Topic-Feature Lattices Construction and Visualization for Dynamic Topic Number

Kai WANG

Bengbu Medical College, Bengbu 233000, China
E-mail: wangkai0552@126.com

Fuzhi WANG

Bengbu Medical College, Bengbu 233000, China
E-mail: w fz.bbmc@foxmail.com

Abstract The topic recognition for dynamic topic number can realize the dynamic update of super parameters, and obtain the probability distribution of dynamic topics in time dimension, which helps to clear the understanding and tracking of convection text data. However, the current topic recognition model tends to be based on a fixed number of topics K and lacks multi-granularity analysis of subject knowledge. Therefore, it is impossible to deeply perceive the dynamic change of the topic in the time series. By introducing a novel approach on the basis of Infinite Latent Dirichlet allocation model, a topic feature lattice under the dynamic topic number is constructed. In the model, documents, topics and vocabularies are jointly modeled to generate two probability distribution matrices: Documents-topics and topic-feature words. Afterwards, the association intensity is computed between the topic and its feature vocabulary to establish the topic formal context matrix. Finally, the topic feature is induced according to the formal concept analysis (FCA) theory. The topic feature lattice under dynamic topic number (TFLDTN) model is validated on the real dataset by comparing with the mainstream methods. Experiments show that this model is more in line with actual needs, and achieves better results in semi-automatic modeling of topic visualization analysis.

Keywords dynamic topic number; infinite latent Dirichlet allocation (ILDA); formal concept analysis; topic feature lattice; topic feature lattice under dynamic topic number (TFLDTN) model

1 Introduction

With the widespread application of Web 2.0, self-media platforms, such as online forums and online communities, have gradually become the main form of information exchange. While users enjoy convenient technology, they also face difficulties in making decisions caused by explosive review data. The topic modeling of the review dataset can realize the “short description” of the document, thus providing the possibility for mining the hidden semantic structure of large-scale datasets. However, in the process of topic recognition and evolution, the dynamic change of the number of topics makes it difficult to quantitatively analyze the relationship between

Received January 22, 2020, accepted July 24, 2020

Supported by the Key Projects of Social Sciences of Anhui Provincial Department of Education (SK2018A1064, SK2018A1072), the Natural Scientific Project of Anhui Provincial Department of Education (KJ2019A0371), and Innovation Team of Health Information Management and Application Research (BYKC201913), BBMC

the content relevance of a document and the number of topics^[1]. In addition, current topic recognition models are mostly based on a fixed number of topics, which cannot represent the semantic relevance between topics. At the same time, the recognition results depend only on the probability between topics, which makes it difficult to characterize the inherent hierarchical relationship of comment events. Therefore, it is extremely urgent to dig deeper topic relationship on the review topic.

After years of research, topic detection and tracking^[2] has gradually formed a relatively complete set of algorithms and systems, the goal of which is to classify massive texts according to topics and track their evolution. According to the different text representation models in the corpus, the current topic evolution methods can be divided into two categories. The first type is cluster evolution analysis based on vector space. This type of method treats high-dimensional corpus text as an unordered set of low-dimensional words. It measures the similarity distance between texts and compares the change of the subject at different times. Lu, et al.^[3] proposed a K-means clustering method (EEAM) based on the multi-vector model. This method constructs topic events by calculating the similarity between sub-topics. The topics at different moments are matched according to the similarity between the event vectors to generate a topic evolution set. Lin, et al.^[4] proposed a news review topic evolution model (WVCA) based on word vectors and clustering algorithms. This model first introduced the word vector model into text stream processing to construct word vectors in time series, and then used the *K*-means cluster to achieve the extraction of topic keywords. Cigarrn, et al.^[5] proposed an unsupervised topic detection algorithm (TDFCA) based on formal concept analysis (FCA). By combining similar content in formal concepts into concept lattices, formal concepts are used as the basic carrier to construct Twitter-based terms. Guesmi, et al.^[6] proposed an event topic selection model (FCACIC) based on FCA. This method uses hierarchical clustering to focus on detecting common interest communities (CIC) in social networks, avoiding the introduction of new topics during the topic detection process. It cannot achieve non-human participation, as this type of method only utilizes the similarity distance between texts to determine the correlation between topics.

In order to cope with the topic detection of massive documents in complex environments, some scholars have proposed probabilistic topic analysis. This type of method considers the topic to be smooth in the time dimension, and uses the topic posterior probability of the $t - 1$ time slice as the t time slice. A priori probability, combined with the calculation of the similarity between topics, reduces the calculation bias caused by part-of-speech differences. For example, the probabilistic index model^[7] (PLSI) and the implicit Dirichlet allocation model (LDA)^[8] map out the process of topic identification and evolution by establishing joint probability between texts, topics, and words. AlSumait, et al.^[9] added the online processing function of the text on the basis of the LDA model, and proposed an online Dirichlet probability model to achieve online tracking of topics. Although the focuses of the above studies are different, a common drawback is that the identification of topics relies heavily on the number of topics in text clustering or classification, and the number of topics needs to be specified in advance or iteratively obtained according to a given threshold, which cannot meet the topic evolution process. To enact this need, Herinrich^[10] proposed the infinite latent Dirichlet allocation (ILDA)

model, which implements topic classification based on the time-dependent relationship of text. However, this method still has the problem of “short sightedness”. Considering the iterative problem of the optimal number of topics, there are more meaningless topics, without considering the weight of different topic feature words under the changing number of topics.

