Received 25 July 2022
Revised 8 September 2022
Accepted 16 September 2022

User-concerned actionable hot topic mining: enhancing interpretability via semantic–syntactic association matrix factorization

Linzi Wang
Institute of Automation, Chinese Academy of Sciences, Beijing, China and
School of Artificial Intelligence, University of Chinese Academy of Sciences, Beijing, China
Qiudan Li
Institute of Automation, Chinese Academy of Sciences, Beijing, China
Jingjun David Xu
Department of Information Systems, City University of Hong Kong, Kowloon Tong, Hong Kong, and
Minjie Yuan
Institute of Automation, Chinese Academy of Sciences, Beijing, China and
School of Artificial Intelligence, University of Chinese Academy of Sciences, Beijing, China

Abstract

Purpose – Mining user-concerned actionable and interpretable hot topics will help management departments fully grasp the latest events and make timely decisions. Existing topic models primarily integrate word embedding and matrix decomposition, which only generates keyword-based hot topics with weak interpretability, making it difficult to meet the specific needs of users. Mining phrase-based hot topics with syntactic dependency structure have been proven to model structure information effectively. A key challenge lies in the effective integration of the above information into the hot topic mining process.

Design/methodology/approach – This paper proposes the nonnegative matrix factorization (NMF)-based hot topic mining method, semantics syntax-assisted hot topic model (SSAHM), which combines semantic association and syntactic dependency structure. First, a semantic–syntactic component association matrix is constructed. Then, the matrix is used as a constraint condition to be incorporated into the block coordinate descent (BCD)-based matrix decomposition process. Finally, a hot topic information-driven phrase extraction algorithm is applied to describe hot topics.

This work was partially supported by the National Key Research and Development Program of China (Grant No. 2020AAA0103405), the National Natural Science Foundation of China (Grant No. 62071467, 71621002), the Research Grants at the City University of Hong Kong (Grant No. 7005595, 9680306), and the Strategic Priority Research Program of Chinese Academy of Sciences (Grant No. XDA27030100).

Journal of Electronic Business &
Digital Economics
Vol. 1 No. 1/2, 2022
pp. 50-65
Emerald Publishing Limited
e-ISSN: 2754-4222
p-ISSN: 2754-4214
DOI 10.1108/JEBDE-07-2022-0023

Findings – The efficacy of the developed model is demonstrated on two real-world datasets, and the effects of dependency structure information on different topics are compared. The qualitative examples further explain the application of the method in real scenarios.

Originality/value – Most prior research focuses on keyword-based hot topics. Thus, the literature is advanced by mining phrase-based hot topics with syntactic dependency structure, which can effectively analyze the semantics. The development of syntactic dependency structure considering the combination of word order and part-of-speech (POS) is a step forward as word order, and POS are only separately utilized in the prior literature. Ignoring this synergy may miss important information, such as grammatical structure coherence and logical relations between syntactic components.

Keywords Phrase-based hot topic mining, User-concerned action element, Word embedding, Matrix factorization

Paper type Research paper

1. Introduction

Mining user-concerned actionable and interpretable hot topics will help business managers and government officers fully grasp the latest events and make timely decisions (Zeng, 2015). The keyword-based hot topics cover extensive information but lack detailed descriptions. Meanwhile, the phrase-based hot topics contain action features and express the deep semantics, reflecting stronger interpretability. Taking the topic of “New Energy Vehicles” as an example, the meaning of high-frequency word “vehicles” is broad. Independent hot words, such as “low carbon,” “clean energy” and “carbon cycle,” lack in-depth details and logical association. Thus, understanding and interpreting the deep semantics behind hot topics is difficult for users due to the aforementioned limitations. The phrase-based hot topics “low carbon life desires clean energy technology” and “New Energy Vehicles promote carbon cycle” contain action verb information, thus further explaining and deepening the semantics of the above keywords. Helping to understand that the “New Energy Vehicles” and “low carbon life” are the current hot concerns is convenient, and New Energy Vehicles have a large market demand to meet people’s desire for low-carbon life. Therefore, such interpretable phrase-based hot topics can help companies locate market demands and thus guide their action. They may strengthen the research of technology and increase product promotion to seize the market share of New Energy Vehicles in time. Overall, considering the actual decision-making application requirements and mining the user-concerned hot topics are important tasks.

Most traditional hot topic mining methods generally use latent Dirichlet allocation (LDA) (Blei, Ng, & Jordan, 2003) and nonnegative matrix factorization (NMF) (Kim, He, & Park, 2014) to identify topics. New trends that enhance the semantic analysis capability of matrix decomposition by integrating word vector representation have recently emerged. For example Shi, Kang, Choo, and Reddy (2018), proposed the SeaNMF model, which combined the semantic association of word pairs with the NMF method, effectively improved the performance of topic analysis based on word vector learning and word-pair modeling methods. Compared with the LDA model, this type of method focuses on the semantic association of global words, and the mined hot words have stronger consistency. However, mining deep semantics is difficult due to the lack of structural information, and only keyword-based hot topics can be generated. These hot topic results are limited in interpretability, which can hardly meet the specific needs of users. Syntactic dependency represented by word order and part-of-speech (POS) can effectively model semantics by providing word position and grammatical information (Cheng, Yue, & Song, 2020; Chotirat & Meesad, 2020; Hahn, Jurafsky, & Futrell, 2020; Liu *et al.*, 2021; Nguyen & Nguyen, 2021; Tan, Wang, & Jia, 2020; Zhu, Li, Sun, & Yang, 2020). Moreover, syntactic dependency has been proven to be one of the important features for relation extraction and abstract generation. The existing studies combine probability-based topic models including LDA with syntactic dependency to improve modeling performance. For example, Darling and Song (2013) integrated POS as a probability parameter into the traditional LDA model, which simultaneously mined short-distance grammatical and long-distance topic patterns in the document collection. However, these studies did not combine POS and word order, which will be explained further in Section 2.2.

This paper proposes a hot topic mining method, namely semantics syntax-assisted hot topic model (SSAHM), which embeds word vectors into NMF and integrates with syntactic dependency structure, to generate actionable and interpretable hot topics. First, the semantic association between word pairs and the syntactic dependency co-occurrence relationship are obtained on the basis of the syntactic parse tree and global word frequency statistics. Hence, the semantic–syntactic component association matrix is constructed. This matrix is further treated as a constraint condition and integrated into the block coordinate descent (BCD)-based matrix decomposition process. The hidden vectors of the hot topics are learned in iteration, and similar content clusters and hot keyword descriptions are also obtained. Finally, a hot topic information-driven phrase extraction algorithm is designed. The deep learning model attention long short-term memory (LSTM) (Bahdanau, Cho, & Bengio, 2014) with pretrained parameters is used for semantic encoding, and the maximal marginal relevance (MMR) scores of candidate phrases are calculated on the basis of the semantic space distance to obtain the hot topic representations with rich semantics.

