J Inf Process Syst, Vol.19, No.1, pp.55~66, February 2023
https://doi.org/10.3745/JIPS.04.0264

ISSN 1976-913X (Print)
ISSN 2092-805X (Electronic)

JOURNAL OF INFORMATION PROCESSING SYSTEMS JIPS

Research on Community Knowledge Modeling of Readers Based on Interest Labels

Kai Wang, Wei Pan, and Xingzhi Chen*

Abstract

Community portraits can deeply explore the characteristics of community structures and describe the personalized knowledge needs of community users, which is of great practical significance for improving community recommendation services, as well as the accuracy of resource push. The current community portraits generally have the problems of weak perception of interest characteristics and low degree of integration of topic information. To resolve this problem, the reader community portrait method based on the thematic and timeliness characteristics of interest labels (UIT) is proposed. First, community opinion leaders are identified based on multi-feature calculations, and then the topic features of their texts are identified based on the LDA topic model. On this basis, a semantic mapping including "reader community-opinion leader-text content" was established. Second, the readers' interest similarity of the labels was dynamically updated, and two kinds of tag parameters were integrated, namely, the intensity of interest labels and the stability of interest labels. Finally, the similarity distance between the opinion leader and the topic of interest was calculated to obtain the dynamic interest set of the opinion leaders. Experimental analysis was conducted on real data from the Douban reading community. The experimental results show that the UIT has the highest average F value (0.551) compared to the state-of-the-art approaches, which indicates that the UIT has better performance in the smooth time dimension.

Keywords

Interest Stability, Knowledge Modeling, Reader Community, Social Label, Topic Feature

1. Introduction

With the rapid development of Amazon Goodreads, Douban and other social networks, community reader-generated content (CRGC) has expanded rapidly, increasing the difficulty of reader recommendation, network marketing and other applications. Community knowledge modeling (CKM) abstracts the group features in social networks by clustering feature attributes of social relations in short labels [1]. Therefore, the realization of CKM is of practical significance for realizing the prediction of panoramic community structure features and seeking the construction of a pan scenario-oriented knowledge graph [2].

CKM studies in the early stage mainly focused on single-user clustering, whose focus was to build a seed user model based on the individual attribute information of network users. The user groups with higher similarity were aggregated by calculating the degree of integration among different users in the community. For example, a boundary aggregation method (TUPRP)-based on the self-labeled preference

Manuscript received June 16, 2022; first revision October 17, 2022; accepted November 28, 2022.
* **Corresponding Author:** Xingzhi Chen (1079132044@qq.com)
School of Health Management, Bengbu Medical College, Bengbu, China (wangkai0552@126.com, panweibbmc@163.com, 1079132044@qq.com)

information of users was proposed, which takes the mass classification as the core to set up a similar feature matrix by clustering the preference characteristics of groups [3]. The overall efficiency of CKM based on a single user is not high, and it is difficult to address medium and large CKM with complex social relations. With the rapid development of online social networking platforms, CKM technology oriented to multiuser information (KMMI) has gradually become a new research direction. The KMMI realizes multi-scene modeling based on multi-attribute characteristics, such as interest similarity, emotional dependence, and behavioral interaction of community members. The KMMI can be divided into three aspects. First, the community topic is proposed to describe the main content of the community. A Weibo hot topic-oriented community identification method (KCCD) was used to identify hot spot topics [4]. Meanwhile, Salehi et al. [5] proposed a method for identifying interest groups in healthy communities (TPUIG), which clustered user topics based on an author topic model. Second, community behaviors were utilized to build the community structure. A community user modeling method (CLH) based on a concept lattice was proposed, which clustered multilevel attribute features based on formal concept analysis [6]. In addition, Tang and Xie [7] proposed an online website user model (OLUP) that built a label system based on consumption behavior. Finally, community emotion was employed to enhance the effectiveness of CKM. Salehi [8] extracted a community knowledge modeling method (SDD), which realized the emotional portrait of community users by labeling semantic information. Liu [9] proposed a community-oriented modeling method for depressed users (TCNN-GRU), which utilized a deep learning model to process the classification information of depressed emotions. In addition, some scholars have described the characteristics of community knowledge from the perspective of multi-attribute feature fusion. For example, Shi et al. [10] proposed a label construction method (BILC) that integrated user behavior and interest tendency. Hu and Chen [11] proposed the user interest modeling method (UIM-LDA), which integrated the topic model with user interest by introducing a forgetting function. However, most methods fail to comprehensively consider the dynamic characteristics of the community, such as user interest transfer, behavior ambiguity and topic evolution, which cannot fully reveal the differences between community members and structures in the time dimension. Specifically, the above methods essentially achieve multi-scene modeling by selecting multi-attribute features based on community members' interest similarity, emotional dependence, and behavioral interaction, which tend to assume that the user interest label is static and unable to realize the dynamic update of user interest. Second, the above methods cannot effectively integrate the topicality and timeliness of reader labels, resulting in a low degree of topic information integration of user nodes.

