IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received December 10, 2018, accepted December 20, 2018, date of publication December 27, 2018,
date of current version January 23, 2019.

Digital Object Identifier 10.1109/ACCESS.2018.2889993

The Joint Framework for Dynamic Topic Semantic Link Network Prediction

ANPING ZHAO[1], LINGLING ZHAO[2], AND YU YU[1]
[1]College of Teacher Education, Institute of Education Informatization, Wenzhou University, Wenzhou 325035, China
[2]College of Computer and Information Science, Chongqing Normal University, Chongqing 401331, China

Corresponding author: Anping Zhao (apzhao@wzu.edu.cn)

This work was supported in part by the Chongqing Research Program of Basic Research and Frontier Technology under Grant cstc2018jcyjAX0708.

ABSTRACT To explore the maximum potential of textual data, a well-organized dynamic semantic structure of the topics is in fact of great importance for effectively supporting the advanced intelligent application. The proposed framework joints the Gaussian mixture model and the Bayesian network to conduct inference and prediction of topic relationships of dynamic topic semantic link network. The approach is to identify the relationships between the topics and to infer the condition-dependent topic relationships for predicting the topic semantic link network structure, which not only describes the relationships between the topics under changing-dependent conditions but also provides a broader understanding of the relationships between the topics in dynamic evolution processes. The results of the evaluation and experimental analysis indicate that the proposed approach is effective, feasible, and well-suited to predict the dynamic and multi-dimensional relationship structure of topics.

INDEX TERMS Semantic link network, Bayesian network, Gaussian mixture models.

I. INTRODUCTION

Topic reveals correlations of words from massive text data, which provides a way to help people to understand and represent a large volume of unstructured texts. The relationships between the topics are a kind of unstructured text information, which is in fact of great importance for supporting the advanced intelligent application in a global perspective. Real world topic semantic link networks, however, are often multidimensional: two topics may be connected by more than one relation, expressing either the different types of topic relationships or the different quantitative values of the same kind of topic relationships. The relationships among topics will dynamically vary with the change in semantic context under different relationship query conditions. The different relationship types and the changing condition with topical relationship add to the classical problems of network analysis. This additional degree of freedom makes it difficult to treat these kinds of dynamic multi-dimensional topic relationship networks with the available tools.

Understanding and dealing with this kind of topic relationship changes is challenging because of the dynamic semantic nature of topics. On the one hand, the topic condition-dependent relationships should be adapted to the dynamic semantic nature in the context evolutionary process. On the other hand, in order to obtain the implicit constructive information from massive textual content, the topics need to be organized and semantically described in a condition-dependent relationship network that exhibits changes on the topic relationships under semantic conditions. Predicting the dynamic changing relationship structure of a multidimensional topic semantic link network requires the new approach of analysis that takes the interplay among topic relationship dimensions and changes into account.

This fact calls for the development of the novel way and model to capture the semantic condition-dependent topic relationship changing and their structure inference and prediction for maximizing the utility of observed textual information. Toward this problem, a Semantic Link Network (SLN) [1] model is used to solve the problem of topic relationship inference and dynamic semantic structure prediction. Its nodes can be any type of resources. Its edges can be any semantic relations. Semantic links reflect the semantic relations among objects. We have developed a Topic Semantic Link Network model (TSLN) that helps users analyze how the correlated topics link by their semantic implied relationships. TSLN explicitly ties the content of the topics with the connection between them. It enables us to specify only part of the inherent topic semantic links to describe the relationships distribution of topics, which focus on exploring related latent topics and relationship links across conditions.

The main contributions of this paper are as follows:

1) We propose an approach to combine the power of Gaussian mixture models with Bayesian network for predicting dynamic topic-level relationship structure in the context evolutionary process. This shows the new insights into condition-dependent topic relationships network while addressing some of the limitations of Bayesian network to adapt to the dynamic semantic nature of the topic relationships.
2) We present joint models as an analytic framework of probabilistic inference to tackle the problem, which provides a comprehensive description of the dynamic relationships between topics in complex context under condition changing processes and predicts an overall network structure.
3) We conduct experiments to demonstrate that the combination of Gaussian mixture models with Bayesian network is well-suited to predict the semantic structure of dynamic topic relationship network TSLN.

The rest of this paper is organized as follows: In section II we review related work. Section III introduces the joint mixture Bayesian network model and the network learning process. We show the effectiveness of the presented approach by evaluation and experimental results in section IV. Discussion is provided in section V. Finally, we summarize our conclusions and future research in section VI.

II. RELATED WORK

Researchers have been looking into the interplay between the topical relationship and semantic hierarchies. Gaul and Vincent [2] investigated the evaluation of the relationship between topics over time based on content (dis) similarity of the underlying textual material. Anandkumar *et al.* [3] presented a principled approach to handle arbitrary correlations among the topics or latent factors. Ceballos *et al.* [4] proposed a methodology predict research groups from collaboration and topical networks based on network analysis. Wang *et al.* [5] presented a kind of user-to-user topic inclusion degree to approach the link prediction problem in social-information network. Zand *et al.* [6] proposed a mixture model-based semantic image segmentation (OBSIS) approach to partition images into non-overlapping regions. Liu *et al.* [7] proposed a new algorithm that can systematically and comprehensively detect relevant relationships useful for the prediction of an arbitrarily given target relationship in heterogeneous networks.