Overall, most prior research focuses on keyword-based hot topics. Thus, the literature is advanced by mining phrase-based hot topics with syntactic dependency structure, which can effectively analyze the semantics. The development of syntactic dependency structure considering the combination of word order and POS is a step forward as word order and POS are only separately utilized in the prior literature. Ignoring this synergy may miss important information, such as grammatical structure coherence and logical relations between syntactic components. The efficacy of the developed model is demonstrated on two real-world datasets, and the effects of dependency structure information on different topics are compared. The qualitative examples further explain the application of the method in real scenarios.

The remainder of the paper is organized as follows. Section 2 first reviews the existing investigations related to the current study. Section 3 formulates the novel task and introduces the structure of the proposed SSAHM. Section 4 shows the quantitative evaluations on two real-world datasets. Section 5 provides two examples as qualitative experiments. Finally, Section 6 concludes the paper and proposes ideas regarding future work.

2. Related work

The current study is related to the following three perspectives: topic modeling, syntactic dependency structure analysis and phrase extraction.

2.1 Topic modeling

The generative probability model and NMF are generally two major groups of topic modeling. Compared with probability-based topic models, such as LDA, the application of NMF can capture the relevant information within the corpus from a global perspective (Bao *et al.*, 2008; Chen *et al.*, 2019; Choo, Lee, Reddy, & Park, 2015; Kim *et al.*, 2015, Kuang, Choo, & Park, 2015, Park, An, Char, & Kim, 2009, Shi *et al.*, 2018). Kim *et al.* (2015) utilized joint NMF for topic modeling to understand large-scale document collections efficiently and find common and discriminative topics simultaneously. Choo *et al.* (2015) proposed the weakly-supervised NMF method by directly combining various forms of prior information, which provided interpretable and flexible results and maintained considerable complexity with standard methods. Kuang *et al.* (2015) proposed a sparse and weakly-supervised NMF model for short text topic modeling by directly factorizing a symmetric term correlation matrix, which is applied to human–computer interaction systems for different scenarios. Shi *et al.* (2018) integrated the semantic representation of words into the NMF framework and proposed the SeaNMF model. This model enriched the associated information of the words and their contexts and alleviated the semantic incompleteness caused by the sparse data of short texts.

2.2 Syntactic dependency structure analysis

Previous work has proven that word order and POS are both analytical perspectives of syntactic structure, which is conducive to mining the deep semantics of texts. On the one hand, topic models that consider word order show strong performance. Jameel, Lam and Bing (2015) were motivated by the capability of word order to capture the semantic fabric of documents and integrated word order structure into a supervised topic model for document classification and retrieval learning, which achieved outstanding performance. On the other hand, POS can also enhance modeling capabilities. Bhowmik, Niu, Savolainen, and Mahmoud (2015) performed POS tagging on the keywords obtained from the LDA model and utilized POS to generate word combination requirements automatically. Mukherjee, Kübler, and Scheutz (2017) introduced the LDA topic model to improve syntactic analysis and found the correlation between words in the topic and POS tags. Hejing (2021) first attempted to integrate POS with semantics in the news reprint scene, and the results showed the feasibility of this idea.

Word order, POS and semantics representation (i.e. word embedding) have been proven to be effective in syntactic dependency structure analysis but are only separately considered, ignoring the combination among them. Systematic work that analyzes syntactic dependency structure in conjunction with such information is currently unavailable. Thus, the relationships among word order, POS and semantics representation (i.e. word embedding) are investigated to fill this gap and enhance the expression of structured information (see Table 1).

	Syntactic		Semantic
Research	POS	Word-order	Word embedding
Jameel *et al.* (2015)		√	
Darling and Song (2013)	√		
Bhowmik *et al.* (2015)	√		
Mukherjee *et al.* (2017)	√		
Shi *et al.* (2018)			√
Hejing (2021)	√		√
The current study	√	√	√

Table 1. Summary of syntactic and semantic dependency applied in topic modeling

2.3 Phrase extraction

The methods for phrase extraction can be categorized into supervised and unsupervised. Compared with the heavy data annotation work, the unsupervised phrase extraction algorithm has stronger practicability. The MMR method proposed by Carbonell and Goldstein (1998) provided a strategy that considers significant and diverse information by calculating MMR scores. The TextRank proposed by Mihalcea and Tarau (2004) was a graph-based phrase extraction method, which treated the document as a graph and considered information recursively drawn from the entire text (graph).

The WordAttracionRank proposed by Wang, Liu, and McDonald (2014) based on the idea of treating text as a graph used word embedding as an external knowledge to guide the generation of new edge weights between words. Bennani-Smires *et al.* (2018) proposed EmbedRank, which represented the document and the candidate phrase as a vector in a high-dimensional space. They also used an improved MMR to calculate the reasonable distance between the candidate phrase and the document to obtain the required phrase cluster according to the ranking.

Different from the above work, the current study focused on mining user-concerned hot topics that are actionable and interpretable based on the needs in real scenarios. The proposed model, namely SSAHM, first integrates word semantic information and syntactic dependency structure, including word order and POS, into the NMF decomposition process, and obtains latent vector representations of diverse hot topic words. The learned parameters and keywords then provide clues for phrase extraction, which helps generate the hot topics with strong interpretability.

3. Method for hot topic mining

3.1 Notations

The frequently used notations in this section are summarized in Table 2.

3.2 Problem formulation

Hot topics usually come from multiple channels, such as news and social platforms. In the real scene, users may provide additional attention to the reports published by news sites or posters with wide influence and high recognition in various channels. The user-concerned hot topic mining aims to conduct deep semantic analysis on the specific event dynamic reports $T = \{T_1, T_2, \ldots, T_N\}$, which originates from the publishers $Pu = \{Pu_1, Pu_2, \ldots, Pu_N\}$, in the above-mentioned channels. Given a (channel, publisher, text) pair corpus $(Ch, Pu, T) = \{(Ch_1, Pu_1, T_1), (Ch_2, Pu_2, T_2), \ldots, (Ch_N, Pu_N, T_N)\}$, the hot topic clustering aims to divide all texts into K clusters $C = \{C_1, C_2, \ldots, C_K\}$ with different hot topic tendencies. Moreover, phrase clusters $L = \{L_1, L_2, \ldots, L_K\}$ with action elements are generated to describe the corresponding hot topic clusters, which have strong interpretability and can effectively meet the needs of users. For example, phrase L_1 is a semantic overview of topic cluster C_1. The text T_i originating from any publisher Pu_i only belongs to a certain cluster C_j, and L_j is the actionable phrase-based hot topics. Among these variables, N and K respectively represent the number of texts and hot topics in the corpus. In addition, M is set as the number of words contained in the corpus V.