To find the hidden pattern in the reader community, some of the interest characteristics of the reader labels based on LDA were found to construct the dynamic community. To fulfill this aim, the following proposals are made in this paper.

(1) A method was proposed for identifying the topic features of reader labels and extracting the text information of community opinion leaders for precise analysis of the topic features.
(2) A method was proposed to mark the interest characteristics of the reader labels and obtain the interest distribution matrix of the user labels in the time dimension.
(3) A method was proposed to calculate the similarity distance of reader labels to describe the dynamic modeling of the community.

To address the need above, a reader CKM method based on the thematic and timeliness characteristics of interest labels (UIT) was proposed, which combines social labels with the dynamic evolution of the

reader's interest. Specifically, the method was divided into three steps. The first step was to identify the topic features of reader labels and extract the text information of community opinion leaders. Afterward, the topic features of opinion leaders were described to establish the probability distribution matrix of user labels in the topic dimension. The second step was to identify the interest characteristics of the reader labels and obtain the interest distribution matrix of the user labels in the time dimension. Finally, the interest label set of the reader community was constructed to describe the dynamic modeling of the community based on the similarity calculation. In summary, the UIT reduces the dependence of CKM on social labels, to a certain extent, by identifying community opinion leaders and establishing target user groups. In addition, the semantic sparsity of reader labels was improved by smoothing the features of reader interest topics in the time dimension.

The remainder of this article is organized as follows: Section 2 carries out topic recognition in the reader community based on topological construction, identification of opinion leaders and LDA topic clustering. Section 3 contains an illustration of updating interest similarity to labels by integrating two kinds of label parameters. Section 4 includes the similarity calculation of reader community topics. Section 5 provides discussions of UIT in the Douban reading community, whereas Section 6 includes conclusions followed by acknowledgment and references.

2. Topic Recognition for Reader Community Labels

2.1 Topological Construction of Reader Community

A reader community is a virtual social network formed by a number of reader nodes around some book resources for thematic comments, content forwarding, behavioral praise and other activities. The dissemination of book resources in the reader community mainly depends on the behavior of users, such as forwarding, commenting and thumbs-up. The recognition of social relations based on readers' online behaviors can objectively reflect the topicality of online communities, thus realizing readers' interest recognition. Therefore, this paper divides readers' social behaviors into three categories: comments, forwarding and thumbs-up, which establish the connection relationship among commenters, carriers and subscribers. The objective is to determine the weight parameters of the above behaviors according to the community environment. The weight calculation of the relationships based on reader behavior is shown in Formula (1), where $W_{edg}(v_i, v_j)$ is the line weight between readers i, j; $\varepsilon_{forward}$, $\varepsilon_{comment}$, and $\varepsilon_{thumbs-up}$ are the weights of comments, forwarding and thumbs-up, respectively; $N_{forward}^{i,j}$, $N_{comment}^{i,j}$, and $N_{thumbs-up}^{i,j}$ are the numbers of comments, forwarding and thumbs-up between readers i, j, respectively; $N_{forward}^{all}$, $N_{comment}^{all}$, and $N_{thumbs-up}^{all}$ are the total numbers of comments, forwarding and thumbs-up by readers i, j, respectively.

$$W_{edg}(v_i, v_j) = \varepsilon_{forward} \cdot \frac{N_{forward}^{i,j}}{N_{forward}^{all}} + \varepsilon_{comment} \cdot \frac{N_{comment}^{i,j}}{N_{comment}^{all}} + \varepsilon_{thumbs-up} \cdot \frac{N_{thumbs-up}^{i,j}}{N_{thumbs-up}^{all}} \quad (1)$$

The reader community can be expressed as a directed graph (R, E, W), where R is the set of reader nodes, E is the set of links between reader nodes, and W is the set of weights on the links, which is used

to represent the closeness of social relations among different nodes. The specific process of topological construction for the reader community is as follows. First, community readers are taken as nodes to construct the reader node set $(v_1, \ldots, v_n)$. Afterward, the association relation between readers $\{W_{edg}(v_i, v_j), \ldots, W_{edg}(v_m, v_n)\}$ is established according to the social behavior information of readers, which is the weighted edge set between nodes.