The approaches mentioned above have two drawbacks. First, these approaches rely on the number of topics of text clustering. Specifically, the recognition of topics are represented in a certain way without considering the semantic changes of the topic feature words under the dynamic number, which fails to avoid false inheritance of the topic. Second, the correlation strength of feature words under different topics in the ILDA model is weak, which is difficult to mine the inherent hierarchical relationship of events.

The motivation of the paper is to establish a partial order constraint relationship between topics and feature words. To achieve this goal, the model for building topic feature lattice under dynamic topic number (TFL_DTN) is proposed, which realizes the dynamic change perception of topics in time series. Specifically, the TFL_DTN model first obtains the topic-feature word probability matrix and the document-topic probability matrix by modeling documents, topics and feature words; then, the topic association matrix is established, and the features under different topics in the document are calculated according to the joint probability among them. Finally, multi-granularity topic networks are identified based on the characteristics of strongly correlated topics.

2 Related Work

2.1 The Theory of ILDA

LDA is an unsupervised probabilistic model based on probabilistic latent semantic analysis (PLSA), which can implement implicit topic mining of documents^[11]. The LDA model is a three-layer Bayesian network, where documents can be viewed as discrete topic words, and different topics converge to a limited mixture of topic feature words with probability, which is shown as Figure 1. However, the hyper parameters α and β of this model need to be set in advance, and after many simultaneous, the number of topics K , which is manually set, is related to the granularity of text division. In extreme cases, an excessively large number of topics will merge too many divided text topics. The ILDA model^[12] generates an empty topic for integration, unable to obtain valid topic description information, which cannot meet the actual needs of topic division. Besides, the time-dependent relationship of the text realizes the topic classification under the dynamic number of topics. The model structure is shown in Figure 2.

Figure 1 LDA model structure

Figure 2 ILDA model structure

There are two main differences between the two models in Figures 1 and 2. First, on the basis of LDA, ILDA changes the value of the topic number K to a dynamic variable that can be arbitrarily selected in the interval $[1, +]$. Second, the document-topic distribution matrix θ in LDA is determined by the Dirichlet distribution of the hyper parameter α polynomial, and θ in ILDA is determined by the joint Dirichlet allocation process (DAP), regardless of the polynomial distribution of the hyper parameter α ^[13]. DAP is a prior distribution based on random probability, which can be obtained by polynomial π . The polynomial π is a polynomial mixture that obeys the Griffiths-Engen-McCloskey (GEM) random measure distribution. The detailed calculation process can be found in [14]. The calculation of the base distribution O is shown in Equation (1)^[15]. The advantage of DAP is that its input is not a fixed number, but a discrete variable that changes dynamically. ILDA is a three-layer Bayesian network. By abstracting a document into a polynomial distribution containing K topics, and abstracting a topic into a polynomial distribution containing multiple feature words, it implements joint modeling of documents, feature words, and topics. At the same time, the number of topics depends on the random prior distribution of the hybrid model. It no longer requires that the topic priors of the document must obey the Dirichlet distribution, thereby reducing the sensitivity of the topic model to the number of topics and improving the ability to model large corpora.

$$O \sim \text{Dir}(\alpha O) \rightarrow O = \sum_{k=1}^k \pi_k \delta(x - k). \quad (1)$$

2.2 The Theory of Formal Concept Analysis

FCA is a formal method that takes the formal context as its domain, which focuses on describing the hierarchical relationship between concepts^[16]. This theory takes the partial order relationship between formal concepts as the core, and realizes the semi-automatic identification of multi-level ordered concept nodes by establishing the mapping relationship between description objects and attributes^[17]. From the perspective of semantic relationship mining, the concept lattice construction process described by FCA theory can be regarded as the process of hierarchical relationship mining between topic nodes. Meanwhile, the association relationship between the topic concepts is obtained to enhance the semantic relationship between the feature words and the topic.

The mathematical foundation of FCA theory is lattice theory and order theory. The modeling process can be described as follows: First, based on the binary membership between objects and attributes, a ternary context is established including (objects, attributes, relationships). Afterwards, the formal context obtains formal concepts that satisfy the partial order relationship. Finally, a formal concept lattice is established based on whether there is an order relationship between the concepts. In the above process, concept nodes at different levels can reflect different generalization and instantiation relationships between objects, which provide new ideas for obtaining the semantic correlation between topics and feature words.

3 Construction of TFL_DTN

Although the ILDA model can realize online topic identification under dynamic topic numbers, it only determines the topic correlation degree through the probability dependence relationship between topics. Besides, it does not take into account the change in the weight of feature words that may be caused by changes in the topic number. At the same time, the model cannot effectively obtain the hidden hierarchical relationships between topics, and lacks the semantic modeling ability of multi-granularity knowledge. Therefore, this paper makes use of the good dynamic topic modeling ability of the ILDA model, by introducing feature word weight parameters into the topic model, and combining the formal concept analysis method to establish a topic recognition model TFL_DTN. The model first utilizes the ILDA model to simulate the dynamic topic generation process. Secondly, the strength of the connection between the topic and its feature words is determined to establish the topic formal context based on the joint probability. Finally, the concept features are used as a guide to construct the topic feature lattice to identify a multi-granular topic network including a document library, a topic array, and a feature word set, so as to realize the conceptual visual modeling of multi-layer network topics.