In the recent years, efforts have been made in combination with the probabilistic graphical model with the mixture model for representing and reasoning the uncertainty of different application fields. Ramos-Líőpez *et al.* [8] proposed a scalable importance sampling algorithm to dynamically update a mixture of Gaussians in conditional linear Gaussian Bayesian networks. Network inference approaches are widely used in biological applications to probe regulatory relationships between molecular components. Ko *et al.* [9] integrated Bayesian network and Gaussian mixture models to describe continuous microarray gene expression measurements. Roos *et al.* [10] presented a dynamic Bayesian network approach to better catch the nonlinear relationships between the variables for short-term passenger flow forecasting. McGeachie *et al.* [11] presented CGBayesNets for predicting an outcome of interest from mixed discrete and continuous variables. An approach was proposed to explore the use of mixtures of Gaussians as a proxy in Hybrid Bayesian Networks for carrying out probabilistic inference [12]. There is also other research related to various research areas. The paper presented a novel model for predicting water leakage in Water Distribution Networks using expert structural expectation-maximization of Bayesian network learning approach [13]. AlJadda *et al.* [14] introduced an extension to Bayesian networks to handle massive sets of hierarchical data. The solutions were proposed to the estimation of finite mixture models with an unknown number of components pertaining to three issues [15]. Reihanian *et al.* [16] proposed a framework for overlapping community detection in social networks by analyzing topics, ratings and links. Gao *et al.* [17] proposed a new method for review expert recommendation using topic relevance and expert relationship. Ozcan and Oguducu [18] studied link prediction of multiple types of nodes and links in the dynamic heterogeneous social networks. They proposed a multivariate method for link prediction in evolving heterogeneous networks using a Nonlinear Autoregressive Neural Network with External Inputs (NARX) [19].

Although the nature of the above work which is models for combining the probabilistic graphical model with the mixture model is similar to ours, the research goals are totally different. Most existing methods are not capable of describing dynamic multi-dimensional varying relationships between the topics under different query correlation conditions. Our approach is more general and flexible than these methods. Mixture Beysian network (MBN) used in TSLN is capable of dynamically learning the multi-dimensional relationship between topics as different topics are emerged. TSLN is based on mixture Beysian network, which can easily be extended to incorporate additional assumptions and information. In our case, we extend the mixture Beysian network to model not only the multi-dimensional relationships between topics, but also the dynamic change between them. It enables us to update the relationship between topics based on the context. MBN tends to be broad and flexible in this context, which allows us to relate topics from differing viewpoints without having to specify in advance what relationship type between topics they expect to find. The way we used gives us a powerful way to understand topics and to capture their multi-dimensional and dynamic relationship.

III. MIXTURE GAUSSIAN BAYESIAN NETWORK MODEL

A. PROBLEM STATEMENT

The graphical models approaches are able to efficiently represent the joint probability distribution of a domain. Bayesian network is an important subclass of graphical models that has

a solid theoretical foundation and includes clear semantics. *A Bayesian network* can be described as a directed acyclic graph with nodes representing a set of random variables $X = \{X_1, \ldots, X_i, \ldots, X_n\}$ and a set of directed edges (links) connects pairs of nodes, $X_i \rightarrow X_j$, representing the relationships between the nodes.

We adopt the Bayesian network approach to represent the structure of TSLN so that the probabilistic graphical model based TSLN can be the semantic model of representing the condition-dependent relationships between topics.

Definition 1: Given a finite set of topics $T = \{t_1, t_2, \ldots, t_N\}$, *a topic semantic link network (TSLN)* is a directed graph G which is a corresponding Bayesian network and represented as the joint probability distribution over all topics in G, denoted as $S_t(TopicSet, LinkSet)$, where S_t is the name of the TSLN, *TopicSet* is a set of topics, and *LinkSet* is a set of semantic links in form of $t_i \rightarrow t_j$, representing the relationship between t_i and t_j, where $t_i, t_j \in TopicSet$.

Let $P(\cdot)$ be a joint probability function over the topics in T, and let X, Y and Z stand for any three disjoint subsets of topics in T, X and Y are said to be conditionally independent given Z, denote as $I(X, Z, Y)$ *iff* $P(x|y, z) = P(x|z)$. Especially, if $Z = \Phi, I(X, \Phi, Y)$ *iff* $P(x|y) = P(x)$ whenever $P(y) > 0$. Based on this conditional independence property and the chain rule of basic probability theory, the joint probability distribution can be represented as a product form of conditional probabilities. Each topic node t_j and its associated parent topic nodes $a(t_j)$ of TSLN, for $j = 1$ *to* N, corresponding to a given topic subnetwork. The overall topic semantic link network is conceived as a set of such topic subnetworks. These definitions are the basis for our later discussion.

In the TSLN, applying the conditional probability distribution form of the Markov property for a stochastic process, the absence of semantic links encodes conditional independencies among topic nodes. Therefore, each topic is independent of non-descendant topics in the TSLN, given the parent topic nodes. It's able to offer a probabilistic framework to describe the structure of TSLN which represents the condition-dependent relationship links between the topics, and it's also well-suited to detect these relationship links by describing the variation typically observed in the topic data and incorporating prior knowledge on the topic relationships.

Understanding and predicting the dynamic multi-dimensional relationship structure is one of the outstanding challenges in the study of TSLN. It is reasonable to model multimodal distribution associated with condition-dependent TSLNs that exhibit changes in the different dimension of topic relationships. Naturally, joint model is one of the adopted approaches. A solution to this problem is to combine Gaussian mixture models with Bayesian network. This mixture Bayesian network model can be used to describe potentially complex distribution of dynamic topic relationships expression across conditions. The topic relationship links can be predicted based on TSLN with the change in the query context across correlation conditions. In this context, we seek an effective graphical representation of TSLN which is embedded in a probabilistic model and inferred using a mixture Bayesian network approach. We introduce the dynamic topic semantic link network prediction problem which focuses on a joint Gaussian mixture models and Bayesian network method to infer and predict the condition-dependent dynamic topic relationships structure.

Definition 2: Given a topic semantic link network TSLN modeled as a mixture Baysian network $G = (TopicSet, LinkSet)$, the *Dynamic TSLN Prediction* problem needs to return a joint probability function $p(G)$ over the topics in TSLN by learning each topic subnetwork structure and estimating the parameters of the Gaussian mixture models.

We use this joint framework to show that it is possible to detect, without prior knowledge of what kinds of structure and semantic links we are looking for, a very broad range of types of topic relationship structure in dynamic topic semantic link networks.