Table 2.
Notations used in this paper

Name	Description
S_w	Order-based word co-occurrence matrix
S_p	Syntactic component co-occurrence matrix
S	Word order-based association matrix
A	Word-content matrix
W	Latent matrix of center words
W_c	Latent matrix of contextual words
H	Latent matrix of texts
N	Number of texts in the corpus
M	Number of distinct words in the corpus
K	Number of hot topics in the corpus
P	Number of all word pairs in the corpus
R_+	Non-negative real numbers

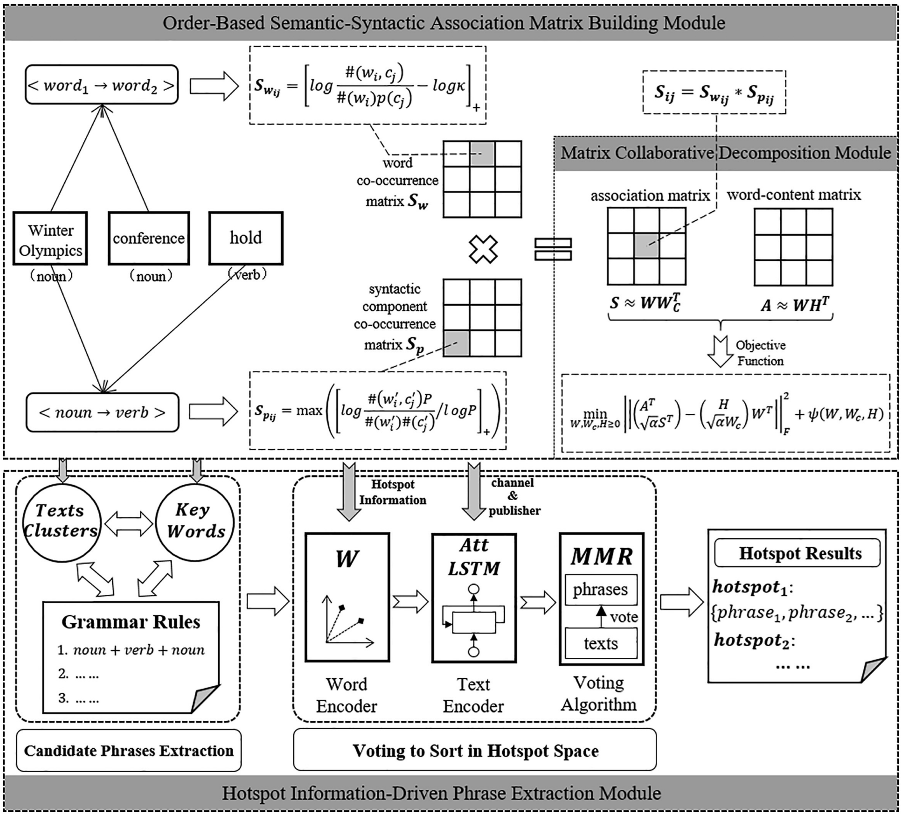
Order-Based Semantic-Syntactic Association Matrix Building Module
$< word_1 \rightarrow word_2 >$
$S_{w_{ij}} = \left[log \frac{\#(w_i, c_j)}{\#(w_i)p(c_j)} - log\kappa \right]_+$
$S_{ij} = S_{w_{ij}} * S_{p_{ij}}$
word co-occurrence matrix S_w
Matrix Collaborative Decomposition Module
association matrix
word-content matrix
Winter Olympics
(noun)
conference
(noun)
hold
(verb)
syntactic component co-occurrence matrix S_p
$S \approx WW_C^T$
$A \approx WH^T$
Objective Function
$< noun \rightarrow verb >$
$S_{p_{ij}} = \max\left(\left[log \frac{\#(w_i', c_j')P}{\#(w_i')\#(c_j')} / l\,ogP\right]_+\right)$
$\min_{W,W_c,H\geq 0} \left\| \begin{pmatrix} A^T \\ \sqrt{\alpha}S^T \end{pmatrix} - \begin{pmatrix} H \\ \sqrt{\alpha}W_c \end{pmatrix} W^T \right\|_F^2 + \psi(W, W_c, H)$
Hotspot Information
channel & publisher
Texts Clusters
Key Words
Grammar Rules
1. noun + verb + noun
2. ……
3. ……
W
Word Encoder
Att LSTM
Text Encoder
MMR
phrases
vote
texts
Voting Algorithm
Hotspot Results
$hotspot_1$:
$\{phrase_1, phrase_2, \ldots\}$
$hotspot_2$:
……
Candidate Phrases Extraction
Voting to Sort in Hotspot Space
Hotspot Information-Driven Phrase Extraction Module

Figure 1.
Framework of the proposed method SSAHM

3.2 Framework of the proposed method

The framework of the proposed method is shown in Figure 1, which contains three modules as a whole, including the order-based semantic–syntactic association matrix building, matrix collaborative decomposition, and hot topic information-driven phrase extraction modules. Specifically, the SSAHM first constructs the order-based word co-occurrence matrix S_w and syntactic component co-occurrence matrix S_p separately, which are both nonnegative and asymmetric. Then above two matrices are multiplied by element positions to construct the word order-based association matrix $S \approx WW_c^T$. Moreover, the model constructs a word-content matrix $A \approx WH^T$ based on all texts in the corpus. Afterward, the matrix S is simultaneously used as a constraint and decomposed with matrix A by the BCD algorithm to obtain the aforementioned matrices, where W and W_c, respectively, denote the latent matrix of center and contextual words, and H is the latent matrix of texts. Furthermore, the cluster set C and hot topic keywords can be obtained from the matrices H and W. Finally, several grammatical rules are formulated in the phrase extraction algorithm to extract candidate phrases. As hot topic information, the matrix W provides the initial embedding representations of the words, which are used by the deep learning model–attention LSTM with pretrained parameters to obtain the semantic vector representations of texts and phrases. The MMR scores of candidate phrases, which simulate the voting process, are calculated by measuring the semantic space distance, and the hot phrase cluster set L is thus generated.

3.3.1 Order-based semantic–syntactic association matrix building module. Syntactic dependency structure helps understand the deep semantics of texts. Verbs generally contain actionable information that is important for users in real scenarios. In addition, order-based POS can find distinguishable popular words and enhance the internal logic of phrase-based hot topics with strong interpretability. However, effectively integrating the above features by designing an order-based semantic-syntactic framework is a key challenge.