3.1 Model Construction based on TFL_DTN

The topic modeling of TFL_DTN can be divided into two sub-models: Self-adaptive topic analysis model (STAM) and Topic feature lattice construction model (TFLCM). First, the STAM model assumes that there is probability dependence between documents, topics, and feature words. Each document converges to a topic with a probability, and each topic extracts feature words with a certain probability, thereby forming a three-layer production probability distribution. Among them, the document is a topic polynomial distribution that obeys the Dirichlet distribution process, and the topic is also a feature polynomial that obeys the Dirichlet distribution, which is shared by the document set containing different mixed topic proportions and feature word weights. For the convenience of explanation, the meanings of the variables and parameters in the model are shown in Table 1. The topic analysis process of the STAM model is depicted as follows. First, we use Gibbs sampling to obtain the dynamic optimal number of topics and establish a document-topic probability matrix and topic-feature word probability matrix to extract topics and feature words respectively. Then, candidate feature words with top N high word frequencies from the document-feature word matrix can be selected, on the basis of extracting feature words with higher weights. Finally, the above steps are iterated to obtain the topic with feature words. The STAM model reconstructs the probabilistic

dependency relationship between the topic and the feature words based on the ILDA model. In essence, the model does not change the generation of documents, topics and feature words, where the relationship still maps topics and feature words to the same semantic space through the probability selection model. Therefore, STAM can still be regarded as a three-layer Bayes network. The functional dependence of the variables and distribution matrix in the model is shown in Figure 3.

Table 1 Parameter comparison in TFL-DTN model

Symbol	Implication	Symbol	Implication
M	Number of corpus documents	P	Topic-feature word probability matrix
N	Number of candidate feature words	Q	Document-topic probability matrix
K	Dynamic topic array	S	Feature word weight matrix
α	Hyper parameters of document-topic probability matrix	O	Dirichlet distribution
β	Hyper parameters of topic-feature word probability matrix	T	Topic collection
γ	Hyper parameters of random parameter probability distribution	D	Document set in the original corpus
π	Joint Dirichlet-Craykey distribution polynomial	D	Document-feature word matrix
z	Topic variable	W	Feature word set
w	Feature word variables	R	Topic association matrix
r	Weight parameter of feature words	I	Formal context association matrix

Figure 3 STAM model structure

The TFLCM model assumes that the probability value of the pair of document and topic has a positive correlation with the correlation strength of the pair of topic and feature word. The greater the probability that a document selects a topic is, and the greater the probability that a topic selects a feature word is. By setting the threshold, the strongly related topic features are

filtered out and mapped into a context matrix of formal context, and the topic feature lattice is finally generated. The generation process of the TFLCM model is expressed as follows. First, the association probability with the highest probability value from the document-feature word probability matrix is extracted, as well as the topic association matrix, by calculating the feature word association strength under different topics in the document. Afterwards, the feature words with strong correlation are generated, and the correlation matrix of topic formal context is generated. Finally, the generated topic feature lattice is reduced through formal concept analysis. The transformation relationship among matrix variables in the model is shown in Figure 4. Based on the above analysis, the relevant definitions are given as follows.

Figure 4 Variable conversion relationships in TFLCM

Definition 1 (Document-Feature Word Matrix) For any document set $D' = \{d_1, d_2, \dots, d_m\}$ containing documents M and feature words V , the frequency vector of the feature word sequence contained in d_i can be computed respectively, then the document-feature word matrix for d_i can be represented as $D_d = \{f_1^d, f_2^d, \dots, f_m^d\}$, where $f_j^d = \sum_{i=1}^V tf_{w_i}$.

Definition 2 (Document-Topic Probability Matrix) For any document $d_i, i \in \{1, 2, \dots, M\}$, if topic probability vectors q_k^z about topic z is generated, the sampling probability of topic z named as $p(z_d|Q_d)$ can be obtained on the basis of the document-topic probability matrix $Q_d = \{q_1^z, q_2^z, \dots, q_k^z\}$ of d_i .

Definition 3 (Topic-Feature Word Probability Matrix) For any topic $z_i, i \in \{1, 2, \dots, N\}$, if the feature word probability vectors p_n^w about the feature word w is generated, the sampling probability $p(w_d|P_z)$ of feature word w in the topic z_i can be obtained on the basis of the topic-feature word probability matrix $P_z = \{p_1^w, p_2^w, \dots, p_n^w\}$.

Definition 4 (Feature Matrix of Feature Words) Let the dependence of the feature word's probability w_i on the topic z_i under the number of topics K' be s_k^z , then the weight matrix of the feature word w_i is $S_z(w) = H(z) - H(z|w)$, where $H(z) = -\sum_{i=1}^n p(z_d|Q_d) \log p(z_d|Q_d)$, $H(z|w) = p(w_d|P_z)H(z) + [1 - p(w_d|P_z)]H(z)$.