B. MODELING TSLN WITH MIXTURE BAYESIAN NETWORK

We seek an effective model to describe both condition-dependent topic relationship structure and potential changes under different correlation conditions in TSLN. We model the joint probability functions of topic subnetworks by integrating mixtures of Gaussian densities into Bayesian networks, which are applicable to TSLN to infer dynamic topic semantic links. Based on the Bayesian network framework, the joint probability distribution of the TSLN is:

$$p(G) = p(t_1, t_2, \ldots, t_N) = \prod_{j=1}^{N} p(t_j|a(t_j)) = \prod_{j=1}^{N} \frac{p(t_j, a(t_j))}{p(a(t_j))} \tag{1}$$

where N is the total number of topics in the G and $a(t_j)$ is the set of parent topic nodes of t_j in the $j-th$ topic subnetwork ($j = 1$ *to* N), each subnetwork corresponding to a given topic node t_j and its parent topic nodes $a(t_j)$ in TSLN. $p(t_j, a(t_j))$ is the joint probability density function of the parent and child topics and $p(a(t_j))$ is the marginal probability density function of the parent topics in the $j-th$ topic subnetwork.

In our problem, we assume that the joint and marginal probability density of topic in TSLN follows a multivariate Gaussian distribution. Therefore, possible topic subnetwork structures are modeled as a Bayesian mixture network model by integrating mixtures of multivariate Gaussian distribution into Bayesian networks. For the $j-th$ topic subnetwork, the joint probability density function with a mixture of K_j multivariate Gaussian distribution is represented as follows:

$$p(t_j, a(t_j)) = \sum_{k=1}^{K_j} \pi_{jk} \mathcal{N}(x_{ji}|\mu_{jk}, \Sigma_{jk}) \tag{2}$$

where π_{jk} is a weight of K_j multivariate Gaussian distributions $\mathcal{N}(x_{ji}|\mu_{jk}, \Sigma_{jk})$, and $\sum_{k=1}^{K_j} \pi_{jk} = 1$. Based on this, the overall likelihood of the data across all topics in the topic

semantic link network G is:

$$p(G) = \prod_{j=1}^{N} \frac{p(t_j, a(t_j))}{p(a(t_j))} = \prod_{j=1}^{N} \frac{\prod_{i=1}^{D} \left(\sum_{k=1}^{K_j} \pi_{jk} \mathcal{N}(x_{ji} | \mu_{jk}, \Sigma_{jk}) \right)}{\prod_{i=1}^{D} \left(\sum_{k=1}^{K_j} \pi_{jk} \mathcal{N}(x^*_{ji} | \mu^*_{jk}, \Sigma^*_{jk}) \right)} \tag{3}$$

where D is the total number of topic semantic link observations in the $j-th$ topic subnetwork. x_{ji} is a vector including the $i-th$ topic semantic link correlation observation for $i = 1\ to\ D$. The multivariate Gaussian distribution

$$\mathcal{N}(x_{ji} | \mu_{jk}, \Sigma_{jk}) = (2\pi |\Sigma_{jk}|)^{1/2} exp \times \left[-\frac{1}{2} (x_{ji} - \mu_{jk})^T (\Sigma_{jk})^{-1} (x_{ji} - \mu_{jk}) \right] \tag{4}$$

where μ_{jk} is a mean vector and Σ_{jk} is a variance-covariance matrix for component k of the K_j Gaussian mixtures. μ^*_{jk} and Σ^*_{jk} are the mean vector and variance-covariance matrix corresponding to the marginal probability density function, respectively.

C. NETWORK LEARNING

Mixture Bayesian network learning of TSLN consists of two main components. The first component is to learn each topic subnetwork. The goal is to identify each topic subnetwork structure and to estimate the parameters of the Gaussian mixture models and the number of Gaussian mixture components best supported by the data. The second component is to integrate the individual topic subnetworks into an overall network, and apply the conditional probability distribution of the Markov property to discover the conditional independence of the topic nodes of the given parent node.

The Bayesian information criterion (BIC) is one of the most widely known and popular tools in statistical models selection because of its computational simplicity and effective performance in many modeling frameworks. It can be used to choose between mixtures with different numbers of Gaussians [20]. The BIC score for a probability model $P(S)$ is as follows:

$$BIC(P) = logP(D_S) - \frac{logN}{2} |P| \tag{5}$$

where D_S is the dataset D restricted to the variables of interest S, N is the number of data points in the dataset D, and $|P|$ is the number of parameters in P. In this work, BIC can be used to identify the optimal number of Gaussian mixture components and associated parameter estimates for topic subnetwork by an asymptotic approximation to a transformation of the Bayesian posterior probability of a candidate model [21]. And it could potentially take the large number of topic relationships and query conditions into consideration in TSLN relationship prediction. The model with the smallest BIC value is preferred over the alternative specifications. The basic BIC based TSLN structure learning algorithm is presented as Table 1. The Expectation-Maximization (EM) is an iterative algorithm which alternates between inferring the missing values given the parameters (E step), and then optimizing the parameters given the "filled in" data (M step), often with closed-form updates at each step. EM exploits the fact that if the data were fully observed, then the Maximum likelihood or Maximum a posterior estimate would be easy to compute. We use the EM algorithm which has been explored in [9] to estimate the mean and variance-covariance parameters of the Gaussian mixture components for each topic subnetwork. The equations in the EM iterations for the subnetwork of the topic t_j are as follows:

TABLE 1. The BIC based network structure learning algorithm.

Input: a set of topics $\{t_1, t_2, \ldots, t_N\}$
Output: the G of TSLN structure
1. ***For each*** topic t_j
2. Evaluate all the edges $t_j \rightarrow a(t_j)$ using BIC score
3. Construct subnetwork G_j of topic t_j
4. ***End For***
5. Integrate all G_j into a overall network G
6. Eliminate cyclic by removing the edge of G with the weakest BIC
7. Reconstruct the G by computing the BIC score for the G
8. ***For each*** G_j of t_j
9. if the nodes of G_j without any edges in G
10. Evaluate edges between the nodes using the BIC score
11. Reconstruct G_j of t_j
12. ***End For***
13. Integrate all reconstructed G_j into the G
14. ***Repeat***
15. Eliminate cyclic in the G
16. Compute the BIC score for the G
17. ***Until*** the G with the best BIC.

