topic are generally stable, this section introduces name entity to characterize the LDA topic. Considering that named entity recognition requires the use of documents as a semantic environment, the calculation process of the relevance between topics and entities requires two factors: the distribution probability of the document on the topic and the importance of the entity word in the document. After this process, each entity can be classified to the topic with the highest relevance and the entities that related to one specific topic are clustered. The relevance ($\xi_{i,k}$) between entity i and topic k is expressed in equations (3), (4), (5):

$$tf_i = \frac{n_{i,j}}{\sum_k n_{k,j}} \tag{3}$$

$$idf_i = \log \frac{|D|}{\left|\left\{j : n_i \in d_j\right\}\right| + 1} \tag{4}$$

$$\xi_{i,k} = tf_i^* idf_i^* \theta_m^k \tag{5}$$

where, θ_m^k is the distribution probability of the document m on the topic ξ_k. The importance of an entity term in a document is calculated by TFIDF algorithm [32].

After calculating the entity-topic relevance, records in the entity database should be updated. If an entity already exists in the entity database, the corresponding entity record needs to be updated; otherwise, the topic-entity relevance needs to be inserted first, and then the corresponding named entity

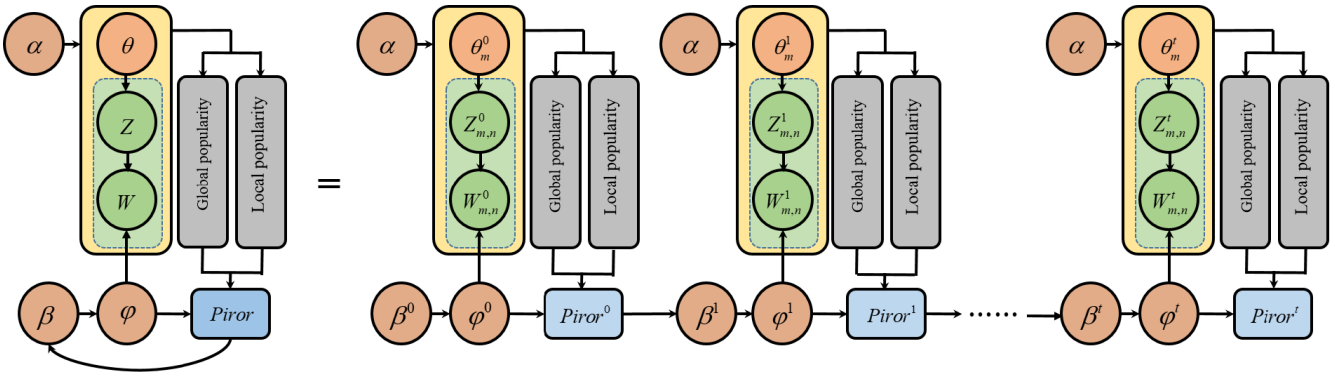

FIGURE 2. Iteration process of tpLDA in one time slice.

record is updated. The detailed process of the algorithm is shown in Algorithm 1.

2) DDCRP ALGORITHM FOR INTERNAL CLUSTERING

The aim of ddCRP algorithm is to cluster entity terms in a topic. According to ddCRP algorithm, we regard each entity term as a customer and the cluster formed by the entity terms as a table. First, we calculate relevance between the entity terms. After extracting named entities, each document has a collection of named entities e_m. Obviously, entities in the same document must have some sort of semantic association. The semantic relevance between entities i and j is calculated as shown in equation (6):

$$r\_entity_{i,j} = \frac{n_{i\cap j}}{n_{i\cup j}} \tag{6}$$

where $n_{i\cap j}$ represents the number of documents containing both entity i and entity j, $n_{i\cup j}$ denotes the number of documents containing entity i or entity j.

Then we divide the associated entity terms as a cluster. Assume that η is scalar parameter, f is the decay function, S is the relevance matrix of all entities and the entity terms of a cluster suffer from the G_0 distribution. The complete probability generation process of the topic-related entity term clustering algorithm based on ddCRP can be described as:

(1) For each entity i, according to semantic relevancy, draw the result of customer distribution $c_i \sim ddCRP(S, \eta, f)$.
(2) The internal cluster class ta_i of entity i is obtained according to the customer assignment result of all entities.
(3) For each cluster k, the entity term it contains $e^* \sim G_0$.
(4) For each entity i, assign it to the final cluster $e_i = e^*_{z_i}$.

3) GLOBAL AND LOCAL POPULARITY CALCULATION

In this section, global topic popularity and local internal cluster popularity are adopted to quantitatively represent the quality and effectiveness of a topic.

The affinity of documents can be measured by information entropy [36], which can be written as:

$$v_m = 1 - \frac{entropy(d_m)}{\max\{entropy(d_1), \ldots, entropy(d_M)\}} \tag{7}$$

$$entropy(d_m) = -\sum_{k=1}^{K} \theta_m^k \log_2 \theta_m^k \tag{8}$$

where θ_m^k indicates the distribution probability of the document m on topic k, K is the number of topics. In general, $entropy(d_m)$ is inversely proportional to v_m, v_m evaluates the concentration of topics in the document m.

The local internal clusters popularity considers two factors: the topic popularity and the ratio of the cluster words in the topic to all the topic words frequencies.