Definition 5 (Topic Association Matrix) Let the association set $R_i = \{r_1^z, r_2^z, \dots, r_k^z\}$ between topic z_i and feature words w_i satisfy the following constraints, then it is called the topic association matrix under the topic z_i . In particular, if $r_i^z \geq r_s^z$, where $s = \arg \max_{i=1 \dots k-1} (r_i^z - r_{i+1}^z)$, R_i is called a strong association matrix (denoted as SR_i) of z_i , recorded as the feature set C_{SR_i} of all topic associations that satisfy the constraints SR_i in the topic set. Constraint 1. $r_i^z = \max_{d_i \in D'} q_i^k, q_i^k \in Q_d$. Constraint 2. For any $1 \leq i \leq j \leq k, r_i^z \leq r_j^z$.

Definition 6 (Topic Formal Context) Let the topic formal context be $F = (T, W, I)$, where $T = \{t_1, t_2, \dots, t_K\}$ represents the topic set and $I \subseteq T \times W$ represents the feature word set. $I \subseteq T \times W$ represents the mapping relationship between the topic and the feature set on condition that $(t_i, w_j) \in I, t_i \in C_{SR_i}$.

Definition 7 (Topic Feature Lattice) Let the topic formal context be $F = (T, W, I)$, when $A \subseteq T$ and $B \subseteq W$, for any two-tuple (A, B) satisfying $A^* = B$ and $B^* = A$, (A, B) can be called a set of formal concepts on condition that when $A_1 \subseteq A_2$ or $B_1 \subseteq B_2$, there is a partial ordering relationship $\prec$ that makes $(A_1, B_1) \prec (A_2, B_2)$ be true, where $*$ operation is defined as Equations (1) and (2). The partial order relationship set of all formal concepts in topic formal context F constitutes the topic feature lattice denoted as $L(T, W, I)$.

$$A^* = \{a_i | a_i \in T, \forall b \in B, (b, a_i) \in I\}, \quad (2)$$

$$B^* = \{b_j | b_j \in W, \forall a \in A, (a, b_j) \in I\}. \quad (3)$$

To sum up, the topic in the TFL_DTN model is a latent variable that depends on a mixture of document-topic polynomials, and the feature words depend on significant variables of the multimodal mixture between (topic, feature word) and (feature word, feature word weight). The core idea of the model is described as follows. Firstly, potential semantic associations among variables can be established, through the probability dependence on documents, topics, feature words and feature word weights. Meanwhile, the Dirichlet stochastic process is viewed as the prior distribution of the Bayes network according to Gibbs sampling. What's more, the sampling algorithm obtains the number of dynamic topics, and establishes a document-topic probability matrix, a topic-feature word probability matrix, and a feature word weight matrix. Finally, the TFL_DTN model calculates the topic association matrix to filter out strongly related topic features and maps them into a formal context association matrix. To enact this need, a binary partial order relationship between topics and feature words is established to generate a topic feature lattice. The overall structure of the TFL_DTN model is shown in Figure 5.

Figure 5 Overall structure of TFL_DTN model

3.2 Model Reasoning and Parameter Iteration

Since the derivation and parameter estimation of variables and distribution matrices in the TFL_DTN model are mainly handled by the STAM model, the TFLCM model mainly

performs secondary filtering and correlation analysis of topic feature words. Therefore, this section mainly discusses the hidden variable z and the matrix P , Q parameter estimation. At the same time, the matrix relationship of the TFLCM model is transformed into the algorithm description in Subsection 3.3.

3.2.1 Model Reasoning

The STAM model first introduces hyper parameters α and β for topic probability distributions to represent mixed documents. Meanwhile, the parameter γ is utilized to represent the probability distribution of feature words for mixed topics. Afterwards, the topic of a word is extracted according to the topic probability distribution, and the characteristic words of the topic are generated on the basis of the characteristic word probability distribution. In the above process, since the α in the STAM model undergo multiple iterations, their initial values have little effect on the calculation of the model. The prior γ can be calculated by the GEM polynomial distribution. Therefore, for the joint probability distribution of the solution model, the posterior conditional probability of the variable w must be obtained first, and then it can be used as the prior conditional probability of the probability matrix P to calculate the topic polynomial distribution. Finally, the Gibbs sampling algorithm is used to approximate the estimation. A steady-state distribution matrix of the probability matrices P , Q is obtained. For the convenience of explanation, the meanings of the variables during parameter iteration are shown in Table 2.

The joint probability of all observable and hidden variables in the model with the hyper parameters is shown in Equation (4).

Table 2 Parameter description in STAM model

Symbol	Implication
$ M $	The number of total documents in the corpus
T	The number of topics in training set
V	The number of words in training set
$n_{-i,j}^{d_i}$	The number of words assigned to topic j in document d_i
$n_{-i,*}^{d_i}$	The number of total assigned topics in document d_i
$n_{-i,j}^{w_i}$	The number of words that feature word w is assigned to topic j
$n_{-i,j}^*$	The number of total words assigned to topic j
$\alpha\gamma$	Mixed hyper parameter distribution
$q_j^{d_i}$	Probability estimation of topic j in document d_i
p_w^j	Probability estimation of feature word w under topic j
$n_j^{d_i}$	Word frequency in document d_i with conditional probability of topic j
$n_*^{d_i}$	Word frequencies of all assigned topics in document d_i
n_j^w	The word frequency of conditional probability of feature word w in document d_i
n_j^*	Word frequencies of all feature words in topic j

$$p(w, z, r, P, Q, S | \alpha, \beta, \gamma) = \prod_{n=1}^N p(w_n | r_n) p(r_n | S, P) p(P | \beta) p(z_n | Q) p(Q | \alpha, \gamma). \quad (4)$$