E step:

$$p(Z_{jk} | x_{ji}, \theta) = \frac{p(Z_{jk} | \theta) p(x_{ji} | Z_{jk}, \theta)}{p(x_{ji} | \theta)} \tag{6}$$

where $p(Z_{jk} | x_{ji}, \theta)$ is the evidence that the semantic link correlation observation vector x_{ji} corresponding to the $j-th$ topic subnetwork refers to the k_{th} multivariate Gaussian mixture component Z_{jk} $(k = 1\ to K)$, and θ is all the unknown parameters across all the Gaussian mixture components.

M step:

$$\pi_{jk} = \frac{\prod_{i=1}^{D} p(Z_{jk} | x_{ji}, \theta)}{D} \tag{7}$$

$$\mu_{jk} = \frac{\prod_{i=1}^{D} p(Z_{jk} | x_{ji}, \theta) x_{ji}}{\prod_{i=1}^{D} p(Z_{jk} | x_{ji}, \theta)} \tag{8}$$

$$\Sigma_{jk} = \frac{\prod_{i=1}^{D} p(Z_{jk} | x_{ji}, \theta)(x_{ji} - \mu_{jk})(x_{ji} - \mu_{jk})^T}{\prod_{i=1}^{D} p(Z_{jk} | x_{ji}, \theta)} \tag{9}$$

In the E step, the probability that each observation refers to each Gaussian mixture component is obtained based on

the parameter values estimated in the M step. In the M step, given the distribution of the observations over the Gaussian mixture components. Parameters are estimated as the values that maximize the log likelihood of the observed data.

The different relationship types of topic subnetwork structure in TSLN can be detected using the machinery of Gaussian mixture models and the EM algorithm. The number of Gaussian mixture components can also be inferred from the data and the vary parameters of the model are used to find the best fit to the observed topic network. We define r_{jn} to be the correlation probability that a semantic link from the parent topic t_n to the topic t_j within a Gaussian mixture component, $t_n \in a(t_j)$.

$$r_{jn} = \sum_{k=1}^{K_j} \pi_k (r_{jn})_k = \sum_{k=1}^{K_j} \pi_k \left(\frac{\sigma_{jn}}{\sigma_j \sigma_n} \right)_k \tag{10}$$

where $\sum_{n=1}^{N} r_{jn} = 1$, σ_j, σ_n and σ_{jn} are the covariance and variance estimates that correspond to the off-diagonal and diagonal entries of Σ_{jk}, π_k is the weight of the Gaussian mixture component.

The directed relationship between the parent and child topics in each topic subnetwork structure can be characterized by the weighted sum of the correlation probability between the parent topic and child topic estimated for the Gaussian mixture component. Therefore, directed graph G of TSLN is said to be changing if its condition-dependent relationship between topics vary across different conditions of the vary parameters of the model. The unknown parameters across the Gaussian mixture components are $\theta = \{\pi_{jk}, r_{jn}\}$. Thus, the network learning problem, in the present case, is transformed to an optimization problem, which requires to maximize the likelihood $p(Z_{jk}|x_{ji}, \theta)$ of the observed data with respect to θ. We use the approach implemented by Newman and Leicht [22] to solve a similar problem to our work in nature. It iterates equations to make the algorithm converge to the global maxima, and infers the network with a range of starting points for each mixture parameter.

IV. EVALUATION AND EXPERIMENTAL RESULT

The evaluation and experimental results in this section demonstrate the utility of combining Gaussian mixture model with Bayesian network for dynamic topic semantic link network prediction.

A. DATASETS

We use the two publicly available datasets for the evaluation of model. The first dataset is from an existing New York Times corpus (NYT).[1] This dataset consists of over 1.8 million articles spanning January, 1987- June, 2007. The dataset excludes wire services articles that appeared during the covered period. We randomly selected a number of articles for use in our experiment, which was pre-processed by separating sentences and removing non-alphabetic characters and single-character words. The final NYT dataset comprises of 10,000 total documents with 7123 unique terms and an average document length of 1865.

The second dataset is the 20 Newsgroups. It is a collection of approximately 20,000 newsgroup documents, partitioned (nearly) evenly across 20 different newsgroups. In our work, for evaluating dynamic topic relationship, we use the existing topic result of 20 Newsgroup data.[2] It includes 30 topics from 13205 of lexicon size.

We utilize LDA model to discover topics from the dataset. Perplexity is used as a measure for a quantitative evaluation of topic models. In our work, we get the optimal number of topics of 100 based on the perplexity of the final NYT dataset. In our previous work, the initial relationship links between topics can be achieved by analyzing the network coherence of a word co-occurrence network, to which a topic refers. Topic model is a content modeling algorithm. It uses words as input, but words are often ambiguous: Style, context can change their meaning. As a result, the meaning of the topic is interpreted by the user. For the both of datasets, the articles already have category labels. We have the following observation: topics are related to a single topic category or many categories, which corresponding to the different semantic context. And if topics are associated with many articles in a particular category, the relationships between topics are likely to belong to that category. Because these topic association relationships are based on the different semantic relatedness of topics under the different semantic context, it is hard to automatically label the relationship types into these relationships. We therefore manually labeled the multi-dimensional relationship type of the relationship links between topics. We have the category labels of news topics. The relative correlation strength of topics can be simply measured by the percentage of topic relationships belonging to that category.