2.2 Identification of Opinion Leaders

Opinion leaders are nodes with relatively high influence that play an important role in mediating and filtering public opinion dissemination. They usually spread network information to form the dissemination of information transmission. By identifying the community opinion leaders, the representative text information of their reader communities can be obtained. On this basis, we can describe the topic characteristics of opinion leaders in the community in the form of short texts with relatively little computational cost to establish the probability distribution matrix of user labels in the topic dimension. Therefore, the current paper takes the interest features of their labels as the keywords of reader communities by integrating non-textual features into the topology construction of reader communities, such as the social behaviors of opinion leaders, which reduces the computation cost of the probability distribution matrix in the topic dimension.

The influence of opinion leaders in the reader community can be quantified by environmental attributes and dynamic attributes. The environmental attributes reflect the static attributes and basic social relations of readers, mainly including the number of fans, friends and centrality of readers [8]. Thus, the environmental index calculation that defines reader *R* is shown in Formula (2).

$$E_t(R) = w_1^E {u_t}^{fan} + w_2^E {u_t}^{friend} + w_3^E {u_t}^{centrality} \tag{2}$$

In Formula (2), w_1^E, w_2^E, and w_3^E represent the weight ratio of the number of fans, the number of friends and the centrality in the environmental index of the reader node at time *t*, respectively. ${u_t}^{fan}$, ${u_t}^{friend}$, and ${u_t}^{centrality}$ represent the number of fans, the number of friends and the feature component of centrality of readers at moment *t*, respectively. The corresponding calculation is shown as ${u_t}^{fan} = (N_t^{fan} - N_{min}^{fan})/(N_{max}^{fan} - N_{min}^{fan})$, ${u_t}^{friend} = (N_t^{friend} - N_{min}^{friend})/(N_{max}^{friend} - N_{min}^{friend})$, and ${u_t}^{centrality} = (N_t^{centrality} - N_{min}^{centrality})/(N_{max}^{centrality} - N_{min}^{centrality})$. N_t^{fan}, N_t^{friend}, and $N_t^{centrality}$ represent the number of fans, the number of friends and the centrality of the reader node at moment *t*, respectively. N_{max}^{fan}, N_{max}^{fan}, and $N_{max}^{centrality}$ represent the maximum value of filament number, friend number and centrality of reader node at moment *t*, respectively.

The dynamic attribute of readers emphasizes leadership in information diffusion. Based on the PageRank algorithm [9], this paper assumes that reader *R* and $U_1 \ldots\ldots U_n$ are correlated, and the leadership index $L_t(R)$ of reader *R* is calculated as shown in Formula (3).

$$L_t(R) = (1-d) + d \cdot \left(\sum_{i=1}^{n} \frac{L_t(U_i)}{\sum_{k \subset friend(j)}^{|friend(j)|} L_t(U_j)}\right) \tag{3}$$

In Formula (3), $L_t(U_i)$ represents the influence weight of U_i on node *R*. *d* represents the damping coefficient. User influence is obtained by integrating the environmental attributes and the dynamic

attributes of readers. The calculation is shown in Formula (4). $INF_t(R)$ represents the user influence index of R.

$$INF_t(R) = E_t(R) + L_t(R) \tag{4}$$

2.3 Community-Node-Text Mapping

The opinion leaders are identified by calculating the influence index of the community nodes. The specific steps are as follows. First, the top-k opinion leaders in different reader communities are identified to match reader communities with opinion leaders according to the influence indices of different readers. Second, the ID of the opinion leader is used as the key word to obtain the text of the related resources, while the content under the same opinion leader is combined to establish the relationship mapping between the opinion leader and the content. Finally, the opinion leader ID is used to map the reader community ID to the dataset of the opinion leader to construct a reader-community-text triplet. In this way, the LDA model establishes a three-layer Bayesian network of document-topic-feature words by sampling the probability distribution of words in the corpus. Based on the LDA model [10], the label set is regarded as a document corpus, and the reader label is regarded as a feature word. The label topic matrix is established by calculating the probability distribution of the reader label and the topic.