The SSAHM respectively constructs the nonnegative and asymmetric order-based word co-occurrence matrix $S_w \in R^{M\times M}$ and syntactic component co-occurrence matrix $S_p \in R^{M\times M}$ according to the frequency of co-occurrence of all ordered word pairs and their semantic components (POS) in the entire corpus. Two aforementioned matrices are fused according to formula (1) by multiplying the corresponding position elements, and the association matrix $S \in R^{M\times M}$ can be easily obtained.

$$S_{ij} = S_{w_{ij}} * S_{p_{ij}} \tag{1}$$

The strategy proposed by Shi *et al.* (2018) is applied to calculate the order-based word co-occurrence matrix S_w. The calculation method is presented among Formulas (2) to (3), which have been proved theoretically feasible by Levy and Goldberg (2014).

$$S_{w_{ij}} = \left[log \frac{\#(w_i, c_j)}{\#(w_i)p(c_j)} - log\,\kappa \right]_+ \tag{2}$$

$$p(c_j) = \frac{\#(c_j)^\gamma}{\sum_{c_j \in V} \#(c_j)^\gamma} \tag{3}$$

where $\#(w_i, c_j)$ indicates the number of occurrences of the ordered word pair (w_i, c_j) in the corpus of texts, and the ordered word pair (w_i, c_j) indicates the appearance of w_i before c_j in the same text. Moreover, $\#(w_i) = \sum_{c_j\in V}\#(w_i, c_j)$ represents the number of word pairs with w_i as the preceding word, and $\#(c_j) = \sum_{w_i\in V}\#(c_j, w_i)$ is the same. Furthermore, the hyperparameters κ and γ, respectively, represent the number of negative samples and the smoothing factor.

The matrix S_p considers the syntactic component information of ordered co-occurring word pairs with the aid of the syntactic analysis tool (Loper & Bird, 2002). Herein, each element $s_{p_{ij}}$ is defined as formula (4).

$$S_{p_{ij}} = \max\left(\left[log \frac{\#(w'_i, c'_j)P}{\#(w'_i)\#(c'_j)} \Big/ logP \right]_+ \right) \tag{4}$$

where $\#(w'_i)$ indicates the number of occurrences of the syntactic component corresponding to word w_i as the preceding word in any word pair, and $\#(c'_j)$ is the same. Moreover, $\#(w'_i, c'_j)$ represents the frequency of the syntactic component pair (w'_i, c'_j), which appears simultaneously in the corpus of texts, and constant P is the number of all word pairs in the corpus V. However, the same word may have multiple syntactic components in different contexts. Thus, the maximum value of $S_{p_{ij}}$ is chosen as the most common syntactic attribute of the word pair. Thus, this value is chosen as the syntactic co-occurrence coefficient for its association matrix.

3.3.2 Matrix collaborative decomposition module. The traditional NMF method (Kim *et al.*, 2014) maps the corpus of texts to the word-content matrix $A \in R_+^{M \times N}$. Each column vector $A_{(:,j)} \in R_+^M$ corresponds to a bag-of-word representation of text j considering N words. The matrix A can be regarded as approximated by two low-rank matrices $W \in R_+^{M \times K}$ and $H \in R_+^{N \times K}$, which can also be formally defined as $A \approx WH^T$. The association matrix $S \approx WW_c^T$ is introduced as a constraint and integrated into the decomposition process of matrix A to use semantic and syntactic information effectively. Therefore, the matrices W, W_c, and H are obtained in the collaborative learning process. The objective function is expressed as formula (5), where $\alpha \in R_+$ is a preset scale factor, and (W, W_c, H) is the penalty function.

$$\min_{W,W_c,H \geq 0} \left\| \begin{pmatrix} A^T \\ \sqrt{\alpha} S^T \end{pmatrix} - \begin{pmatrix} H \\ \sqrt{\alpha} W_c \end{pmatrix} W^T \right\|_F^2 + \psi(W, W_c, H) \tag{5}$$

Finally, the BCD algorithm is incorporated to solve the above formula, and three matrices, namely W, W_c, and H, are updated in accordance with formulas (6) to (8) (Shi *et al.*, 2018) based on random and nonnegative initialization. The update will be iterated repeatedly until the algorithm converges.

$$\mathrm{W}_{(:,k)} \leftarrow \left[W_{(:,k)} + \frac{(AH)_{(:,k)} + \alpha(SW_c)_{(:,k)} - \left(WH^TH\right)_{(:,k)} - \alpha\left(WW_c^TW_c\right)_{(:,k)}}{\left(H^TH\right)_{(k,k)} + \alpha\left(W_c^TW_c\right)_{(k,k)}} \right]_+ \tag{6}$$

$$\mathrm{W}_{c(:,k)} \leftarrow \left[W_{c(:,k)} + \frac{(SW)_{(:,k)} - \left(W_cW^TW\right)_{(:,k)}}{\left(W^TW\right)_{(k,k)}} \right]_+ \tag{7}$$

$$\mathrm{H}_{(:,k)} \leftarrow \left[H_{(:,k)} + \frac{\left(A^TH\right)_{(:,k)} - \left(HW^TW\right)_{(:,k)}}{\left(W^TW\right)_{(k,k)}} \right]_+ \tag{8}$$

Additional detail indicates that the matrices W and H, respectively, associate M words and N texts with K hot topics. For one thing, the column vector $W_{(:,k)} \in R_+^{M \times 1}$ denotes the latent representation of k-th hot topic considering M words, and its elements are the weights of the corresponding words. For another thing, the row vector $H_{(j,:)} \in R_+^{1 \times K}$ is the latent representation of j-th text considering K hot topics, and its elements represent the probability that the text belongs to each hot topic. The time complexity of this module in a single iteration is $O((M + N)MK)$, which shows its applicability in practical scenarios.

3.3.3 Hot topic information-driven phrase extraction module. The mined hot topic information shows the distribution of words in the hot topic space and provides clues of semantic distance for phrase extraction.

The preamble modules provide the k keywords of each hot topic j and the cluster C_j contained therein. As hot topic information and guided by multiple hot topics, the word latent vector representation matrix W is also obtained as a parameter in the iterative process. Thus, the phrase extraction module first constructs phrase grammar rules based on the aforementioned information, which are inspired by Yin & Lina (2017) and are inclined to mine phrases with action elements. As a set of candidate hot phrases, all phrases that match the rules are extracted. Moreover, the matrix W from the hot topic learning process is rich in hidden hot topic information. This matrix initializes the embedding representations of all words and embeds hot topic information into semantic representation, which is used to extract the descriptive phrases. The deep learning model attention–LSTM with pretrained parameters, which are guided by source channel Ch_i and publisher Pu_i corresponding to each text T_i, is applied in the W space to obtain the semantic embedding representations of candidate phrases and texts. Finally, the sub-algorithm MMR, which is based on normalized cosine similarity and combined with voting rules, is used to extract and sort the candidate phrases of each hot topic. A set of hot phrases related to hot topics and diverse in semantics are obtained. The detailed algorithm process is shown in Algorithm 1.