B. EVALUATION METRICS

To evaluate the approach, we apply it to the above two datasets and we choose to compare different topic relationship schemes to experimentally analysis of our approach. The evaluation comprises two parts. First, we evaluate the effect of the model of two aspects: number of mixture components and semantic links prediction, which is a basis for the dynamic TSLN prediction. Second, we evaluate the quality of the resulting TSLN to show the ability of the approach to detect a dynamic multi-dimensional relationship structure of TSLN under different correlation conditions.

The approach encompasses a wide range of network models, from multiple Gaussian distribution to Bayesian networks with variable number of the mixture Gaussian components across sub-networks of TSLN. The different parameter setting of TSLN allows us to compare the results of the approaches to analyze the effectiveness of each. We use

[1] http://archive.ics.uci.edu/ml/

[2] http://www.cs.cmu.edu/ tmalisie/lda/

BIC score to evaluate the mixture components estimation. It is used to determine the optimal number of mixture Gaussian components best supported by the data for each sub-network, which accounts for model complexity and data available. Assessment of the mixture model parameter estimation can be utilized to gain insights into each the topic semantic link sub-networks.

One aspect of the approach is the capacity to predict new semantic links in the presence of a dynamic TSLN. Link prediction is a natural generalization task in the TSLN, which is the problem of predicting new upcoming connection in the network and another way to measure the quality of our approach. A quantitative evaluation of this ability is how well the predictive power of the TSLN can be improved by mixture Bayesian network model, which predicts the implicit topic semantic links after observing a portion of the links between topics of TSLN. We use *link perplexity* to evaluate semantic link prediction. It can be described as follows. Suppose we observe topic semantic links $l_{1:P}$ from a topic semantic link network and are interested in which model provides a better semantic link predictive distribution $p(l|l_{1:P})$ of the remaining dependency semantic links. Link perplexity shows the predictive distributions of set of semantic links L of topic semantic link network:

$$linkPerp = exp - \frac{\sum_{i=P+1}^{L} \log p(\tilde{l}_i|l_{1:P})}{|L|} \tag{11}$$

Models that give lower link perplexity to the unseen topic semantic link better capture the semantic relationship structure of topics.

The mixture Bayesian network used in TSLN provides the description of the dynamic multi-dimensional relationship between topics. We focus on tackling with the dynamic topic semantic link changes under different parameter conditions, which are not addressed in the other previous studies. This evaluation aims to gain insights into general and condition-dependent dependency relationships between topics, and demonstrate the ability and flexibility of the mixture Bayesian network model to detect dynamic multi-dimensional relationship structure of TSLN without knowing the kinds of semantic links. Correlation probability distribution θ_r shows the connection from topics with a kind of topic relationship type to each other topics. The values of the θ_r behave as a kind of relationship measure, indicating how important a particular semantic link is to a particular correlation type between topics. Thus, given the TSLN, we use *the correlation probability distribution* function to analyze the potential changes of TSLN structure under different mixture component.

C. EXPERIMENTAL RESULTS

We use the BIC score as a criterion for estimating the number of mixture components on the NYT dataset. Fig. 1 present BIC scores of the TSLN under the different number of mixture components. We can see that the overall trends indicate that for the both of datasets used in our work,

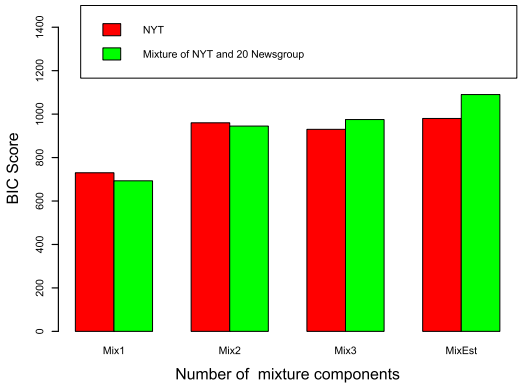


FIGURE 1. BIC scores with the different number of mixture components. Mix1– 1 mixture component. Mix2– 2 mixture components. Mix3–3 mixture components. MixEst–the number of mixtures is estimated from the data.

Gaussian mixtures give a better description of the semantic links between topics than single distribution. This also shows that the adequacy of the approach for describing multi-dimensional semantic relationships of TSLN. However, increasing the number of Gaussian mixture components does not necessarily lead to the increase of BIC score. The approach automatically selects the optimal number of mixture component based on the data. In the TSLN, the advantage of the mixture Bayesian network in estimating the optimal number of mixture components is obvious to method with a fixed number of mixture components. BIC-based methods become computationally expensive when the number of covariates is large. In our case, the number of covariates is limited and the computational complexity is acceptable.

In the link prediction, Correlated Topic Model (CTM) is an extension of the Latent Dirichl *et al.* location model to model both topics and the correlations between them [23]. Our evaluation compares the link prediction results of the CTM approach on the NYT dataset and the mixture of NYT and 20 Newsgroup. We use the metric link perplexity to measure the predictive power of the TSLN and to compare the results from the MBN with CTM. The 10-fold cross-validation is utilized to confirm the reliability of the TSLN, where the original links of TSLN are divided into 10 subgroups with equal size. For each fold, we utilize one subset as the test links, and use the other 9 folds to train the model. The result of the predictive distribution comparison of the NYT is shown in Fig. 2(a). It shows the link perplexity under the varying number of the topics in the TSLN. We can find that the MBN model has more predictive power than the CTM when there are more topics. It can be seen that MBN obtains lower perplexities than the CTM as the number of topics is increased to the optimal number. This is because the TSLN includes directional and non-directional semantic relationships between topics. In directional relationships, the child and parent topics are identified, meanwhile in non-directional relationships, topics are linked but there is neither a parent nor a child topic, which was related through intermediate topics.