By solving the integrals for P and Q in the above formula, the probability dependence between variables can be further solved, as shown in Equation (5).

$$p(w, z, r, d) = p(w, r | z, d) p(z | d). \quad (5)$$

The above formula can be further expressed as shown in Equation (6).

$$p(w, r | z, d) = \iint p(Q | \alpha, \gamma) p(P | \beta) p(S | P) \prod_{n=1}^N \sum_z p(w_n | P) p(z_n | Q) dP dQ, \quad (6)$$

where $p(P | \beta)$ represents the probability that a super parameter β generates its feature words according to the probability under the topic. $p(S | P)$ represents the probability that the feature word weight matrix depends on the feature word distribution. $p(Q | \alpha, \gamma)$ represents the prior distribution of a Bayes network that depends on the Dirichlet random process. Equation (6) can be further expressed as follows (Equation (7)).

$$p(Q | \alpha, \gamma) = \frac{\Gamma[\sum_{i=1}^K (\alpha_i)]}{\prod_{i=1}^K \Gamma(\alpha_i)} \prod_{i=1}^K \Gamma(\pi_i \delta(x - k)). \quad (7)$$

The posterior probability of the available document library $D = \{w_i\}_{d=1}^{|M|}$ is shown in Equation (8).

$$p(D | \alpha, \beta, \gamma) = \prod_{d=1}^{|M|} p(w | \alpha, \beta, \gamma). \quad (8)$$

From the above formula, the Gibbs sampling formula can be further obtained as shown in Equation (9).

$$p(z_i = j | z_{-i}, w_i, r) = \frac{n_{-i,j}^{d_i} + (\alpha\gamma)}{n_{-i,*}^{d_i} + T(\alpha\gamma)} \cdot \frac{n_{-i,j}^{w_i} + (\beta)}{n_{-i,j}^* + V(\beta)} \bigg/ \left(\sum_{j=1}^T \frac{n_{-i,j}^{d_i} + (\alpha\gamma)}{n_{-i,*}^{d_i} + T(\alpha\gamma)} \cdot \frac{n_{-i,j}^{w_i} + (\beta)}{n_{-i,j}^* + V(\beta)} \right). \quad (9)$$

3.2.2 Parameter Estimation

The STAM model first assigns random topics to feature candidates, and then iteratively calculates the probability distribution of feature words w until the probability is stable (Equation (9)). After that, topic j is extracted from the Q matrix (Equation (10)), and feature words w can be extracted from the P matrix with the probability of the formula (Equation (11)).

$$q_j^{d_i} = (n_j^{d_i} + (\alpha\gamma)) / (n_*^{d_i} + T(\alpha\gamma)), \quad (10)$$

$$p_w^j = (n_j^w + (\beta)) / (n_*^{d_i} + V(\beta)). \quad (11)$$

3.3 Algorithm Description

According to the description mentioned above, the parameter iterative process of the STAM model, as well as the matrix relationship conversion process of the TFLCM model can be described in Algorithm 1.

Algorithm 1 Topic feature lattice construction algorithm

Input: α, β, γ , Document set after initial tokenization $D' = \{ph_1, ph_2, \dots, ph_t\}$, Number of initial topics k_0 , Initial topic weight r_0 , Iteration threshold ϖ , Mapping function ϕ

Output: Topic feature lattice L , matrix P, Q, S , Number of topics K , Topic association matrix R , Topic set T

Step 1: For each $d_i \in D'$

Step 2: For each topic $k \in [1, K]$

Step 3: $K = K_0$

Step 4: $Q \sim \text{Dir}(\alpha O)$, $O : \pi \sim \text{GEM}(\gamma)$

Step 5: $P \sim \text{Dir}(\beta)$

Step 6: end for

Step 7: For each ph_i in D'

Step 8: $ph_i \rightarrow tf_i$

Step 9: $D = [D' : tf_i]$

Step 10: End for

Step 11: For all documents $D' \in [1, K]$

Step 12: for all words $w \in [1, N]$ in M

Step 13: Sample $k \sim p(z|z_{-i}, \alpha, \beta, \gamma, k)$

Step 14: Get v

Step 15: Create a new topic in M

Step 16: Sample α, β, γ

Step 17: end for

Step 18: if $v < \varpi$

Step 19: Re-sampling

Step 20: end if

Step 21: End for

Step 22: For each word $w \in [1, N]$ in M

Step 23: Calculate r

Step 24: Generate S

Step 25: Choose a z from Q where $z \sim \text{Multi}(z|Q)$

Step 26: Choose a w from P where $w \sim \text{Multi}(w|P)$

Step 27: Choose a r from S where $r \sim \text{Multi}(r|S)$

Step 28: End for

Step 29: Get Q, P, S

Step 30: For each z in Q

Step 31: $(D', P, \phi) \rightarrow R$

Step 32: $R \rightarrow I$

Step 33: $(T, W, I) \xrightarrow{\text{FCA}} L(T, W, I) \ R \rightarrow I.$

Step 34: End for

Step 35: Get $\{L(T, W, I), T, P, Q, R\}$

The proposed algorithm of TFL-DTN can be divided into two sub-models: STAM and TFLCM. STAM starts initializing the number of topic number and generating the matrix Q, P on the basis of Dirichlet distribution as shown from steps 1 to 6. The proposed algorithm then gets the document-feature word matrix by calculating vector of feature word frequency as shown from steps 7 to 10. The model iteratively samples the feature words, and calculates the feature word weight matrix under different topic numbers to obtain the topic-feature word probability