Algorithm 1	**Hot topic information-driven phrase extraction algorithm**
Input:	➢ **Hot topic information:** Word latent vector representation matrix W based on multiple hot topics. ➢ Similar clusters $C = \{C_1, C_2, \ldots, C_K\}$ of different hot topics. ➢ The k keywords and j-th cluster $C_j = \left\{T_1^j, T_2^j, \ldots, T_{\lvert C_j \rvert}^j\right\}$ contained in each of the K hot topics, where $j \in [1, K]$. ➢ The source channel Ch_i and publisher Pu_i corresponding to each text T_i, where $i \in [1, N]$.
Output:	Hot phrase cluster set $\{L_1, L_2, \ldots, L_K\}$ for all K hot topics, where L_j is the hot phrase cluster extracted from the cluster C_j.
Step1:	Build phrase grammar rules: ➢ (Adjective) + noun/noun phrase ➢ (Adjective) + noun/noun phrase + (adverb) + intransitive verb ➢ (Adjective) + noun/noun phrase + (adverb) + transitive verb + (adjective) + noun/noun phrase
Step2:	Select the j-th hot topic. According to the POS information of k keywords, all phrases that meet the grammatical rules are extracted to form a set of candidate hot phrases $R^j = \{s_1^j, s_2^j, \ldots, s_{\lvert R^j \rvert}^j\}$. If the set $R^j = \emptyset$, then go to Step3, otherwise, go to Step4.
Step3:	Locate the positions of k keywords of hot topic in each cluster C_j, and then extract clauses containing keywords and match grammatical rules to form a set of candidate hot phrases $R^j \neq \emptyset$.
Step4:	In the matrix W space, the row vectors of W are the initial embedding representations of the words, which successfully integrate hot information into semantic representation. Attention–LSTM with pre-trained parameters, which are guided by source channel Ch and publisher Pu, is applied to respectively encode the texts and candidate phrases to obtain their semantic vector embedding representations.
Step5:	Select the i-th text T_i^j in the cluster C_j. Use the sub-algorithm MMR based on normalized cosine similarity to select t hot phrases from the set of candidate hot phrases R^j. The selection process means that these t hot phrases get votes for T_i^j.
Step6:	$i \leftarrow i + 1$. Repeat Step5 until the traversal of the cluster C_j ends, indicating that all texts in C_j have completed the voting. Sort the candidate hot phrases according to the total number of votes for each phrase. The ordered phrase description cluster L_j of j-th hot topic, in which phrases with additional votes indicate an effective fit with the hot topic, is thus generated.
Step7:	$j \leftarrow j + 1$. Repeat Step2–Step6 until the traversal of clusters C ends. Finally, the hot phrase cluster set $\{L_1, L_2, \ldots, L_K\}$ of all hot topics can be obtained.

4. Experimental analysis

The performance of the proposed SSAHM is evaluated on two real-world datasets in this section. First, the description of real-world datasets is presented. Then, the baseline methods, evaluation metrics and parameter settings are introduced. Finally, the experimental results on two real-world datasets are analyzed and discussed.

4.1 Dataset description

Chinese news related to the "New Energy Vehicles" and the "Big Data Industry Expo" are collected from 20 news sites. More detailed descriptions of the datasets including the data collection time range, the number of texts, distinct words and different syntactic elements (POS) are listed in Table 3.

4.2 Baseline methods

We utilize NMF (Kim *et al.*, 2014) and SeaNMF (Shi *et al.*, 2018) as baseline methods. Specifically, the first three methods are classic and representative in topic modeling and tend to present good results in some scenarios. SeaNMF integrates word embedding into NMF, which is able to verify the effectiveness of word embedding for hot topic mining.

NMF (Kim *et al.*, 2014): This model divides the hot topics of texts by decomposing the nonnegative document-term matrix into document-topic and topic-term matrices.

SeaNMF (Shi *et al.*, 2018): This model is based on global word co-occurrence modeling to divide the hot topics.

SSAHM-POS: A variant of SSAHM, which only considers POS to model syntactic dependency structure, ignoring word order information.

4.3 Parameter settings

The number of top keywords for each hot topic is set as $k \in [3, 4, 5, 6]$. The number of hot topics considering the different influences of various events is assumed to be $K_O = 7$ per day for the New Energy Vehicles data, but $K_E = 4$ is set for the Big Data Industry Expo. Moreover, the hyperparameters γ, κ and α are respectively initialized to 1, 1 and 0.1, and the penalty function $\psi(W, W_c, H)$ is set to 0.

4.4 Evaluation metrics

Pointwise mutual information (PMI) (Röder, Both, & Hinneburg, 2015) and hot topic quality (HQ) are adopted as the evaluation metrics to access the interpretability and user acceptance from word and phrase-based hot topics, respectively.

The PMI evaluates the coherence score S_k of each hot topic with top k keywords by formula (9), which is the basis for enhancing the interpretability of the composed phrase.

$$S_k = \frac{2}{k(k-1)} \sum_{i=1}^{k-1} \sum_{j=i+1}^{k} \log \frac{p(w_i, w_j)}{p(w_i)p(w_j)} \tag{9}$$

	Time range	Number of texts	Number of distinct words	Number of different POS
New Energy Vehicles	2022.02.19–2022.02.23	14,883	6,164	44
Big Data Industry Expo	2021.05.25–2021.06.03	26,905	6,744	45

Table 3. Detailed descriptions of the datasets

where $p(w)$ represents the probability that the word w appears in any text, and $p(w_i, w_j)$ is the probability of two words w_i and w_j in the same text. Finally, the average coherence score over all hot topics works as the PMI score to evaluate models.

Furthermore, the evaluation metric of HQ, which represents the clustering quality of similar text clusters corresponding to each hot topic, is designed to measure the user acceptance of mined phrase-based hot topics in practical applications. The annotator will give a high score to indicate a high-quality clustering result when the description of the phrase-based hot topics and their texts are intuitively consistent. Three annotators are invited to score the hot topic clustering results, and the average score is scaled to the [0,1].

4.5 Experimental results and discussions

4.5.1 Coherence results of word-based hot topics. Table 4 shows the coherence score PMI of different methods on the two datasets. The bold font and the underline are, respectively, used to highlight the best performance and second-best values.