Based on the above analysis, the topic popularity and internal cluster popularity are shown as follows:

$$GTP(k) = \theta^k V \tag{9}$$

$$LTP(\vec{e}_{k,i}) = \frac{word\_fp(\vec{e}_{k,i})}{word\_fp(\vec{e}_k)} GTP(k) \tag{10}$$

where $V = (v_1, v_2, \ldots, v_M)^T$, $v_i (1 \le i \le M)$ measures the concentration of topics, θ^k is a dimensional matrix within K rows and M columns, $GTP(k)$ is the popularity of the k-th topic, $\vec{e}_k$ denotes the related entities according to topic k, $word\_fp(\vec{e}_{k,i})$ is the sum of the word frequencies of the i-th cluster of the k-th topic, $word\_fp(\vec{e}_k)$ is the sum of word frequencies in the k-th topic, $LTP(\vec{e}_{k,i})$ is the popularity value of the i-th cluster of the k-th topic.

4) PARAMETER GENERATION

In this section, we adjust the hyper-parameter β by the global and local topic popularity to improve the LDA model. The parameter generation algorithm is described in Algorithm 2 and the adjustment method of parameter β

is as shown follows:

$$B = \begin{bmatrix} b_{1,1}\ b_{2,1}\ b_{3,1} \ldots b_{k,1} \\ b_{1,2}\ b_{2,2}\ b_{3,2} \ldots b_{k,2} \\ \cdot \\ \cdot \\ b_{1,n}\ b_{2,n}\ b_{3,n} \ldots b_{k,n} \end{bmatrix} \tag{11}$$

$$wtp(w_{j\in n}) = \begin{cases} LTP(w_j), & w_j \in E \\ 1, & w_j \notin E \end{cases} \tag{12}$$

$$WTP(W) = diag(wtp(1), wtp(2), \ldots, wtp(n)) \tag{13}$$

$$ETP = \exp\left(\lambda * \left(GTP_i - GTP_{i-1}\right)/M\right) \tag{14}$$

$$\tilde{\beta} = WTP(W) * B * ETP \tag{15}$$

where n denotes the number of unique words on a corpus, k denotes the number of topics, B that consists of $b_{k,n}$ represents the probability distribution of each topic. The value of $wtp(w)$ denotes the entity popularity, which depends on the word w is a name entity or not. If w is name entity and the entity belongs to the i-th internal cluster of topic k, $wtp(w)$ is the cluster popularity $LTP(w)$. Otherwise, we set the value as 1. $WTP(W)$ is a $n*n$ diagonal matrix and the value of each entry on its main diagonal corresponds to entity popularity $wtp(w)$. ETP measures the variable intensity and direction of the global topic popularity in topic mining process, and i is the iteration step during training, here the maximize value is set as 15. If ETP value is larger than 1, the topic popularity increases, otherwise, it decreases. In each time slice, the enhanced LDA model is trained with 15 iterations. Here, the choice of the number of iterations is based the ETP value. Experiments show that after 15 iterations, ETP is basically stable to 1, reflecting that topic popularity topic of this iteration is the same as the value of the last iteration, which indicates that the topic is stable. ETP can reflect topic popularity and WTP is related to entity popularity. Both of them are related to topic-word distribution, then new $\tilde{\beta}$ can be generated based on the equation (13)∼(15).

5) DOCUMENT CLUSTERING

The outcome of LDA algorithm, document-topic distribution θ, is adopted to represent a document. The topic with the highest probability is selected as the topic label for a document. Since the crawled dataset is divided into 10 data subsets by temporal characteristic, documents in adjacent time slices have semantic and content consistency. The output of the tpLDA of the previous time slice can be utilized as the Dirichlet prior for LDA model at the subsequent time slice. We run tpLDA at each time stage with the output distribution β of the previous tpLDA and the experiment is repeated 10 times with different split datasets.

IV. EXPERIMENT AND ANALYSIS

In this section we describe the dataset extraction and processing step, evaluation measures, experiments results and related analysis.

Algorithm 2 Topic Distribution Iteration

Input: The dataset D, the distribution probability of topics on terms β, and the name entity set E
Output: Topic mining results

// If the inputs don't contain β, use the default β parameter
$\alpha = (0.1, \ldots, 0.1)$
if β *is None*
  $\beta = (0.1, \ldots, 0.1)$
for i = 1 to N_{iter}
  $\theta^i, \varphi^i = GibbsSampling(\alpha, \beta, D)$
  for $d_m \in D$
    Set W_m^i by (8) and (9)
  end for
  for each topic k
  Get ξ_k the correlation of entities and topic k by(4),(5) and (6)
  Get the related entities $\vec{e}_k$ in topic k sorted by ξ_k
//Calculate the internal correlation of the entity term.
// Store the results in a two-dimensional array $S[i][j]$
  $S = NULL$
  for $e_{k,a} \in \vec{e}_k$
    for $e_{k,b} \in \vec{e}_k$
      $S[a][b] = r\_entity_{a,b}$
    end for
  end for
  // get the topic internal cluster by ddcrp
 Get the topic internal cluster C_k from $ddCRP(S, \eta, f)$
    Set global topic popularity of topic k by (9)
    Set local topic popularity of each cluster c in topic k by (10)
    end for
    Set $WTP(w)$ of each word by (12) and (13)
    $ETP^i = \exp\left(\lambda * \left(GTP^i - GTP^{i-1}\right)/M\right)$
    $\tilde{\beta}^i = WTP(W) * B * ETP^i$
  end for
end if
$\tilde{\beta} = \tilde{\beta}^{i=N_{iter}}$
$\theta, \varphi = GibbsSampling(\alpha, \tilde{\beta}, D)$
return $\theta, \varphi, \tilde{\beta}$

A. DATA EXTRACTION AND PREPROCESSING

The dataset used in this study is named as wtArticles and was collected between 13 June and 21 June 2019. It includes 2815 documents of 10 categories from the website www.washingtonpost.com/. The total number of terms is 559149 and the number of unique terms is 21579. The average number of words among all documents is 199, which is large enough to be analyzed by LDA. The categories are: Business, Politics, World, Society, Soccer, Middle east, China & world, Markets, Organizations and Finance. It is available online at https://github.com/xal2019/tpLDA/tree/master/dataset.