matrix and document-topic probability matrix as shown from steps 11 to 21. Consequently, topics, feature words with weights are extracted on the basis of extracting feature words with higher weights as shown from steps 22 to 29. To build the topic feature lattice, TFLCM starts calculating topic association matrix by extracting the association probability with the highest probability value from the document-feature word probability matrix, as well as generating the correlation matrix of topic formal context as shown from steps 30 to 35.

4 Results and Discussions

4.1 Preprocessing

We randomly select 1,583,275 online review data of the 20 automobile brand forums from the two websites of Auto Home and Netease Auto, from August 1, 2019 to September 20, 2019. First, the initial document is segmented, and the standard document corpus is obtained by removing data such as stop words, special symbols, and useless tags. Then, the text is converted into a set of review phrases, and a document-word matrix is established. Afterwards, the TF-IDF vector can be calculated to obtain attribute feature words of comment data.

4.2 Results Analysis

4.2.1 Comparison of the Optimal Number of Topics

In order to verify the rationality of the number of dynamic topics in the STAM model, $\alpha = 0.1$, $\beta = 0.1$, $\gamma = 0.2$, while the initial topic weight $r_0 = 0.5$, and the iteration threshold is 0.2. The number of topics in the Baseline method (LDA model) is set to $K = 40$, whose value is set manually in advance. The STAM model only specifies the number of algorithm iterations (200 and 280 respectively), and determines the number of topics in the document by week.

It can be seen in Figure 6 that although the content of the events described in the corpus is relatively fixed, the number of topics in different periods is dynamically changed, which reflects the correlation between the evolution of topics and the number of topics. In addition, the real data is summarized by the method of manual annotation, and the number of topics varies in the interval [15, 60], which is consistent with the experimental results of the STAM model. At the same time, there is no positive correlation between the document size and the number of topics, but related to the degree of clustering of the actual topics. For example, during the 200 iterations of the STAM model, the document set of the second week contains 847 texts, while the document set for Week 3 is composed of 561 texts. In contrast, the topic number of the former is only 32 but the one of the latter is 49.

In addition, in order to test the capabilities of topic prediction and text representation in STAM model, the perplexity of the above models in the corpus document is calculated. The smaller the value of the perplexity is, the stronger topic prediction capability for the document and it has. The calculation of perplexity is shown as Equation (12). And the experimental results are shown in Figure 7.

$$\text{perplexity} = \exp \left\{ - \left(\frac{\sum_{d \in D} \ln p(w_d)}{\sum_{d \in D} N_d} \right) \right\}, \quad (12)$$

where, $p(w_d) = \prod_{n=1}^{N_n} \sum_{k=1}^K p(w_d|z_k)p(z_k|d_m)$.

Figure 6 Dynamic curve for topic number

Figure 7 Perplexity curve

Figure 7 shows that the perplexity curve of the STAM model is lower than that of the Baseline method as a whole, and the perplexity of the dataset gradually decreases as the number of topics increases. Second, when the number of topics is $K = 70$, the changeable degree is small, which indicates that the topic distribution under this topic number tends to be stable and the model achieves optimal performance, while the STAM model achieves the best performance when the number of topics is $K = 62$, indicating that the model has relatively few topics. In that case, the ability to capture the correlation between topics under a dynamic number of topics is stronger, which reduces the model's dependence on the number of topics K and improves the data representation ability for small sample data sets.

4.2.2 The Construction of Topic Feature Lattice

When the model's iterative probability threshold is set to 0.01, the document-feature word matrix can be acquired in the corpus. At the same time, when a relatively stable state is reached in the STAM model, both the document-topic probability matrix and topic-feature word probability matrix are output. The top 10 feature words with higher probability are extracted, and their feature word weights are calculated separately. Due to the large number of topics, Table 3 lists only six relatively concentrated topics.

Table 3 Results of topic feature words (Partial)