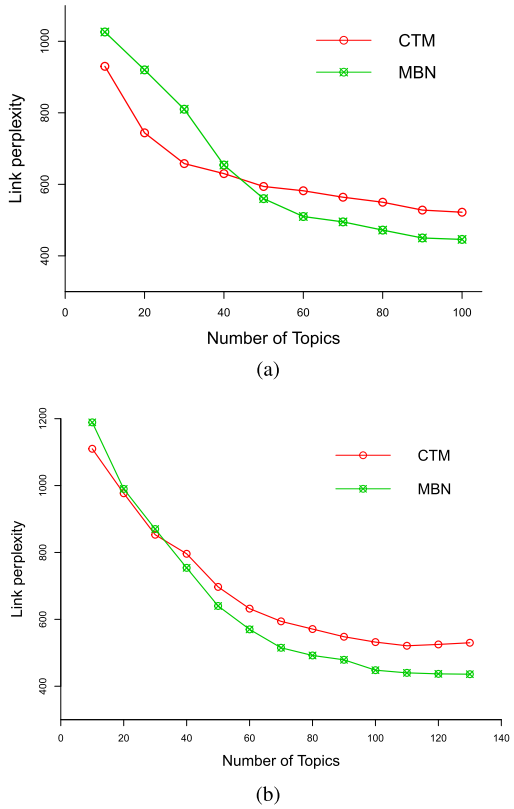


FIGURE 2. The link perplexity comparison under MBN and CTM. (a) The link perplexity comparison on NYT. (b) The link perplexity comparison on 20 Newsgroup.

The MBN model uses both of directional and non-directional semantic links information in the TSLN for links prediction. Thus, it is able to detect a range of different semantic link types without knowing in advance what type to expect and can be applicable to TSLN for capturing the multidimensional semantic link information about the topics.

As an extended example, we run our approach supposing that all topic relationships from NYT and 20 Newsgroup can be represented as a mixture of Gaussian. Under these constraints, we randomly mixed 130 topics from the two datasets. We can think of a new link as an unobserved relationship between newly added topics that affect the topic relationships observed in the TSLN. The result of the predictive power comparison with the mixture of the NYT and the 20 Newsgroup is shown in Fig. 2(b). We can see that the MBN provide more predictive power than the CTM. This also indicates the stability of the approach. Fig. 3 summarizes part of the TSLN predicted by the mixture Bayesian network approach on the mixture dataset of NYT and 20 Newsgroup. Some new links between the topics from the NYT and the topics from the 20 Newsgroup are predicted (e.g. link between "Computer" and "Technology", "Security" and "Technology"). We can note that the model can identify newly emerged relationship structure of TSLN. The most representative words of each topic of the part of the TSLN from the NYT and 20 Newsgroup is shown in Table 2.

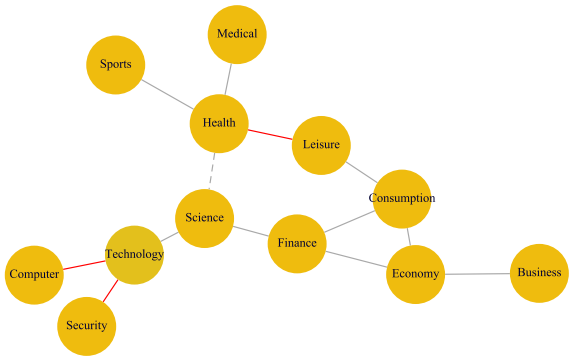


FIGURE 3. The part of the TSLN predicted by the MBN on the mixture dataset of NYT and 20 Newsgroup.

TABLE 2. The most representative words for the part of topics from NYT and 20 Newsgroup.

Topics	The most representative words
Economy	economy tax technology stock income
Finance	finance stock investment fund bank
Consumption	consumer sell consumption price market
Technology	software manufacture technology internet product
Health	health desease insurance research medical
Medical	drug treatment doctor medicine cancer
Leisure	lifestyle exercise wellbeing dress fat
Sports	game team hockey play games
Science	space snsa earth henry launch
Computer	drive card system scsi hard

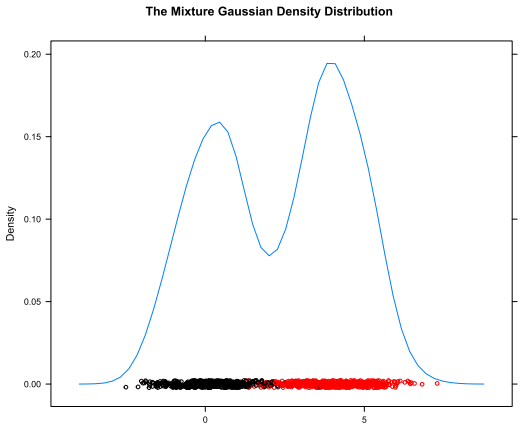


FIGURE 4. An example of two mixture Gaussian components.

Our approach starts by modeling TSLN as a mixture Gaussian of the relationship between topics. It offers novel information on dynamic multi-dimensional relationship types by the different correlation probability between topics under the different mixture components. Thus, there is another question that needs to be answered. Can this mixture Bayesian model capture dynamically topic relationship changes? We analyze potential changes of TSLN structure by changing the correlation probability distribution under different mixture component, which can enhance the understanding of the relationship difference between topics. In this mixture of multi-dimensional relationship space, as an example of two mixture Gaussian components, we might describe this kind of dynamics of multi-dimensional relationship in the Fig. 4. We can see from the figure that the parameters (density mean)

of correlation probability distribution θ_r between topics are different under the different mixture components. In this context, the topic relationship structure in TSLN corresponding to the different query could be dynamically changed and be described by the different mixture components. For example, if a query context is about the ''Health'', we would expect the strong relationship between topic ''Sports'' and ''Medical''. There would be the different correlation probability between topic ''Sports'' and ''Medical'' under different mixture components for another query context. Relationship strength is spatially aligned with their query context and mixture component.