Text preprocessing is a very common but important step in document clustering and the aim is to standardize the representation of texts. Data preprocessing commonly includes four steps:

(1) convert to lowercase. All the documents are converted into lowercase to keep the whole data in a uniform format.
(2) remove special characters as they essentially do not contribute any meaningful information to our topic modeling.
(3) tokenize the sentences into terms. The aim of tokenizing the sentences is to obtain informative words, such as nouns and adjectives readily.
(4) remove stop words. Stop words are generally a set of frequently used words and carry no information that should be removed.
(5) stemming the word to its origin. The goal of stemming is to reduce the variation and obtain the lexical root or stem for words.
(6) lemmatization using the WordNet Lemmatizer.
(7) construct term-document matrix. This matrix presents the distribution/frequency of words within documents which is used as the main input to LDA algorithm.

B. EVALUATION MEASURES

Validation measures used for evaluating document clustering algorithms are basically divided into intrinsic and extrinsic measures. Extrinsic measures are based on external information about clusters to evaluate accuracy of clustering while intrinsic measures are used when class labels are not available.

To provide further evidence for the proposed method, we use both intrinsic measures (topic coherence) and extrinsic measures (precision, recall and F-score, Rand index, Adjusted Rand Index (ARI), Jaccard Index (JC) and Fowlkes-Mallows index (FMI) [37]) to validate the quality of the resulted clustering. The values of these measures are between 0 and 1 and higher is better, except ARI, for which the value is between -1 and 1. The extrinsic metrics are according to the following equations:

$$Rand - index = \frac{TP + TN}{TP + FP + TN + FN} \tag{16}$$

$$Jaccard - index = \frac{TP}{TP + FN + FP} \tag{17}$$

$$FMI = \sqrt{\frac{TP}{TP + FP} \cdot \frac{TP}{TP + FN}} \tag{18}$$

$$ARI = \frac{RI - E\,[RI]}{\max(RI) - E\,[RI]} \tag{19}$$

$$\Pr ecision = \frac{TP}{TP + FP} \tag{20}$$

$$\mathrm{Re}\,call = \frac{TP}{TP + FN} \tag{21}$$

$$F - \mathrm{score} = \frac{2TP}{2TP + FP + FN} \tag{22}$$

where TP denotes the number of pairs of documents that the two documents are similar and belong to the same cluster, FN denotes the number of pairs of documents that the two documents are similar and belong to different clusters, FP indicates the number of pairs of documents that the two documents are different and belong to the same cluster and TN indicates the number of pairs of documents that the two documents are different and belong to different clusters.

Besides the above extrinsic measures, we also introduce an intrinsic measure topic coherence [38] to assess the meaningfulness of the underlying topics captured. Topic coherence provides a quantitative measure of the interpretability of a topic [39]. We obtain the coherence by taking the average pointwise mutual information of two words drawn randomly from the same document [40].

C. EXPERIMENT RESULTS AND ANALYSIS

In this section, we conduct experiments on tpLDA model and provide an analysis of the performance of several methods to demonstrate the effectiveness of our proposed approach. In general, four different topic models include standard LDA [7], Dynamic Topic Models (DTM) [41], sentenceLDA [25] and ProdLDA method [26]. The resulted clustering results that validated using different measures are shown in Tables 1-7.

1) PERFORMANCE EVALUATION WITH SIX EXTRINSIC CLUSTERING MEASURES

In this section we first provide three extrinsic measures (FMI, JC and ARI) with the four clustering methods on the wtArticles dataset. It is clear from these tables that our proposed method outperformed the other methods on all three evaluation measures. tpLDA is well ahead of the other four methods and achieves a high FMI at each individual time step with 15% higher FMI than the previous best one and 4% higher FMI at least. Comparatively, for both Jaccard Index and Adjusted Rand Index, overall the performance is much higher than the other four methods. For Jaccard Index (JC) from the overall perspective, tpLDA outperforms best among 8 of the 9 time slices. Our proposed method achieves the second-best value (28.06 %) of Jaccard Index (JC) at the sixth time slice and is slightly lower (1.87%) than the best one. Table 3 show the quality of clusters according to Adjusted Rand Index (ARI). Among all nine time slices, tpLDA is superior than other topic models in seven time slices. Generally, Table 1 to 3 show that the quality of clusters is dramatically improved by introducing topic popularity to topic modeling techniques in the document clustering process based on external clustering measures.

In addition, we report the clustering accuracy based on the recall, precision and F-score to verify the applicability between tpLDA and four previous proposed algorithms.

Since precision and recall are working in opposite direction, F-score is taken as the weighted harmonic mean of precision and recall to reflect the overall clustering ability. From Table 5 and 6, tpLDA shows strong advantages in terms

TABLE 1. FMI of different models on wtArticles dataset.

FMI	LDA [7]	DTM [41]	senLDA [25]	prodLDA [26]	tpLDA
0	17.36	42.28	38.16	25.14	**52.23**
1	21.44	45.75	40.55	33.61	**65.89**
2	18.00	39.50	40.09	25.04	**54.21**
3	16.69	34.02	34.95	27.74	**38.53**
4	17.37	42.82	41.89	35.17	**50.27**
5	17.09	34.40	44.15	31.32	**44.64**
6	16.20	37.54	41.65	28.01	**54.22**
7	16.44	34.57	38.00	37.26	**43.57**
8	19.36	33.92	42.63	36.68	**54.07**

TABLE 2. JC of four different topic models on wtArticles dataset.