Topic name	Topic feature words with probabilities (descending order)
Topic 22	brake 0.0923 / sideslip 0.0776 / blind zone 0.0681 / resonance 0.0635 / vision 0.0489 / vehicle warning 0.0437 / stability 0.0332 / weight 0.0332 / vehicle stall 0.0274 / loose parts 0.0165
Topic 8	fuel consumption 0.0851 / cost performance 0.0739 / price 0.0722 / comprehensive performance 0.0696 / per hundred kilometers 0.0696 / high speed 0.0584 / working condition 0.0492 / wind resistance 0.0477 / auto parts zero ratio 0.0368 / idle speed 0.0295
Topic 34	acceleration 0.0654 / maximum torque 0.0554 / power 0.0512 / vehicle climbing 0.0477 / idle speed 0.0461 / engine 0.0313 / turbine 0.0313 / transmission 0.0296 / performance 0.0296 / new energy 0.0212
Topic 68	peculiar smell 0.1284 / ride space 0.0937 / suspension system 0.0735 / NVH 0.0667 / Soundproof 0.0545 / vehicle seat 0.0479 / tire noise 0.0448 / assisted driving 0.0379 / seat ventilation 0.0345 / human-computer interaction 0.0307
Topic 71	maintenance 0.0762 / after-sales service 0.0754 / 4S 0.0694 / engine oil 0.0516 / vehicle failure rate 0.0507 / vehicle inspection 0.0472 / vehicle paint 0.0367 / tire 0.0286 / accessories 0.0286 / working hours 0.0104
Topic 79	vehicle stalled 0.1374 / jitter 0.1238 / steering 0.0863 / clutch 0.0794 / automatic 0.0794 / exhaust 0.0634 / vehicle 0.0432 / gear shift 0.0415 / brake 0.0364

Based on the identification results of topic feature words in Table 3, the content of the topic sets is analyzed manually, which is summarized as the following comment topics: Topic 1 (Topic 22) is security evaluation; Topic 2 (Topic 8) is economy evaluation; Topic 3 (Topic 34) is dynamic performance evaluation; 4 (Topic 68) is comfort evaluation; 5 (Topic 71) is service evaluation; Topic 6 (Topic 79) is manipulative evaluation. The top 10 associated topics of the reviews are listed in Table 4, in which the main characters of security-related topic are No. 13 and No. 22. The feature words contain the topic of the braking, sideslip, blind area, early warning etc. These words are highly associated with vehicle safety of the vehicles, which is strongly aligned with the classification results of manual annotation. In addition, according to algorithm 1, the document set is mined for strongly correlated topic features, and the relational matrix of formal context is established to construct the topic feature lattice. The corresponding part of Hasse structure of topic feature lattice is shown in Figure 8, which shows that the closer the concept is to the top-level root node, the more generalization features of topic words are, such as vehicle length, wheelbase, weight. The term specialization is usually more prominent, such as the acceleration, torque, vehicle power, and vehicle hill climbing associated with node Topic 34. The conclusions show that the topic feature lattice based on the TFLCM model can intuitively find the hierarchical relationships of different topic feature words, with a good modeling ability in obtaining generalization of topic words and semantic relationships.

Table 4 Strongly related topic features (Partial)

Topic category	Strongly related topics (in descending order)
Security evaluation	Topic 22 / Topic 13 / Topic 46 / Topic 7 / Topic 21 / Topic 88 / Topic 74 / Topic 62 / Topic 107 / Topic 95
Economy evaluation	Topic 8 / Topic 11 / Topic 33 / Topic 40 / Topic 56 / Topic 75 / Topic 99 / Topic 7 / Topic 61 / Topic 115
Dynamic performance evaluation	Topic 34 / Topic 124 / Topic 81 / Topic 73 / Topic 84 / Topic 113 / Topic 18 / Topic 22 / Topic 51 / Topic 92
Comfort evaluation	Topic 68 / Topic 16 / Topic 53 / Topic 63 / Topic 77 / Topic 127 / Topic 137 / Topic 12 / Topic 64 / Topic 76
Service evaluation	Topic 71 / Topic 5 / Topic 107 / Topic 143 / Topic 19 / Topic 67 / Topic 55 / Topic 112 / Topic 17 / Topic 35
Manipulative evaluation	Topic 79 / Topic 24 / Topic 61 / Topic 19 / Topic 107 / Topic 46 / Topic 93 / Topic 40 / Topic 122 / Topic 51

Figure 8 Hasse diagram of the topic feature lattice (partial)

4.3 Discussion

In order to verify the rationality of the TFL_DTN model, the accuracy rate, recall rate, F1 value, and mean absolute error (MAE) are selected as the evaluation indicators. Meanwhile, a comparison experiment is performed with the TFIDF algorithm^[18], the TDFCA algorithm^[15], and the ILDA algorithm^[12] on the same data set. Tables 5 and 6 are the comparison results of the evaluation indexes of the above algorithms. The results show that the prediction performance of the TFL_DTN model is significantly better than the other methods on the six review topics. The accuracy, recall and F1 value of the measured data can be maintained around 0.65, and the MAE value can be below 0.85. The reason is that the TFL_DTN model combines the

probabilistic relationship and partial order relationship between topic feature words and topics, which not only effectively reduce the dimensionality, but also improve the topic awareness of the document in the changing topic word K .

Table 5 Comparison of accuracy and recall of different algorithms

	TFIDF		TDFCA		ILDA		TFL_DTN	
	Accuracy	Recall	Accuracy	Recall	Accuracy	Recall	Accuracy	Recall
Security evaluation	55.86	56.31	56.88	59.64	58.34	61.47	65.88	60.53
Economy evaluation	57.24	58.14	59.37	61.16	62.76	62.83	67.52	61.04
Dynamic performance evaluation	58.33	58.67	60.55	60.44	61.53	63.35	68.19	62.49
Comfort evaluation	57.75	56.07	59.31	57.33	62.45	62.51	66.67	63.39
Service evaluation	54.38	55.98	55.76	57.28	57.77	59.08	62.98	61.17
Manipulative evaluation	59.15	57.34	61.98	58.46	62.49	60.46	65.74	60.61