Suppose $U = \{u_1, \dots, u_s\}$ is the first set of opinion leaders identified in a reader community, and $B = \{b_1, \dots, b_r\}$ is the set of book resources marked by opinion leaders. $L = \{l_1, \dots, l_t\}$ is the set of reader labels after text cleaning. When LDA topic recognition is carried out, the reader-community-text triad needs to be decomposed into a binary relation matrix $R(i, j)$ containing only the reader label, and the matrix element r_{ij} is the frequency of labeling l_j marked by readers u_i in a certain period. Second, the optimal number of topics K is determined by calculating the confusion degree under a different number of topics [11]. After the iteration, the label-topic distribution matrix $M(l_i, t_j)$ under different label sets is obtained, and the matrix element m_{ij} is the probability that label l_i belongs to topic t_j. Finally, the probability distribution of the labels l_i with readers u_i under different topics is obtained by combining the reader-label distribution matrix $R(i, j)$ and label-topic distribution matrix $M(l_i, t_j)$.

3. Label Interest Recognition in the Reader Community

Readers' interest in book resources can be reflected by the frequency of their access to relevant labels. In a period of time, the more frequently readers u_i visit label l_i, the higher the intensity of readers' interest [12]. Therefore, the reader's interest intensity in the relevant resources based on the frequency of labels is shown in Formula (5).

$$\text{Intensity}(u_i, l_j) = -\frac{1}{m}\sum_{k=1}^{m}\sum_{j=1}^{n} fre(u_i, l_j) \cdot log_2(fre(u_i, l_j)) \tag{5}$$

In Formula (5), m is the total number of labels contained in the set, n is the total number of labels marked by user u_i, and $fre(u_i, l_j)$ is the frequency of accessing labels l_j by user u_i.

The label interest stability calculation model is put forward on the basis of the dynamic evolution

characteristics of readers' interest [13], which assumes that the reader label matrix at t moment is influenced by which lies at t–1 moment. Meanwhile, the reader interested in forgetting factor of t–1 time can be set up. On this basis, the stability of label interest under adjacent time slices is calculated. Therefore, the stability of readers' interest in labels at time t is defined as $Stability_t^{u_i,l_j}$, as shown in Formula (6).

$$Stability_t^{u_i,l_j} = Stability_{t-1}^{u_i,l_j} e^{\frac{log\,2}{hl_{u_i}}(tagged_t - tagged_{t-1})} + \Delta Stability_t^{u_i,l_j} \tag{6}$$

In Formula (6), $Stability_{t-1}^{u_i,l_j}$ is the stability of readers' interest in labels l_j at moment t–1. hl_{u_i} is the half-life of the reader's interest, which is determined by the user's labeling cycle in the community. $tagged_t - tagged_{t-1}$ is the interval between two annotations for the label l_j. $\Delta Stability_t^{u_i,l_j}$ is the label increment at moment t under the reader's interest in the label increment. $\Delta Stability_t^{u_i,l_j} = fre_t^{u_i,l_j} / \sum_{j=1}^{m} fre_t^{u_i,l_j}$, $fre_t^{u_i,l_j}$ is the frequency of user u_i accessing label l_j at time t. $\sum_{j=1}^{m} fre_t^{u_i,l_j}$ is the frequency with which user u_i accesses all labels at time t.

In this paper, the interest recognition of reader labels mainly utilizes the label intensity of readers in a period of time, which is transformed into the characteristic prior of interest intensity at time t. Meanwhile, the weight of interest labels can be dynamically updated by comprehensively considering the label interest intensity and label interest stability. Therefore, the weight calculation of the interest label at time t is shown in Formula (7).

$$Weight_{u_i,l_j} = \delta \mathrm{Intensity}(u_i, l_j) + (1-\delta) Stability_t^{u_i,l_j} \tag{7}$$

In Formula (7), δ is the reader's interest intensity factor, and its value reflects the proportion of the reader's access frequency in the label weight.

4. Similarity Calculation of Reader Community Topic

4.1 Experimental Setup

The research process is mainly divided into the following specific steps. The text data of book reviews published by readers are collected on the Douban reader community platform to initialize the review database for post-text analysis. The initial data need to be preprocessed, including text word segmentation, removal of stopping words and pos tagging. A social network is constructed based on the relationships between feature words to realize the topology construction of the reader community and the identification of opinion leaders. The book reviews are clustered based on the LDA topic model to extract the topics. The similarity of opinion leaders under different interest topics is calculated by combining the topic label matrix and interest label matrix.