JC	LDA [7]	DTM [41]	senLDA [25]	prodLDA [26]	tpLDA
0	9.08	25.83	22.21	13.68	**35.31**
1	10.98	28.96	23.68	18.46	**47.95**
2	9.56	23.73	23.98	13.62	**36.08**
3	9.08	19.95	20.87	15.96	**22.75**
4	9.37	26.75	26.13	21.14	**31.63**
5	9.18	20.48	**28.06**	18.19	26.19
6	8.61	22.65	25.74	15.70	**33.76**
7	8.92	20.88	23.40	22.90	**23.70**
8	10.68	20.32	27.00	22.44	**32.58**

TABLE 3. ARI of different models on wtArticles dataset.

ARI	LDA [7]	DTM [41]	senLDA [25]	prodLDA [26]	tpLDA
0	2.34	29.89	25.52	10.81	**38.88**
1	3.62	31.17	26.05	18.60	**50.38**
2	5.64	27.71	28.49	11.86	**38.72**
3	3.20	23.10	**23.74**	15.13	20.06
4	4.09	32.63	31.44	23.43	**33.31**
5	2.38	22.61	**33.72**	19.35	23.30
6	2.74	25.79	30.59	15.51	**34.00**
7	2.61	24.56	28.80	27.40	22.69
8	1.21	22.09	32.17	24.65	**33.66**

TABLE 4. Recall of different models on wtArticles dataset.

Rec	LDA [7]	DTM [41]	senLDA [25]	prodLDA [26]	tpLDA
0	23.74	**53.90**	52.29	33.89	50.14
1	29.22	55.51	**57.07**	49.86	54.97
2	25.86	50.41	**52.49**	33.75	43.92
3	20.14	**42.08**	40.85	31.55	29.14
4	21.31	**50.72**	48.62	39.83	37.16
5	19.80	40.09	**49.89**	37.69	36.20
6	21.30	44.99	**50.22**	36.06	36.98
7	16.06	35.71	**40.55**	37.59	25.96
8	17.76	37.10	**45.95**	37.89	34.70

of precision and F-score. As for recall, tpLDA is able to provide a comparable clustering results with the enhanced LDA-based models. Taken together, the promising results

TABLE 5. Precision of different models on wtArticles dataset.

Pre	LDA [7]	DTM [41]	senLDA [25]	prodLDA [26]	tpLDA
0	12.92	33.16	27.85	18.65	**54.42**
1	14.60	37.71	28.81	22.66	**78.99**
2	15.23	30.95	30.62	18.57	**66.91**
3	13.71	27.50	29.90	24.40	**50.94**
4	15.19	36.15	36.09	31.06	**68.02**
5	14.10	29.51	39.07	26.02	**80.00**
6	13.65	31.33	34.55	21.75	**79.51**
7	14.78	33.47	35.61	36.93	**73.13**
8	12.13	31.01	39.55	35.51	**84.27**

TABLE 6. F1-score of different models on wtArticles dataset.

F-score	LDA [7]	DTM [41]	senLDA [25]	prodLDA [26]	tpLDA
0	16.73	41.06	36.34	24.06	**52.19**
1	19.47	44.91	38.29	31.15	**64.82**
2	19.17	38.35	38.68	23.96	**53.03**
3	16.31	33.26	34.53	27.52	**37.07**
4	17.74	42.21	41.43	34.90	**48.06**
5	16.47	34.00	43.82	30.79	**49.85**
6	16.64	36.94	40.93	27.13	**50.48**
7	15.39	34.55	37.92	37.26	**38.31**
8	14.42	33.78	42.51	36.66	**49.15**

TABLE 7. Topic coherence (NMPI) scores of different models on wtArticles dataset.

TC	LDA [7]	DTM [41]	senLDA [25]	prodLDA [26]	tpLDA
0	26.42	53.31	52.52	46.00	**54.98**
1	33.74	52.12	51.80	49.94	**59.88**
2	28.45	50.87	52.74	49.63	**60.34**
3	35.74	48.68	51.85	46.57	**60.00**
4	34.43	49.44	53.03	53.32	**57.87**
5	30.05	46.53	53.44	38.28	**49.84**
6	29.25	46.64	**53.03**	35.29	47.38
7	30.28	46.19	50.48	34.92	**52.20**
8	48.73	43.71	53.64	25.70	**56.13**

demonstrate that the topic and topic internal cluster popularity information could help in generating higher clustering accuracy.

2) PERFORMANCE EVALUATION WITH INTRINSIC CLUSTERING MEASURES

Beside of extrinsic measures, we also conduct experiments on a intrinsic clustering measure, topic coherence, to further validate the rationality of the introduction of topic and its internal cluster popularity information for document clustering. Table 7 shows that tpLDA outperforms the other four topic models in terms of topic coherence over eight time slices. The topics extracted by tpLDA appear visually more