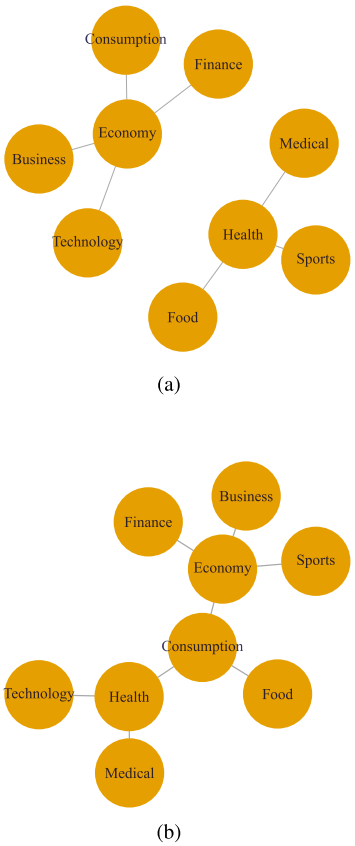


FIGURE 5. The different topic semantic link sub-networks with the different mixture component. (a) The topic sub-networks with the mixture component 1. (b) The topic sub-networks with the mixture component 2.

The two sub-network structure of TSLN detected by the MBN under the different mixture component is shown in Fig. 5(a). We can see from Fig. 5(a) that there is the two topical sub-network of TSLN between several topics. The topic of Economy depends on the topics of Technology, Business, Consumption and Finance, while the Health topic is dependent on the Medical, Sports and Food topics. However, there are different semantic links between these topics when they are described by another mixture component. The result is presented in Fig. 5(b). The topic of Health does not depend on the topic of Food but is closely related to the topic of Technology. The reason is that the parameter corresponding to each mixture component of a topic sub-network offers the different correlation pattern information on the TSLN. Therefore, these results show that condition-dependent topic semantic links are best described by the MBN which can capture dynamically topic relationship change.

This finding supports the different correlations of semantic links between the topics detected by the MBN approach. The experimental results further advocates the capability of the MBN approach to uncover the dynamic and multi-dimensional topic semantic link structure contained in the data.

V. DISCUSSION

The analyses presented here were intended for better understanding joint framework of mixture Bayesian network approach for dynamic TSLN. We showed how a joint framework of the Gaussian mixture model and a Bayesian network approach are related to relationship dynamics at the topic level. We discuss some aspects of condition depended relation inference and prediction in the evolution process of dynamic relationships.

The joint framework was able to identify different mixture components of the relationship between topics depending on the context conditions under consideration. We assume that multimodal probability distribution associated with TSLN can capture condition-dependent topic relationships changing. The optimal number of Gaussian mixture components was estimated to precise characterize the relationships of topics, which varied across topic subnetworks. Therefore, prediction performance of dynamic TSLN depends in part on the high degree of interconnectedness among the topics in this network. Inadequate number of topic relationships also affects performance. The degree of relationship between these topics was expected to vary across different conditions.

The joint framework showed the flexibility to detect dynamic multi-dimensional relationship structure of TSLN without knowing the kinds of semantic links. It provides the ability to capture changes on topical relationships across the multiple context conditions. The flexibility was granted by the mixture Bayesian network learning process. Dynamic TSLN was learnt by the use of potentially variable numbers of mixture components across the network. The optimal number of mixture components for each topic sub-network was estimated from the data. This data-driven mixture Bayesian network approach can be used in different scenarios.

There is potential inadequacy of the fixed number of mixture components to describe TSLN. The description of the topic sub-networks supported by the data based on the BIC score favored mixture over single distribution. This indicates that for the TSLN, mixture distribution offers a better description of the relationship between topics than single distribution. Thus, a higher number of iteration was required to estimate the model parameters and predict the TSLN structure. The model uses the EM algorithm to obtain parameter estimates and the BIC score to identify the optimal number of mixture components supported by the data. This also enabled

the evaluation of the joint model to be independent of the specific algorithm used to learn the structure of TSLN.

VI. CONCLUSION AND FUTURE WORKS

Modeling dynamic topic semantic link network with the multi-dimensional relationship change across conditions is an important way to implement advanced intelligent application in text mining and analysis. In this paper, we proposed and explored the use of joint Gaussian mixture models and Bayesian network, a mixture graphical model method, to infer the dynamic condition-dependent multi-dimensional relationships between topics and to predict the semantic structure of TSLN under changing condition. The main motivation is to facilitate users' understanding and to cope with dynamic vary relationships among topics for supporting an real intelligent application.

We have presented a joint mixture Gaussian and Bayesian network approach for exploratory analysis of TSLN in which topic semantic links are multi-dimensional types based on the observed patterns of connection between them. A comprehensive characterization of the dynamic TSLN was obtained using the joint approach which is able to uncover changes in the network. The methodology revolves around a Bayesian network based Gaussian mixture models to describe potentially complex distribution of topic relationships expression and its changes in TSLN across conditions. The work presented here demonstrates the potential of combining Bayesian network with Gaussian mixture model in predicting particular relationship type-dependent dynamic topic relationships and structure from text data.

Our framework makes assumptions that multimodal probability distribution associated with TSLN can capture condition-dependent topic relationships changing. Multi-dimensional semantic links of TSLN enhance the understanding of the differences in the relationships between the topics, which pertain to different mixture model components. While because of heterogeneity of the relationship in the real data, the approach should be extended to use mixtures of other distribution, or incorporate prior information on the topics in the semantic link network to offer a more comprehensive understanding of dynamic topic relationship network. This is a worth investigating issue for our future work.

REFERENCES

[1] H. Zhuge, *The Knowledge Grid:Toward Cyber-Physical Society*, 2nd ed. River Edge, NJ, USA: World Scientific, 2012.

[2] W. Gaul and D. Vincent, "Evaluation of the evolution of relationships between topics over time," *Adv. Data Anal. Classification*, vol. 11, no. 1, pp. 159–178, Mar. 2017.

[3] Z. Pan, Y. Liu, G. Liu, M. Guo, and Y. Li, " Topic network: Topic model with deep learning for image classification," *J. Electron. Imag.*, vol. 27, no. 3, p. 033009, 2018.