To achieve the above purpose, the operating environment of this experiment is Windows 10, and the memory capacity is 6 GB. The UIT uses Python 3.x to implement LDA topic identification. In addition, the software of UIT, including tools, is specified according to the corresponding experimental steps as follows. In the first stage, we used Pyspider, a web crawler system based on Python, to write scripts on the browser interface. Then, the crawling results were stored in the SQL database in real time. In the

second stage, the word segmentation software ICTCLAS was used to remove the stopping words. Meanwhile, the Jieba package in Python was used to perform the word segmentation of text messages. Afterward, Pajek, a social network analysis tool, was used to complete the reader community topology construction and realize the relationship clustering among nodes. In the final stage, the Python visualization tool LDAvis package was used to analyze the feature words under the topic.

4.2 Model Parameter Analysis

Reader behavior parameters mainly include forwarding behavior parameters, comment behavior parameters and upvoting behavior parameters. In the process of resource dissemination, the "VP vote" behavior indicates that readers have a strong sense of identity for labeling. The more times readers like the same hash label, the more likely it is that readers agree on their opinions. Therefore, the highest weight rating is given to this kind of behavior. The behavior of comment parameters is that readers express their attitudes by forwarding information containing labels to establish the correlation. Therefore, a medium weight is given to such behavior. Forwarding behavior refers to the social behavior in which readers express their feelings and opinions in response to a certain interest. The scope of influence is usually limited to a certain type of label, and the propagation influence of such behavior is limited. Therefore, this paper assigns a lower weight level to such behavior.

Attribute parameters mainly include the number of readers' fans, the number of friends and the weight proportion of centrality in the environment index, which reflects the influence degree of different feature components on the static attributes of readers. In the absence of environmental catastrophe, the evolution of the static attribute characteristics of opinion leaders is time-smooth [14]. Therefore, the attribute parameters of the above feature components are regarded as the growth rate of their characteristic values in unit time, while the value at time t is determined by calculating the growth rate of the number of readers' fans and friends and centrality at time t–1. The eigenvalue growth rate is calculated as $\alpha_t^w = \Delta f_{t-1}^w / f_{t-1}^w$. α_t^w is the eigenvalue growth rate of readers at time t. Δf_{t-1}^w represents the change in the reader's eigenvalue at time t–1. f_{t-1}^w represents the eigenvalue of the reader at time t–1.

The half-life of interest reflects the forgetting cycle of readers' acquisition, processing and dissemination of community knowledge, which determines the decay rate of reading interest in the time dimension [15]. Similarly, the attenuation degree of readers' interest in literature resources is associated with the label information generated by reading new literature. Therefore, the half-life of interest is defined as the interval between the current time and the time when the newer half of the labels are marked in a period of time. In this paper, "week" is taken as the basic measurement unit, and the validity period of the label is set as 12 weeks. In that case, the half-life of interest decay of readers is calculated. Finally, the expected half-life of interest is obtained as $E(hl_{u_i})$=1.75.

In general, the evolution of readers' interest is time dependent, which means that the intensity of interest at time t+1 will be affected by moment t. It shows that readers' interest in labels slowly decayed or gradually strengthened. Therefore, there should be a positive correlation between the reader's labeling behavior of interest labels at t and t+1 moment. By assigning various interest intensity factors (0, 0.2, 0.4, 0.6, 0.8, 1), the correlation coefficient of reader prediction labels at t and t+1 moment are calculated as $\rho_\delta = fre(Tag_t^n) / fre(Tag_{t+1}^n)$. $fre(Tag_t^n)$ represents the ratio of the total frequency of the first n labels, with high intensity marked by readers at time t to the total frequency of all labels. In the experiment, 100 readers with a higher half-life of interest are selected, and their values in adjacent time intervals are

calculated in 12 consecutive weeks. The correlation coefficients of the prediction labels are calculated by plugging them into Formula (13). The results show that when ρ_δ is 0.6, the average correlation coefficient is the largest, indicating that the prediction effect of reader labels is the highest at this time.