Table 6 Comparison of F1 values and MAE of different algorithms

	TFIDF		TDFCA		ILDA		TFL_DTN	
	F1	MAE	F1	MAE	F1	MAE	F1	MAE
Security evaluation	56.08	1.671	58.23	1.434	59.86	1.198	63.09	0.788
Economy evaluation	57.69	1.931	60.25	1.552	62.79	1.371	64.12	0.862
Dynamic performance evaluation	58.50	1.656	60.49	1.274	62.43	1.154	65.22	0.796
Comfort evaluation	56.90	1.637	58.30	1.394	62.48	1.144	64.99	0.835
Service evaluation	55.17	1.671	56.51	1.485	58.42	1.291	62.06	0.884
Manipulative evaluation	58.23	1.937	60.17	1.576	61.46	1.242	63.07	0.927

5 Conclusions

The proposed method TFL_DTN designs a topic recognition visualization to optimize the topic semantic correlation feature generated by the ILDA model. The model iteratively generates the topic-feature word probability distribution matrices and document-topic probability distribution matrices, based on the conditional probabilistic dependency relationship among topics, documents, and feature words. Through the calculation of feature word weights and strong correlation matrix, a visual concept lattice for topic feature is constructed, which realizes the generalization and specialization of semantic relationships between topic features.

Experiments show that the TFL_DTN model has a good ability of topic recognition under dynamic subject numbers. To enact this need, the following innovative points are made in this paper:

- 1) A method is proposed for calculating the correlation strength of feature words under different topics using joint probability of topic-feature words.

2) A method is proposed to construct topic feature lattice in formal context association matrix at multi-granularity.

In order to improve the calculation accuracy of the topic prediction model, the future research will focus on the semantic analysis of topic sentiment, to deeply dig the online users' sentiment tendencies, and establish text sentiment for the hidden features of topics.

References

- [1] Li X, Ruan T, Pan L. Research on the framework of news topic analysis based on fusion denoising and dynamic topic. *Information Science*, 2018, 36(4): 14–21.
- [2] Yu M, Luo W, Xu H, et al. Research on hierarchical topic detection in topic detection and tracking. *Journal of Computer Research and Development*, 2006, 43(3): 489–495.
- [3] Lu N, Luo J, Liu Y, et al. Effective event evolution analysis algorithm. *Application Research of Computers*, 2009, 26(11): 4101–4103.
- [4] Lin J, Zhou Y, Yang A, et al. Analysis on topic evolution of news comments by combining word vector and clustering algorithm. *Computer Engineering and Science*, 2016, 38(11): 2368–2374.
- [5] Cigarrn J, Castellanos Á, Garca-Serrano A. A step forward for topic detection in Twitter: An FCA-based approach. *Expert Systems with Applications*, 2016, 57: 21–36.
- [6] Guesmi S, Trabelsi C, Latiri C. FCA for common interest communities discovering. 2014 International Conference on Data Science and Advanced Analytics (DSAA). IEEE, 2014: 449–445.
- [7] Christidis K, Mentzas G. Using probabilistic topic models in enterprise social software. Abramowicz W, Tolkdorf R. *Business Information Systems. BIS 2010. Lecture Notes in Business Information Processing*, vol 47. Springer, Berlin, Heidelberg, 2010: 23–34.
- [8] Pei C, Xiao S, Jiang M. Research on microblog user clustering based on improved LDA topic model. *Information Studies: Theory & Application*, 2016, 39(3): 135–139.
- [9] AlSumait L, Barbar D, Domeniconi C. On-line LDA: Adaptive topic models for mining text streams with applications to topic detection and tracking. 2008 Eighth IEEE International Conference on Data Mining. IEEE, 2008: 3–12.
- [10] Heinrich G. “Infinite LDA” — Implementing the HDP with minimum code complexity. Technical note TN2011/1, 2011: 1–20.
- [11] Blei D M, Ng A Y, Jordan M I, et al. Latent dirichlet allocation. *Journal of Machine Learning Research*, 2003, 3(4–5): 993–1022.
- [12] Fang Y, Huang H, Jian P, et al. Self-adaptive topic model: A solution to the problem of “rich topics get richer”. *China Communications*, 2014, 11(12): 35–43.
- [13] Gershman S J, Blei D M. A tutorial on Bayesian nonparametric models. *Journal of Mathematical Psychology*, 2012, 56(1): 1–12.
- [14] Feng S. Reversible measure-valued processes associated with the Poisson-Dirichlet distribution. *Scientia Sinica Mathematica*, 2019, 49(3): 377–388.
- [15] Huillet T. Random partitioning problems involving poisson point processes on the interval. *International Journal of Pure and Applied Mathematics*, 2005, 24(2): 143–179.
- [16] Tang X, Li P. Model construction of secondary organization of Weibo search results based on concept lattice. *Information Studies: Theory & Application*, 2014, 37(10): 115–120.
- [17] Pang B, Gou J, Mu W. Extracting topics and their relationship from college student mentoring. *Data Analysis and Knowledge Discovery*, 2018, 2(6): 92–101.
- [18] Xu W, Wen Y. A Chinese keyword extraction algorithm based on TFIDF method. *Information Studies: Theory & Application*, 2008, 31(2): 298–302.