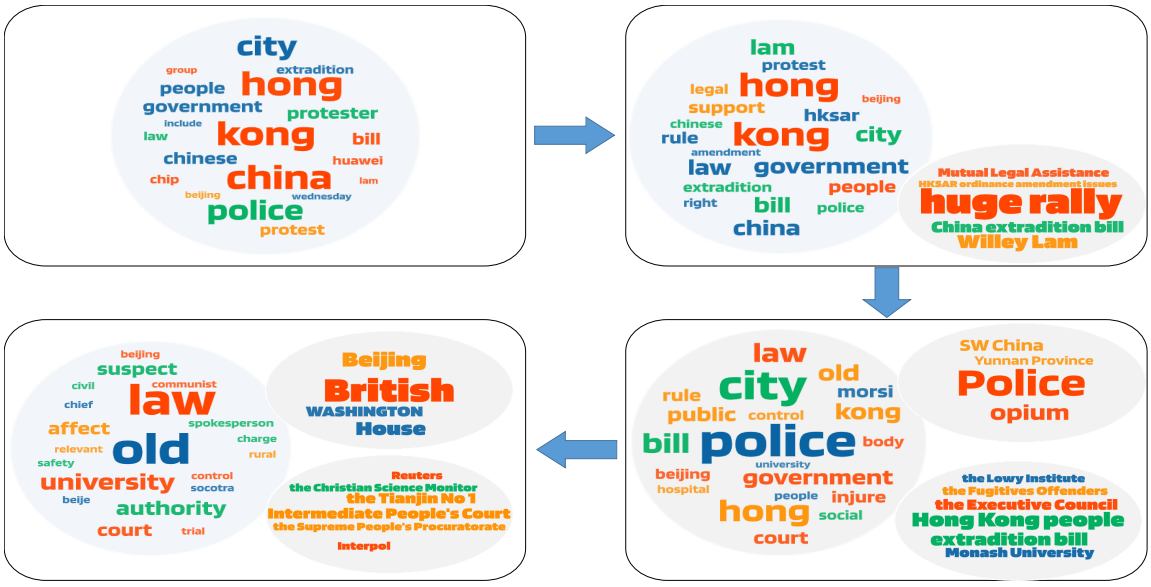

FIGURE 3. An example topic and the main internal clusters over time.

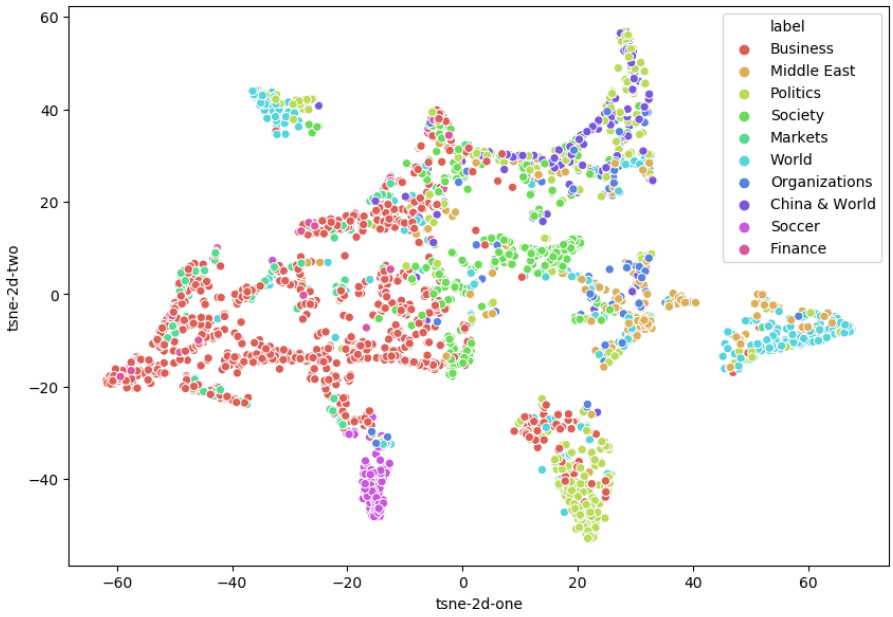

FIGURE 4. T-SNE projection of document on wtArticle dataset.

coherent than others. A little lower topic coherence produced at the seventh time slice might be caused that there do not contain enough documents to accurately approximate the pointwise mutual information. The overall results suggest that our contribution of introdu-cing topic and topic cluster-level popularity into topic model improves topic coherence.

As pointed out by [42], the choice of Dirichlet hyper-parameter is important to the topic quality of LDA. The main advantage of tpLDA topic models is that the incor-poration of topic and cluster-level popularity allows us to design a specific Dirichlet prior. The increased measures in our method reflect that tpLDA can consistently produce better topics, whether measured by automatically determined topic coherence or qualitative examination.

3) TOPIC AND THE CORRESPDING INTERNAL CLUSTERS REPRESENTATION OVER TIME

Each document is assigned with a mixture of topics, hence another benefit of LDA analysis on documents clustering is that it would be possible to illustrate the frequency/popularity of the local and global topics over time. Fig.3 shows the detected topics and the corresponding components of the aggregated internal clusters. The larger ellipse denotes the topics over all documents and the remaining smaller clusters