[4] H. Ceballos, S. E. Garza, and F. J. Cantu, "Factors influencing the formation of intra-institutional formal research groups: group prediction from collaboration, organisational, and topical networks," *Scientometrics*, vol. 114, no. 1, pp. 181–216, 2018.

[5] Z. Wang, J. Liang, and R. Li, "Exploiting user-to-user topic inclusion degree for link prediction in social-information networks," *Expert Syst. Appl.*, vol. 108, pp. 143–158, Oct. 2018.

[6] M. Zand S. Doraisamy, A. A. Halin, and M. R. Mustaffa, "Ontology-based semantic image segmentation using mixture models and multiple CRFs," *IEEE Trans. Image Process.*, vol. 25, no. 7, pp. 3233–3248, Jul. 2016.

[7] Y. Liu, S. Xu, and L. Duan, "Relationship emergence prediction in heterogeneous networks through dynamic frequent subgraph mining," in *Proc. 23rd ACM Int. Conf. Conf. Inf. Knowl. Manage. (CIKM)*. New York, NY, USA: ACM, 2014, pp. 1649–1658, doi: 10.1145/2661829.2661916.

[8] D. Ramos-López *et al.*, "Scalable importance sampling estimation of Gaussian mixture posteriors in Bayesian networks," *Int. J. Approx. Reasoning*, vol. 100, pp. 115–134, Sep. 2018.

[9] Y. Ko, C. X. Zhai, and S. Rodriguez-Zas, "Inference of gene pathways using mixture Bayesian networks," *BMC Syst. Biol.*, vol. 3, no. 1, p. 54, 2009.

[10] J. Roos, S. Bonnevay, and G. Gavin, "Dynamic Bayesian networks with Gaussian mixture models for short-term passenger flow forecasting," in *Proc. 12th Int. Conf. Intell. Syst. Knowl. Eng. (ISKE)*, Nov. 2018, pp. 1–8, doi: 10.1109/ISKE.2017.8258756.

[11] M. J. McGeachie, H.-H. Chang, and S. T. Weiss, "CGBayesNets: Conditional Gaussian Bayesian network learning and inference with mixed discrete and continuous data," *PLoS Comput. Biol.*, vol. 10, no. 6, p. e1003676, 2014.

[12] A. Salmerón and F. Reche, "Mixtures of Gaussians as a proxy in hybrid Bayesian networks," in *The Mathematics of the Uncertain* (Studies in Systems, Decision and Control), vol. 142. 2018, pp. 367–374.

[13] S.-S. Leu and Q.-N. Bui, "Leak prediction model for water distribution networks created using a Bayesian network learning approach," *Water Resour. Manage.*, vol. 30, no. 8, pp. 2719–2733, 2016.

[14] K. AlJadda *et al.*, "Mining massive hierarchical data using a scalable probabilistic graphical model," *Inf. Sci.*, vol. 425, pp. 62–75, Jan. 2018.

[15] Z. van Havre, N. White, J. Rousseau, and K. Mengersen, "Overfitting Bayesian mixture models with an unknown number of components," *PLoS ONE*, vol. 10, no. 7, p. e0131739, 2015.

[16] A. Reihanian, M. R. Feizi-Derakhshi, and H. S. Aghdasi, "Overlapping community detection in rating-based social networks through analyzing topics, ratings and links," *Pattern Recognit.*, vol. 81, pp. 370–387, Sep. 2018.

[17] S. Gao, Z. Yu, L. Shi, X. Yan, and H. Song, "A method to review expert recommendation using topic relevance and expert relationship," *Int. J. Cooperat. Inf. Syst.*, vol. 27, no. 1, p. 1741004, 2018.

[18] A. Ozcan and S. G. Oguducu, "Multivariate time series link prediction for evolving heterogeneous networks," *Int. J. Inf. Technol. Decision Making*, 2018, doi: 10.1142/S0219622018500530.

[19] A. Ozcan and S. Oguducu, "Link prediction in evolving heterogeneous networks using the NARX neural networks," *Knowl. Inf. Syst.*, vol. 55, no. 2, pp. 333–360, 2018.

[20] S. Davies and A. Moore, "Mix-nets: factored mixtures of Gaussians in Bayesian networks with mixed continuous and discrete variables," *Proc. 16th Conf. Uncertainty Artif. Intell. (UAI)*, San Francisco, CA, USA: Morgan Kaufmann, 2000, pp. 168–175.

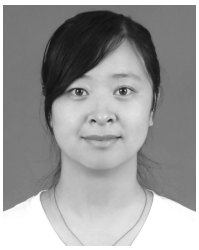
[21] A. A. Neath and J. E. Cavanaugh, "The Bayesian information criterion: Background, derivation, and applications," *Wiley Interdiscipl. Rev., Comput. Statist.*, vol. 4, no. 2, pp. 199–203, 2012.

[22] M. E. J. Newman and E. A. Leicht, "Mixture models and exploratory analysis in networks," *Proc. Nat. Acad. Sci. USA*, vol. 104, no. 23, pp. 9564–9569, 2007.

[23] D. M. Blei and J. D. Lafferty, "A correlated topic model of science," in *Proc. 23rd Int. Conf. Mach. Learn.* Cambridge, MA, USA: MIT Press, 2006, pp. 113–120.

ANPING ZHAO graduated from Southwest University, China. He received the Ph.D. degree in artificial intelligence from the Department of Computer and Information Science, Southwest University, in 2011. He was a Postdoctoral Researcher with the Artificial Intelligence Research Group, University of York, U.K., from 2015 to 2017. He is currently an Associate Professor with the Institute of Education Informatization, Wenzhou University. His main research interests include machine learning and text mining.

LINGLING ZHAO is currently pursuing the master's degree in computer science with Chongqing Normal University. Her main research interests include text mining and deep learning.



YU YU received the Ph.D. degree in logics from Southwest University, China, in 2010. In 2016, she was an Academic Visitor with the Department of Computer Science, University of York. She is currently an Associate Professor with the College of Teacher Education, Wenzhou University. Her main research interests include language logic and semantic analysis.