For New Energy Vehicles data, the SSAHM and SSAHM-POS outperform other traditional methods considering syntactic dependency structure. Compared with SeaNMF, these methods have the most significant performance improvement considering PMI when k is 3, and the value increases from 1.1105 to 1.9846 and 1.7344. This result demonstrates the effectiveness of structure information. Among the methods without syntactic dependency, SeaNMF achieves the best performance in the case of the top four hot words, where the PMI value ranges from 0.0421 to 1.0638, indicating that the global semantic association can discover additional co-occurrences between words and improve the performance. Compared with SSAHM-POS ignoring order information, the SSAHM performs effectively considering order POS. The PMI value varies from 0.8402 to 1.1570 when k is 5. This comparison further implies that the synergy between word order and POS is crucial in user-concerned hot topic mining.

Compared with the methods without syntactic dependency structure, the proposed SSAHM and SSAHM-POS for Big Data Industry Expo data achieve improved performance. Compared with SeaNMF, the methods increased the PMI value from 0.0546 to 1.8897 and 1.3694 when k is 3 and obtained the maximum improvement. For the traditional methods, SeaNMF performs better than others, where the PMI value among the top four hot words varies from −1.0033 to −0.2319, proving the importance of global word association for coherence. For the methods with structure information, SSAHM outperforms SSAHM-POS, demonstrating that the combination of word order and POS can effectively construct the syntactic structure and contribute to user-concerned hot topic mining.

The above analysis reveals that the proposed methods with structure information achieve the best performance. Furthermore, these methods have different adaptability in various topic datasets. Benchmarking with SeaNMF, the performance of our methods applied to the Big Data Industry Expo data has a more significant improvement than the New Energy

Table 4.
PMI score comparisons of word-based hot topic results

Dataset	Method	$k = 3$	$k = 4$	$k = 5$	$k = 6$
New Energy Vehicles	NMF	0.4100	0.0421	−0.2634	−0.4989
	SeaNMF	1.1105	1.0638	0.3571	0.1631
	SSAHM	*1.9846*	*1.5971*	*1.1570*	*0.7077*
	SSAHM-POS	*1.7344*	*1.4937*	*0.8402*	*0.5516*
Big Data Industry Expo	NMF	−0.6051	−1.0033	−0.8690	−1.0479
	SeaNMF	0.0546	−0.2319	−0.4932	−0.9210
	SSAHM	*1.8897*	*1.2290*	*0.6344*	*−0.2228*
	SSAHM-POS	*1.3694*	*1.1186*	*0.6098*	*−0.4036*

Note(s): The large-signed PMI value verifies the large coherence between top k keywords

Vehicles data with simple syntactic forms. This comparative result shows that our proposed methods have the ability to process data with complex and diverse syntactic forms, like Big Data Industry Expo, by modeling structured information.

4.5.2 Practical application results of phrase-based hot topics. The same module is leveraged as the SSAHM to extract phrases and further measure the feasibility of hot topic mining results in practical applications. The top two in the ordered phrase cluster are used as the hot topic results, and the HQ score comparisons of various methods are listed in Table 5. The best performance and second-best values are highlighted with bold font and underline, respectively.

The proposed SSAHM and SSAHM-POS with syntactic dependency structures demonstrate the best performance among all models, proving that syntactical structure information helps enhance the quality and user acceptance of phrase-based hot topics. Compared with the performance on the Big Data Industry Expo data, the models achieve slightly improved performance on the New Energy Vehicles data. The analysis result is due to the slightly difficult accurate summarization of the hot topic phrases caused by the wide distribution of hot topics with semantic differences in the Big Data Industry Expo data. However, the New Energy Vehicles data with concentrated hot topics help the mined phrases in easily describing their semantics.

5. Qualitative experiments

5.1 Case 1: the comparison of hot topic results between different models

Figure 2 lists the representative hot words mined from the New Energy Vehicles dataset by SeaNMF and SSAHM. The figure reveals that high-frequency words in specific fields, such as “New Energy” and “Vehicles”, are common in various hot topic categories. Thus, these words can be easily identified by the above models. Compared with SeaNMF, SSAHM can also

	New Energy Vehicles	Big Data Industry Expo
NMF	0.6967	0.6767
SeaNMF	0.7533	0.7300
SSAHM	*0.7800*	*0.7667*
SSAHM-POS	*0.7733*	*0.7533*

Table 5.
HQ score comparisons of phrase-based hot topic results

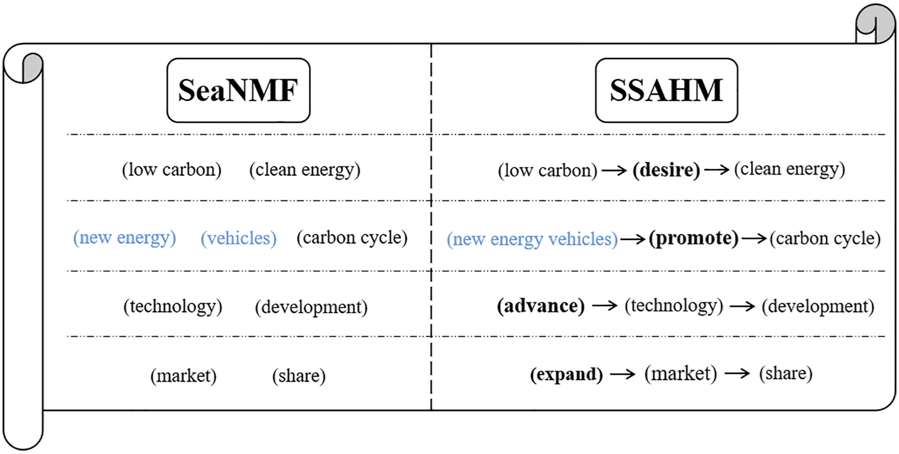
SeaNMF
SSAHM
(low carbon) (clean energy)
(low carbon) → (desire) → (clean energy)
(new energy) (vehicles) (carbon cycle)
(new energy vehicles) → (promote) → (carbon cycle)
(technology) (development)
(advance) → (technology) → (development)
(market) (share)
(expand) → (market) → (share)

Figure 2.
Representative hot words mined from the New Energy Vehicles dataset by two various methods

obtain action-specific words, such as “desire” and “promote”, by considering syntactic structure features, such as POS. These words substantially boost the extraction of actionable phrases and lay the foundation for enhancing the interpretability of hot results.

5.2 Case 2: the evolution of hot topics

Accurately mining the hot topics in each period can track the hot topic evolution trend of the event. The evolution reflects the popularity, and the changes in the public opinion focus on different periods. It can provide information support for relevant management departments to strengthen supervision and formulate policies.