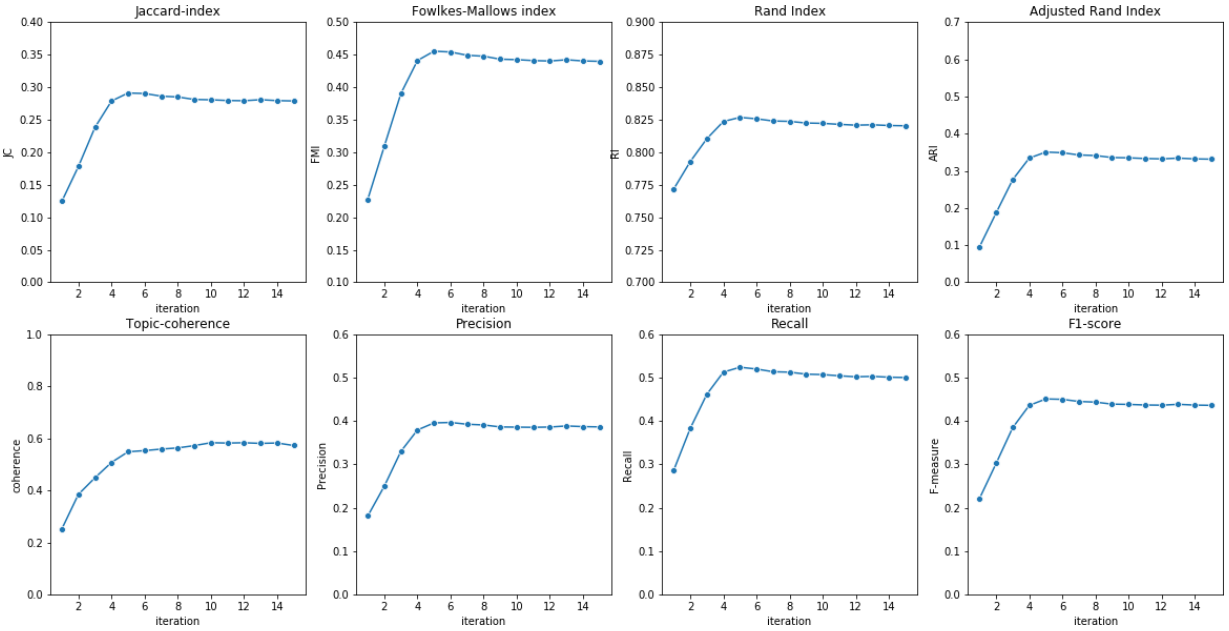

FIGURE 5. Iteration number of tpLDA.

TABLE 8. Topic representation and assignment on wtArticles dataset.

ID	Topics	Label
0	Stock market company price rate share trade investor high buy	Markets
1	China chinese billion million market country bank accord company yuan	Finance
2	Hong kong police bill government city china law lam chinese	Politics
3	Group world team game minute goal win woman player cup	Soccer
4	Iran attack oil iranian saudi tanker state nuclear military gulf	World
5	Trump president house white campaign tell boee american think airline	Organizations
6	China chinese country dprk cooperation state people development party international	China&World
7	People province china city earthquake hit local accord quake area	Society
8	Government israel palestinian force israeli refugee yeman border houthi syrian	Middle east
9	Company service new include report people accord facebook court technology	Business

are fine-grained results after clustering. The upper left part of Fig.3 depicts one cluster and the corresponding aggregated entities at the third time slice. With the evolution of the topic, the topic is divided into two internal clusters at the sixth time slice. Then at the eighth time slice, there still extists two internal clusters while topic components change. The specific content of the topic has changed from ''Hong Kong politics'' to ''World and China affairs''.

4) TOPIC REPRSENTATION AND ASSIGNMENT ON wtArticle DATASET

To further demonstrate the quality of the topics, we use one of the outcomes of LDA algorithm, topic-word distribution β, to present LDA analysis. Topics are distributions over words. The top ten words with the highest probability are derived from posterior distribution and the topic with the highest probability is chosed as the topic label for a document. Extracted topic representations during document clustering process using LDA are depicted in Table 8. The words that included in a topic displays the similarity in each single topic with respect to their probability distribution over words. For example, the words ''Hong kong'', ''police'', ''government'', ''china'', ''law'' in the second time slice explicitly reflect the semantic information of ''politics'', which is the word that with the highest probability and choosed as the label for the document. The validated and labeled topics at each time slice evaluated the ability of the proposed tpLDA for documents clustering.

5) DOCUMENT DISTRIBUTION VISUALIZATION

Documents distribution is presented by t-SNE projection in an intuitive way in Fig.4. Topic distribution is utilized as the document feature in t-SNE method. The t-SNE method is repeated 1000 times on wtArticle

dataset and each color corresponds to one group from the 10 different groups of the dataset. From Fig.4, the words in label of ''Business'', ''politics'', ''Society'', ''World'', ''Chain & World'', ''Soccer'', ''Middle East'' are aggregated together obviously, which demonstrate that tpLDA is appropriate for document clustering.

6) TPLDA UNDER DIFFERENT INTERATIONS

We construct an experiment to research the convergence behavior of LDA. The experiment is performed on wtArticle dataset in terms of the evaluation criteria described above. tpLDA was run for 15 iterations to ensure that the log-likelihood converges. From Fig.5, we can see that the accuracy of documents clustering on wtArticle dataset increases sharply before the iteration number reaches 6 and then tpLDA achieves a stable convergence, which indicates that the most appropriate value for the iteration number should be taken as 6.

V. CONCLUSION

Due to the difficulty to define prior parameters, utilizing probabilistic models for document clustering is a challenging task. In this article, we present a new probabilistic model tpLDA, which leverages global and local topic popularities for document clustering. tpLDA introduces prior-knowledge (topic and its internal cluster popularity) to the construction of LDA hyper-parameters which are always difficult to define. The generated distribution guides the learning of a new generative process that reflects the dynamic changes in the data at successive time slices. Through comparing different performances of several document clustering and topic modeling methods in terms of several measures, we can demonstrate that tpLDA outperformed traditional LDA-based techniques and are comparable for some neural networks. With the ever-growing abundance of online data, tpLDA can serve as an alternative tool for document clustering.

ACKNOWLEDGMENT

The authors would like to thank editors and reviewers for careful reading of the manuscript. The authors would also like to thank the anonymous reviewers for their valuable comments and suggestions, which were helpful for improving the quality of the paper.

REFERENCES

[1] D. M. Blei, ''Hierarchical topic models and the nested Chinese restaurant process,'' in *Proc. 16th Int. Conf. Neural Inf. Process. Syst.*, Whistler, BC, Canada, May 2003, pp. 17–24.

[2] A. Guille, H. Hacid, C. Favre, and D. A. Zighed, ''Information diffusion in online social networks: A survey,'' *ACM SIGMOD Rec.*, vol. 42, no. 2, pp. 17–28, 2013.