4.3 Community Topic Similarity Calculation

The calculation of community topic similarity was mainly divided into three parts. First, according to the identified opinion leaders, the labeling frequency was calculated, and the opinion leader-label matrix $R(i,j)$ was constructed. Then, the topic features were extracted to obtain the label-topic distribution matrix $M(l_i,t_j)$ based on the LDA topic model. Second, the opinion leader-interest label matrix $I(u_m,l_n)$ was obtained by integrating the interest intensity and interest stability to update the interest weight of the label dynamically. The matrix element i_{mn} represents the opinion leaders' interest value in the label l_n, and the inner product operation between matrix $R(i,j)$ and $I(u_m,l_n)$ was used to obtain the opinion leaders' topic matrix $O(u_m,t_j)$. The matrix element o_{mj} represents the probability that u_m is assigned to topic t_j. Afterward, the similarity distance between the opinion leader and the interest topic was calculated based on the similarity model shown in Formula (8), according to the set of interest topics I_{topic}. j represents the number of labels. n represents the number of opinion leaders; m represents the number of interest topics. Finally, the dynamic interest set of opinion leaders was obtained.

$$Sim(U,L)=\frac{\sum_{j=1}^{q}(u_{mj}\times l_{jn})}{\sqrt{\sum_{j=1}^{q}(u_{mj})^2}+\sqrt{\sum_{j=1}^{q}(l_{jn})^2}-\sum_{j=1}^{q}(u_{mj}\times l_{jn})} \tag{8}$$

5. Experimental Analysis

The experimental data came from the Douban reader Community (https://book.douban.com/). First, the web crawler was used to extract the text data of book reviews published by readers from January 1 to April 1, 2021, involving 70,648 readers and 2,328 book resources. There were 81,480 labels and 357,264 valid texts. Second, the data were classified and stored in the SQL database after preprocessing, Chinese word segmentation and other operations. It includes three tables, namely, the reader behavior table, static property table and dynamic property table. The reader behavior table records the text content published, reprinted, commented on and upvoted on by readers. The static property table records the number of readers' fans, friends and centrality. The dynamic property table records the text information associated with readers' forwarding and comments.

The topological construction of the reader community is to realize the clustering of relationships among nodes based on the social behaviors of readers. The specific process is as follows. First, the reader *ID* in the reader behavior table is used as the key word to retrieve the reader containing comments, forwarding and upvoting. Behavioral relationship mapping among readers is established. Second, the weight values of the behavior relations are marked to calculate the edge weights between reader nodes according to the weight grades of behavior parameters. Finally, the reader node table and weighted edge table are initialized with the social network analysis tool Pajek [16] to obtain the reader community, as shown in Table 1.

First, the parameter value of attributes at *t* moment is determined by combining the static attribute table and the dynamic attribute table, based on the growth rate of the number of fans, friends and centrality of readers at *t*–1 moment. Then, the influence of readers at different sampling moments is calculated by plugging it into Formulas (2)–(7). Afterward, 20% of *top-n* are selected as the final opinion leaders by combining the number of readers in different communities. The identification results of some opinion leaders in Community 105 are shown in Table 2.

Table 1. Douban reader community information

Sampling date	No. of communities	Community number	No. of nodes	No. of edges
1.15	10	105	2,629	30,156
		46	1,772	9,758
2.15	15	74	3,577	24,405
		13	768	5,607
3.15	12	273	3,008	16,721
		144	2,794	18,630

Table 2. Information of opinion leaders in Community 105

ID	Name	No. of label	Environmental index	Leadership index	Influence index
1541	Tiny Kitty	4,772	0.0672	0.1472	0.2144
2799	Poetry and piece	3,580	0.0359	0.1269	0.1628
582	Whistleword	6,403	0.0584	0.1037	0.1621
7890	Tipsy sunshine V	2,887	0.0617	0.0884	0.1501

According to the identified opinion leaders, the reader community *ID* is mapped to the dataset of opinion leaders. To address this need, the opinion leader label matrix is constructed, and the LDA topic model is recognized by Pathon language. The calculations show that the model has the least confusion when the number of topics is K=45. After 1,000 iterations, the probability distribution of the label topic-feature of Community 105 is obtained. According to the parameter analysis in Section 4.2, when hl_{u_i}=1.75 and δ=0.6, the weight of the interest label can be calculated accurately. Therefore, the above parameters are substituted to calculate the weight of the interest labels of the *top-n* opinion leaders in the community at moment *t*, as shown in Table 3.

The topic label matrix and interest label matrix of opinion leaders are substituted to obtain the similarity of opinion leaders under different interest topics. Afterward, the similarity distance of interest labels in different communities can be identified. Table 4 is a partial result of Community 105.