As shown in Figure 3, the proposed method has discovered the hot topics of the “Big Data Industry Expo” in five different development periods, namely “Industry Expo is about to begin”, “Industry Expo is about to begin”, “fantastic technology lights up the Industry Expo”, “Industry Expo closes perfectly” and “outstanding results review the Industry Expo”. For example, the hot discussion during propaganda and preparation stage mainly includes setting up the scene and welcoming the attendees, like “Brilliant lighting creates a strong atmosphere for the Expo” and “Welcome to 2021 Industry Expo.” Besides, during summary and feedback stage, outstanding results are displayed to review the event. Representative hot topics include “Fruitful results of the 2021 Industry Expo Investor Conference,” “The contracted value of the 2021 Industry Expo exceeded 50 billion yuan” and so on. The rich evolutionary trend may help management department fully grasp the real-time hot topics.

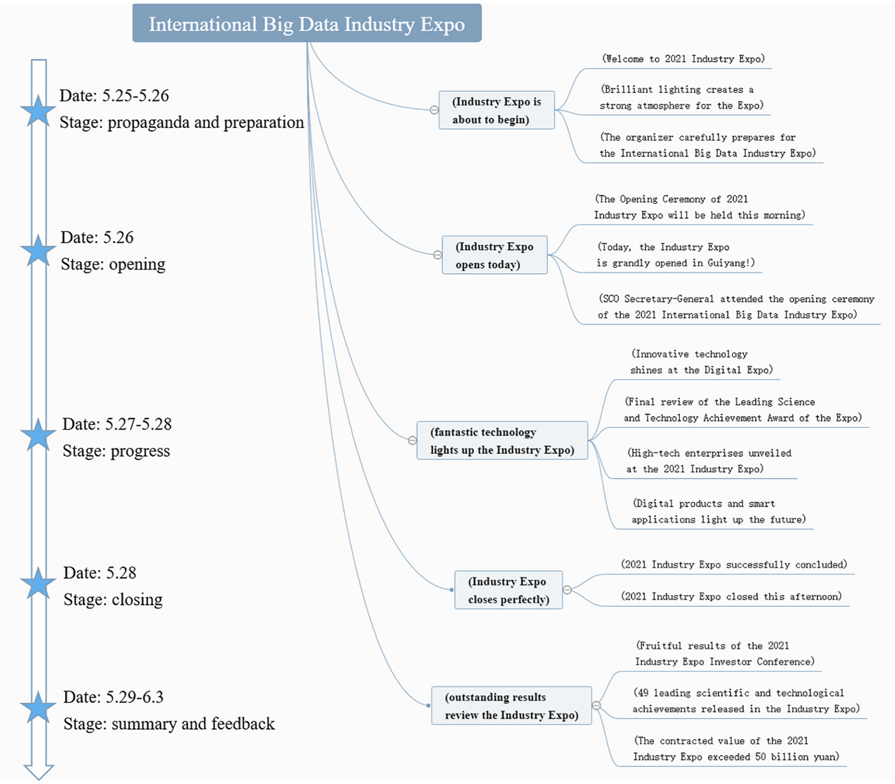
International Big Data Industry Expo
Date: 5.25-5.26
Stage: propaganda and preparation
Date: 5.26
Stage: opening
Date: 5.27-5.28
Stage: progress
Date: 5.28
Stage: closing
Date: 5.29-6.3
Stage: summary and feedback
(Industry Expo is about to begin)
(Welcome to 2021 Industry Expo)
(Brilliant lighting creates a strong atmosphere for the Expo)
(The organizer carefully prepares for the International Big Data Industry Expo)
(Industry Expo opens today)
(The Opening Ceremony of 2021 Industry Expo will be held this morning)
(Today, the Industry Expo is grandly opened in Guiyang!)
(SCO Secretary-General attended the opening ceremony of the 2021 International Big Data Industry Expo)
(fantastic technology lights up the Industry Expo)
(Innovative technology shines at the Digital Expo)
(Final review of the Leading Science and Technology Achievement Award of the Expo)
(High-tech enterprises unveiled at the 2021 Industry Expo)
(Digital products and smart applications light up the future)
(Industry Expo closes perfectly)
(2021 Industry Expo successfully concluded)
(2021 Industry Expo closed this afternoon)
(outstanding results review the Industry Expo)
(Fruitful results of the 2021 Industry Expo Investor Conference)
(49 leading scientific and technological achievements released in the Industry Expo)
(The contracted value of the 2021 Industry Expo exceeded 50 billion yuan)

Figure 3.
Hot topic evolution of the big data industry expo

6. Conclusion and future work

The SSAHM method is developed in this paper to mine user-concerned phrase-based hot topics with action elements. The semantic expression of these hot topics with interpretability can meet the specific needs of users. Relevant management departments can also benefit from this task and fully grasp the latest developments of events to make decisions accurately and timely. The SSAHM simultaneously integrates word semantic association and syntactic dependency structure, including word order and POS, based on the NMF framework. The phrase extraction algorithm driven by hot topic information uses the deep learning model attention–LSTM for semantic encoding and obtains phrase-based hot topics containing action elements that are fused into the method. The experimental results on two constructed datasets prove the effectiveness and practicability of the proposed method. Two qualitative experiments are further conducted to demonstrate the performance of the model and the role of hot topic mining in practical applications. Future works will be devoted to conceiving a hot topic prediction framework, which can accurately and timely predict upcoming hot topics. Furthermore, we will investigate matrix sparsity to further reduce the time complexity of the model, thereby enhancing application performance in practical scenarios. The companies and management departments can take precautions for emergencies and make informed decisions based on accurate predictions to maximize benefits.

References

Bahdanau, D., Cho, K. H., & Bengio, Y. (2014). Neural machine translation by jointly learning to align and translate. *ArXiv E-Prints*. Available from: https://arxiv.org/abs/1409.0473.

Bao, L., Tang, S., Li, J., Zhang, Y., & Ye, W.-P. (2008). Document clustering based on spectral clustering and non-negative matrix factorization. *International Conference on Industrial, Engineering and Other Applications of Applied Intelligent Systems* (pp. 149-158). doi: 10.1007/978-3-540-69052-8_16.

Bennani-Smires, K., Musat, C.-C., Hossmann, A., Baeriswyl, M., & Jaggi, M. (2018). Simple unsupervised keyphrase extraction using sentence embeddings. *Proceedings of the 22nd Conference on Computational Natural Language Learning* (pp. 221-229). Available from: https://infoscience.epfl.ch/record/255278.

Bhowmik, T., Niu, N., Savolainen, J., & Mahmoud, A. (2015). Leveraging topic modeling and part-of-speech tagging to support combinational creativity in requirements engineering. *Requirements Engineering*, *20*(3), 253-280. doi: 10.1007/s00766-015-0226-2.