[3] E. Sherkat, S. Nourashrafeddin, E. E. Milios, and R. Minghim, ''Interactive document clustering revisited: A visual analytics approach,'' in *Proc. Conf. Human Inf. Interact. Retr. (IUI)*, Tokyo, Japan, Mar. 2018, pp. 281–292.

[4] A. Kelaiaia and H. F. Merouani, ''Clustering with probabilistic topic models on arabic texts,'' in *Modeling Approaches and Algorithms for Advanced Computer Applications*. Cham, Switzerland: Springer, 2013, pp. 65–74.

[5] T. K. Landauer, P. W. Foltz, and D. Laham, ''An introduction to latent semantic analysis,'' *Discourse Process.*, vol. 25, nos. 2–3, pp. 259–284, Jan. 1998.

[6] T. Hofmann, ''Probabilistic latent semantic analysis,'' in *Proc. 15th Conf. Uncertainty Artif. Intell.*, Stockholm, Sweden, Jan. 1999, pp. 289–296.

[7] D. M. Blei, A. Y. Ng, and M. I. Jordan, ''Latent Dirichlet allocation,'' *J. Mach. Learn. Res.*, vol. 3, pp. 993–1022, Mar. 2003.

[8] M. D. Hoffman, D. M. Blei, and F. Bach, ''Online learning for latent Dirichlet allocation,'' in *Proc. 23rd Int. Conf. Neural Inf. Process. Syst.*, Vancouver, BC, Canada, vol. 1, Nov. 2010, pp. 856–864.

[9] G. E. Hinton and R. Salakhutdinov, ''Replicated softmax: An undirected topic model,'' in *Proc. Adv. Neural Inf.Processing Syst.*, Jan. 2009, pp. 1607–1614.

[10] S. Lauly, ''Document neural autoregressive distribution estimation,'' *J. Mach. Learn. Res.*, vol. 18, pp. 1–24, Mar. 2016.

[11] H. Larochelle and S. Lauly, ''A neural autoregressive topic model,'' in *Proc. 25th Int. Conf. Neural Inf. Process. Syst.*, Lake Tahoe, NV, USA, vol. 2, Jan. 2012, pp. 2708–2716.

[12] S. Feuerriegel, A. Ratku, and D. Neumann, ''Analysis of how underlying topics in financial news affect stock prices using latent Dirichlet allocation,'' in *Proc. 49th Hawaii Int. Conf. Syst. Sci. (HICSS)*, Jan. 2016, pp. 1072–1081.

[13] C. S. Ziqiang, L. Yang, L. Wenjie, and L. Heng, ''A novel neural topic model and its supervised extension,'' in *Proc. AAAI Conf. Artif. Intell., 29th AAAI Conf. Artif. Intell.*, Austin, TX, USA, Feb. 2015, pp. 2210–2216.

[14] Y. Bengio, P. Lamblin, D. Popovici, and H. Larochelle, ''Greedy layer-wise training of deep networks,'' in *Proc. 19th Int. Conf. Neural Inf. Process. Syst.*, Vancouver, BC, Canada, Dec. 2006, pp. 153–160.

[15] C. Wartena and R. Brussee, ''Topic detection by clustering keywords,'' in *Proc. 19th Int. Conf. Database Expert Syst. Appl.*, Sep. 2008, pp. 54–58.

[16] L. Hou, J. Li, Z. Wang, J. Tang, P. Zhang, R. Yang, and Q. Zheng, ''NewsMiner: Multifaceted news analysis for event search,'' *Knowl.-Based Syst.*, vol. 76, pp. 17–29, Mar. 2015.

[17] K. H. Lim, S. Karunasekera, and A. Harwood, ''ClusTop: A clustering-based topic modelling algorithm for Twitter using word networks,'' in *Proc. IEEE Int. Conf. Big Data (Big Data)*, Boston, MA, USA, Dec. 2017, pp. 2009–2018.

[18] I. Vulić, W. De Smet, J. Tang, and M.-F. Moens, ''Probabilistic topic modeling in multilingual settings: An overview of its methodology and applications,'' *Inf. Process. Manage.*, vol. 51, no. 1, pp. 111–147, Jan. 2015.

[19] X. Pu, R. Jin, G. Wu, D. Han, and G.-R. Xue, ''Topic modeling in semantic space with keywords,'' in *Proc. 24th ACM Int. Conf. Inf. Knowl. Manage. (CIKM)*, Melbourne, VC, Australia, Oct. 2015, pp. 1141–1150.

[20] T. Lin, W. Tian, Q. Mei, and H. Cheng, ''The dual-sparse topic model: Mining focused topics and focused terms in short text,'' in *Proc. 23rd Int. Conf. World Wide Web (WWW)*, Seoul, South Korea, Apr. 2014, pp. 539–550.

[21] J. Mcauley and J. Leskovec, ''Hidden factors and hidden topics: Understanding rating dimensions with review text,'' in *Proc. 7th ACM Conf. Recommender Syst. (RecSys)*, Hong Kong, Oct. 2013, pp. 165–172.

[22] Y. Song, D. R. Gnyawali, M. K. Srivastava, and E. Asgari, ''In search of precision in absorptive capacity research: A synthesis of the literature and consolidation of findings,'' *J. Manage.*, vol. 44, no. 6, pp. 2343–2374, Jul. 2018.

[23] D. Ramage, D. Hall, R. Nallapati, and C. D. Manning, ''Labeled LDA: A supervised topic model for credit attribution in multi-labeled corpora,'' in *Proc. Conf. Empirical Methods Natural Lang. Process. (EMNLP)*, Singapore, vol. 1, Aug. 2009, pp. 248–256.