Table 3. Weight of opinion leaders' interest labels at *t* moment (Part)

ID	Interest label	Label intensity	Label stability	Interest weight
2799	movie	0.075	0.133	0.098
	novel	0.071	0.126	0.093
346	essays	0.063	0.123	0.083
	politics	0.051	0.114	0.076
1227	literature	0.082	0.145	0.107
	prose	0.068	0.132	0.094

Table 4. Interest label similarity of opinion leaders in Community 105

Opinion leader ID	Topics with similarities
2799	topic2/0.855 topic1/0.749 topic16/0.712 topic32/0.629 topic36/0.445
582	topic4/0.747 topic22/0.661 topic44/0.405 topic13/0.388 topic24/0.152
7890	topic20/0.654 topic7/0.582 topic10/0.346 topic38/0.225 topic41/0.187
3341	topic19/0.663 topic5/0.576 topic31/0.542 topic12/0.466 topic38/0.371
164	topic26/0.901 topic4/0.753 topic37/0.694 topic32/0.433 topic22/0.378

To verify the rationality of the UIT, the topological dataset of the community recorded for 5 consecutive weeks was named Dt_1 – Dt_5 based on the sampled data on January 15. Using the five-fold cross-validation method, four of them are successively taken as training datasets, and the rest are taken as test datasets. At the same time, the results of the UIT are compared with those of the TUPRP [3], KCCD [4], BILC [6], and UIM-LDA [7]. Afterward, the mean value of the five results is taken as the final value. The comparison of different algorithms on the F value is shown in Fig. 1.

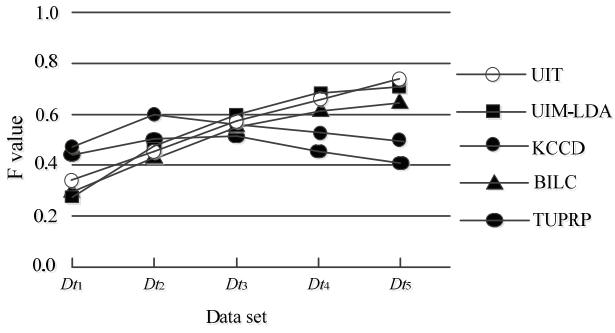

Fig. 1. Comparison of the model validation.

The following conclusions can be drawn from Fig. 1 in four aspects. First, although the TUPRP has a high recognition quality of readers' new interests on the dataset (D_{t1}, D_{t2}) at the initials of the window, the overall F value of the algorithm is the lowest. The reason is that the social relationship and interest drift between readers are not taken into account. Second, the reader model based on multiple users, including KCCD and BILC, performs better than TUPRP on the F value, but the improvement is limited. The main reason is that although KCCD extracts the multidimensional topic characteristics of readers to realize topic clustering, it does not integrate the behaviors between the reader features into CKM. BILC combines the behavior label with the interest label, which improves the modeling ability on the behavior interest feature, but it ignores the dynamic change in the reader's interest in the time dimension. Therefore, it is easy to distort the interest cluster. Third, UIM-LDA integrates the characteristics of the topic and user interests, which improves the effectiveness of CKM. However, it ignores the dependence between user behavior and community structure, thus cutting off the intrinsic semantic association between different datasets. Finally, the UIT has the highest average F value (0.551) in the five datasets. The main reason is that the UIT can jointly model community structure, reader influence and interest topics in the smoothing time dimension by introducing model parameters, such as reader parameters, interest half-life, and interest intensity factors. Hence, the accuracy and timeliness of group modeling are improved in adjacent time slices.

6. Summary

This paper proposes a reader CKM method (UIT) based on user interest labels by extracting the topicality and timeliness of reader labels. The model combines the dynamic labeling of reader interest with the LDA topic model and obtains the set of interest labels of the reader community by calculating the reader topic similarity distance to realize the knowledge discovery of the reader community.

In general, the following improvements can be obtained in contrast to the state-of-the-art approaches. (1) The UIT combines dynamic labeling of readers' interests with the LDA topic model to characterize the dependence association between implicit topic features of communities and readers' interests, which can dynamically perceive the changes in readers' interest intensity and interest stability to realize the topic description of users' interest in the time dimension. (2) The UIT obtains the interest labels of the reader community by calculating the topic similarity distance at different moments. On this basis, the influence of opinion leaders is used as a feature vector to quantify the environmental attributes and dynamic attributes of readers. Therefore, it is helpful to predict the characteristics of the panoramic community structure in seeking pan scene knowledge graph construction, which is an important and central contribution of this study. The later research direction of this paper can improve the extensibility of the model in community differentiation modeling regarding aspects of community location and communication modeling.