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *The Journal of Machine Learning Research*, *3*, 993-1022. doi: 10.5555/944919.944937.

Carbonell, J., & Goldstein, J. (1998). The use of MMR, diversity-based reranking for reordering documents and producing summaries. In *Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval* (pp. 335-336). doi: 10.1145/290941.291025.

Chen, Y., Wu, J., Lin, J., Liu, R., Zhang, H., & Ye, Z. (2019). Affinity regularized non-negative matrix factorization for lifelong topic modeling. *IEEE Transactions on Knowledge and Data Engineering*, *32*(7), 1249-1262. doi: 10.1109/TKDE.2019.2904687.

Cheng, K., Yue, Y., & Song, Z. (2020). Sentiment classification based on part-of-speech and self-attention mechanism. *IEEE Access*, *8*, 16387-16396. doi: 10.1109/ACCESS.2020.2967103.

Choo, J., Lee, C., Reddy, C. K., & Park, H. (2015). Weakly supervised nonnegative matrix factorization for user-driven clustering. *Data Mining and Knowledge Discovery*, *29*(6), 1598-1621. doi: 10.1007/s10618-014-0384-8.

Chotirat, S., & Meesad, P. (2020). Effects of part-of-speech on Thai sentence classification to wh-question categories using machine learning approach. *Proceedings of the 11th International Conference on Advances in Information Technology* (pp. 1-5). doi: 10.1145/3406601.3406648.

Darling, W. M., & Song, F. (2013). Probabilistic topic and syntax modeling with part-of-speech LDA. *ArXiv E-Prints*. Available from: https://arxiv.org/abs/1303.2826.

Hahn, M., Jurafsky, D., & Futrell, R. (2020). Universals of word order reflect optimization of grammars for efficient communication. *Proceedings of the National Academy of Sciences*, *117*(5), 2347-2353. doi: 10.1073/pnas.1910923117.

Hejing, L. (2021). Analyzing media reprint effect based on multi-source data. University of Chinese Academy of Sciences.

Jameel, S., Lam, W., & Bing, L. (2015). Supervised topic models with word order structure for document classification and retrieval learning. *Information Retrieval Journal*, *18*(4), 283-330. doi: 10.1007/s10791-015-9254-2.

Kim, J., He, Y., & Park, H. (2014). Algorithms for nonnegative matrix and tensor factorizations: a unified view based on block coordinate descent framework. *Journal of Global Optimization*, *58*(2), 285-319. doi: 10.1007/s10898-013-0035-4.

Kim, H., Choo, J., Kim, J., Reddy, C. K., & Park, H. (2015). Simultaneous discovery of common and discriminative topics via joint nonnegative matrix factorization. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 567-576). doi: 10.1145/2783258.2783338.

Kuang, D., Choo, J., & Park, H. (2015). Nonnegative matrix factorization for interactive topic modeling and document clustering. *Partitional Clustering Algorithms*. Cham: Springer.

Levy, O., & Goldberg, Y. (2014). Neural word embedding as implicit matrix factorization. In *Proceedings of the 27th International Conference on Neural Information Processing Systems* (pp. 2177-2185).

Liu, Z., Winata, G.I., Cahyawijaya, S., Madotto, A., Lin, Z., & Fung, P. (2021). On the importance of word order information in cross-lingual sequence labeling. *Proceedings of the AAAI Conference on Artificial Intelligence* (pp. 13461-13469). Available from: https://ojs.aaai.org/index.php/AAAI/article/view/17588.

Loper, E., & Bird, S. (2002). NLTK: the natural language toolkit. In *Proceedings of the ACL-02 Workshop on Effective tools and methodologies for teaching natural language processing and computational linguistics* (pp. 63-70). doi: 10.3115/1118108.1118117.

Mihalcea, R., & Tarau, P. (2004). Textrank: Bringing order into text. In *Proceedings of the 2004 conference on empirical methods in natural language processing* (pp. 404-411). Available from: https://digital.library.unt.edu/ark:/67531/metadc30962/.

Mukherjee, A., Kübler, S., & Scheutz, M. (2017). Creating POS tagging and dependency parsing experts via topic modeling. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics* (pp. 347-355), available from: https://aclanthology.org/E17-1033/.

Nguyen, L. T., & Nguyen, D. Q. (2021). PhoNLP: a joint multi-task learning model for Vietnamese part-of-speech tagging, named entity recognition and dependency parsing. *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Demonstrations* (pp. 1-7). doi: 10.18653/v1/2021.naacl-demos.1.

Park, S., An, D. U., Char, B., & Kim, C.-W. (2009). Document clustering with cluster refinement and non-negative matrix factorization. *International Conference on Neural Information Processing* (pp. 281-288). doi: 10.1007/978-3-642-10684-2_31.

Röder, M., Both, A., & Hinneburg, A. (2015). Exploring the space of topic coherence measures. *Proceedings of the eighth ACM international conference on Web search and data mining* (pp. 399-408). doi: 10.1145/2684822.2685324.

Shi, T., Kang, K., Choo, J., & Reddy, C. K. (2018). Short-text topic modeling via non-negative matrix factorization enriched with local word-context correlations. *Proceedings of the 2018 World Wide Web Conference* (pp. 1105-1114). doi: 10.1145/3178876.3186009.

Tan, Y., Wang, X., & Jia, T. (2020). From syntactic structure to semantic relationship: hypernym extraction from definitions by recurrent neural networks using the part of speech information. *International Semantic Web Conference* (pp. 529-546). doi: 10.1007/978-3-030-62419-4_30.

Wang, R., Liu, W., & McDonald, C. (2014). Corpus-independent generic keyphrase extraction using word embedding vectors. *Software Engineering Research Conference* (pp. 1-8).

Yin, K., & Lina, Z. (2017). RubE: rule-based methods for extracting product features from online consumer reviews. *Information and Management*, *54*(2), 166-176. doi: 10.1016/j.im.2016.05.007.

Zeng, D. (2015). Crystal Balls, statistics, Big data, and psychohistory: predictive analytics and beyond. *IEEE Intelligent Systems*, *30*(02), 2-4. doi: 10.1109/MIS.2015.24.

Zhu, M., Li, H., Sun, X., & Yang, Z. (2020). BLAC: a named entity recognition model incorporating part-of-speech attention in irregular short text. *2020 IEEE International Conference on Real-time Computing and Robotics (RCAR)* (pp. 56-61). doi: 10.1109/RCAR49640.2020.9303256.

Corresponding author
Qiudan Li can be contacted at: qiudan.li@ia.ac.cn