[24] F. Lu, B. Shen, J. Lin, and H. Zhang, ''A method of SNS topic models extraction based on self-adaptively LDA modeling,'' in *Proc. 3rd Int. Conf. Intell. System Design Eng. Appl.*, Hong Kong, Jan. 2013, pp. 112–115.

[25] G. Balikas, M.-R. Amini, and M. Clausel, ''On a topic model for sentences,'' in *Proc. 39th Int. ACM SIGIR Conf. Res. Develop. Inf. Retr. (SIGIR)*, Pisa, Italy, Jul. 2016, pp. 921–924.

[26] A. Srivastava and C. Sutton, ''Autoencoding variational inference for topic models,'' *arXiv:1703.01488*, Mar. 2017. [Online]. Available: https://arxiv.org/abs/1703.01488

[27] L. F. Da Cruz Nassif and E. R. Hruschka, ''Document clustering for forensic computing: An approach for improving computer inspection,'' in *Proc. 10th Int. Conf. Mach. Learn. Appl. Workshops*, Honolulu, HI, USA, vol. 1, Dec. 2011, pp. 265–268.

[28] B. S. Kumar and V. Ravi, ''LDA based feature selection for document clustering,'' in *Proc. 10th Annu. ACM India Comput. Conf. (ZZZ)*, Bhopal, India, Nov. 2017, pp. 125–130.

[29] L. M. Abualigah, A. T. Khader, M. A. Al-Betar, and O. A. Alomari, "Text feature selection with a robust weight scheme and dynamic dimension reduction to text document clustering," *Expert Syst. Appl.*, vol. 84, pp. 24–36, Oct. 2017.
[30] C. Yang, L. Xie, and X. Zhou, "Unsupervised broadcast news story segmentation using distance dependent Chinese restaurant processes," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Pisa, Italy, May. 2014, pp. 4062–4066.
[31] J. Song, Y. Huang, X. Qi, Y. Li, F. Li, K. Fu, and T. Huang, "Discovering hierarchical topic evolution in time-stamped documents," *J. Assoc. Inf. Sci. Technol.*, vol. 67, no. 4, pp. 915–927, Apr. 2016.
[32] C. Li, S. Rana, D. Phung, and S. Venkatesh, "Data clustering using side information dependent Chinese restaurant processes," *Knowl. Inf. Syst.*, vol. 47, no. 2, pp. 463–488, May 2016.
[33] D. M. Blei, "Probabilistic topic models," *Commun. ACM*, vol. 55, no. 4, p. 77, Apr. 2012.
[34] Y. Chen, H. Zhang, R. Liu, Z. Ye, and J. Lin, "Experimental explorations on short text topic mining between LDA and NMF based Schemes," *Knowl.-Based Syst.*, vol. 163, pp. 1–13, Jan. 2019.
[35] Z.-Y. Shen, J. Sun, and Y.-D. Shen, "Collective latent Dirichlet allocation," in *Proc. 8th IEEE Int. Conf. Data Mining*, Pisa, Italy, Dec. 2008, pp. 1019–1024.
[36] P. Niu and D. G. Huang, "TF-IDF and rules based automatic extraction of Chinese keywords," *J. Chin. Comput. Syst.*, vol. 37, no. 4, pp. 711–715, Apr. 2016.
[37] E. B. Fowlkes and C. L. Mallows, "A method for comparing two hierarchical clusterings," *J. Amer. Stat. Assoc.*, vol. 78, no. 383, pp. 553–569, Sep. 1983.
[38] M. RÖder, A. Both, and A. Hinneburg, "Exploring the space of topic coherence measures," in *Proc. 8th ACM Int. Conf. Web Search Data Mining (WSDM)*, Shanghai, China, Feb. 2015, pp. 399–408.
[39] D. Mimno, "Optimizing semantic coherence in topic models," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, Edinburgh, Scotland, Jul. 2011, pp. 262–272.
[40] J. H. Lau, D. Newman, and T. Baldwin, "Machine reading tea leaves: automatically evaluating topic coherence and topic model quality," in *Proc. 14th Conf. Eur. Chapter Assoc. Comput. Linguistics*, Gothenburg, Sweden, Apr. 2014, pp. 530–539.
[41] D. M. Blei and J. D. Lafferty, "Dynamic topic models," in *Proc. 23rd Int. Conf. Mach. Learn. (ICML)*, Pittsburgh, PA, USA, Jun. 2006, pp. 113–120.
[42] H. M. Wallach, D. Mimno, and A. McCallum, "Rethinking LDA: Why priors matter," in *Proc. Adv. Neural Inf. Process. Syst.*, Vancouver, BC, Canada, Dec. 2009, pp. 1973–1981.

PENG YANG received the Ph.D. degree from Southeast University, in 2006. He was a Research Scientist with CERN, where he was participated in the alpha magnetic spectrometer (AMS) experiment, from 2007 to 2009, which is led by Nobel Laureate P. S. Ting. He is currently an Associate Professor with the School of Computer Science and Engineering, Southeast University, where he is also the Deputy Director of the Future Network Research Center. His research interests include new-generation Internet architecture, natural language processing, and cyber content governance. He is also a member of the National Technical Committee, Standardization Administration of China.

YU YAO is currently pursuing the Ph.D. degree with the School of Computer Science and Engineering, Southeast University, Nanjing, China. Her research interests include natural language processing and machine learning.

HUAJIAN ZHOU is currently pursuing the M.S. degree with the School of Computer Science and Engineering, Southeast University, Nanjing, China. His research interests include natural language processing and machine learning.