Acknowledgement

This work is supported by the Key Project of Humanities and Social Science in Anhui Education Department (Grant No. 2022AH051412 and 2022AH051405), the Major Humanities and Social Science Project of Education Department of Anhui Province (Grant No. SK2021ZD0066), and the Key Project of Humanities and Social Sciences of Bengbu Medical College (Grant No. 2020byzd223sk).

References

[1] L. Liu, S. Wang, and Z. Hu, "A literature review on community profiling," *Library and Information Service*, vol. 63, no. 23, pp. 122-130, 2019.

[2] L. Sun, "Approach of multi-source competitive intelligence fragments fusion based on intelligence element similarity," *Information Studies: Theory & Application*, vol. 41, no. 10, pp. 8-14, 2018.

[3] Y. Cai and Q. Li, "Personalized search by tag-based user profile and resource profile in collaborative tagging systems," in *Proceedings of the 19th ACM International Conference on Information and Knowledge Management*, Toronto, Canada, 2010, pp. 969-978.

[4] S. Ding, N. Wang, and J. C. Wu, "Hot topic detection of Weibo based on keyword co-occurrence and community discovery," *Modern Information*, vol. 38, no. 3, pp. 10-18, 2018.

[5] A. Salehi, M. Ozer, and H. Davulcu, "Sentiment-driven community profiling and detection on social media," in *Proceedings of the 29th on Hypertext and Social Media*, Baltimore, MD, 2018, pp. 229-237.

[6] R. Wang and W. Zhang, "Behavior and interest labeling construction and application of academic user portraits," *Modern Information*, vol. 39, no. 9, pp. 54-63, 2019.

[7] X. Tang and L. Xie, "Construction and dynamic update of theme-based user interest model," *Information Studies: Theory & Application*, vol. 39, no. 2, pp. 116-123, 2016.
[8] H. Li and Y. Liang, "Time series clustering method with label propagation based on centrality," *Control and Decision,* vol. 33, no. 11, pp. 1950-1958, 2018.
[9] F. Chung and A. Tsiatas, "Finding and visualizing graph clusters using PageRank optimization," *Internet Mathematics*, vol. 8, no. 1-2, pp. 46-72, 2012.
[10] J. Shi, M. Fan, W. L. Li, "Topic analysis based on LDA model," *Acta Automatica Sinica*, vol. 35, no. 12, pp. 1586-1592, 2009.
[11] J. Hu and G. Chen, "*Mining and evolution of content topics based on Dynamic LDA*," Library and Information Service, vol. 58, no. 2, pp. 138-142, 2014.
[12] J. A. Nunez, P. M. Cincotta, and F. C. Wachlin, "Information entropy," *Celestial Mechanics and Dynamical Astronomy*, vol. 64, pp. 43-53, 1996.
[13] G. Zhu and L. Zhou, "Hybrid recommendation based on forgetting curve and domain nearest neighbor," *Journal of Management Sciences in China*, vol. 15, no. 5, pp. 55-64, 2012.
[14] Y. Liu, K. Wang, and Y. Liu, "Online recognition approach for opinion leaders using influence heredity," *Information Studies: Theory & Application*, vol. 42, no. 7, pp. 126-131, 2019.
[15] S. Zhu and X. Jiang, "Analysis of Literature obsolescence for humanities and social sciences journals based on CSSCI data," *Journal of the China Society for Scientific and Technical Information*, vol. 36, no. 10, pp. 1031-1037, 2017.
[16] W. Meng and J. Pang, "Application of Pajek in visualization of coauthored networks in information science," *Information Studies: Theory & Application*, vol. 31, no. 4, pp. 573-575, 2008.

Kai Wang **https://orcid.org/0000-0002-9098-3017**

He graduated from Anhui Agricultural University with a master's degree in 2011. He is currently a lecturer in Bengbu Medical College, whose current research focuses on information processing and information retrieval.

Wei Pan **https://orcid.org/0000-0003-0444-973X**

He graduated from Jilin University with a doctor degree in 2015. He is currently an associate professor in Bengbu Medical College, whose current research focuses on data mining and information integration.

Xingzhi Chen **https://orcid.org/0000-0003-3703-5306**

He graduated from Bengbu Medical College with a bachelor's degree in 1992. He is currently a professor in Bengbu Medical College, whose current research focuses on data mining and social computing.