Multi-grained alignment method based on stable topics in cross-social networks (#93969)

First submission

Guidance from your Editor

Please submit by **3 Jan 2024** for the benefit of the authors (and your token reward) .

Structure and Criteria
Please read the 'Structure and Criteria' page for general guidance.

Raw data check
Review the raw data.

Image check
Check that figures and images have not been inappropriately manipulated.

If this article is published your review will be made public. You can choose whether to sign your review. If uploading a PDF please remove any identifiable information (if you want to remain anonymous).

Files
Download and review all files from the materials page.

7 Figure file(s)
3 Table file(s)
2 Raw data file(s)

Structure and Criteria

Structure your review

The review form is divided into 5 sections. Please consider these when composing your review:

1. **BASIC REPORTING**
2. **EXPERIMENTAL DESIGN**
3. **VALIDITY OF THE FINDINGS**
4. General comments
5. Confidential notes to the editor

You can also annotate this PDF and upload it as part of your review

When ready submit online.

Editorial Criteria

Use these criteria points to structure your review. The full detailed editorial criteria is on your guidance page.

BASIC REPORTING

- Clear, unambiguous, professional English language used throughout.
- Intro & background to show context. Literature well referenced & relevant.
- Structure conforms to PeerJ standards, discipline norm, or improved for clarity.
- Figures are relevant, high quality, well labelled & described.
- Raw data supplied (see PeerJ policy).

EXPERIMENTAL DESIGN

- Original primary research within Scope of the journal.
- Research question well defined, relevant & meaningful. It is stated how the research fills an identified knowledge gap.
- Rigorous investigation performed to a high technical & ethical standard.
- Methods described with sufficient detail & information to replicate.

VALIDITY OF THE FINDINGS

- Impact and novelty not assessed. *Meaningful* replication encouraged where rationale & benefit to literature is clearly stated.
- All underlying data have been provided; they are robust, statistically sound, & controlled.
- Conclusions are well stated, linked to original research question & limited to supporting results.

Standout reviewing tips

The best reviewers use these techniques

Tip	*Example*
Support criticisms with evidence from the text or from other sources	*Smith et al (J of Methodology, 2005, V3, pp 123) have shown that the analysis you use in Lines 241-250 is not the most appropriate for this situation. Please explain why you used this method.*
Give specific suggestions on how to improve the manuscript	*Your introduction needs more detail. I suggest that you improve the description at lines 57- 86 to provide more justification for your study (specifically, you should expand upon the knowledge gap being filled).*
Comment on language and grammar issues	*The English language should be improved to ensure that an international audience can clearly understand your text. Some examples where the language could be improved include lines 23, 77, 121, 128 – the current phrasing makes comprehension difficult. I suggest you have a colleague who is proficient in English and familiar with the subject matter review your manuscript, or contact a professional editing service.*
Organize by importance of the issues, and number your points	*1. Your most important issue* *2. The next most important item* *3. …* *4. The least important points*
Please provide constructive criticism, and avoid personal opinions	*I thank you for providing the raw data, however your supplemental files need more descriptive metadata identifiers to be useful to future readers. Although your results are compelling, the data analysis should be improved in the following ways: AA, BB, CC*
Comment on strengths (as well as weaknesses) of the manuscript	*I commend the authors for their extensive data set, compiled over many years of detailed fieldwork. In addition, the manuscript is clearly written in professional, unambiguous language. If there is a weakness, it is in the statistical analysis (as I have noted above) which should be improved upon before Acceptance.*

Multi-grained alignment method based on stable topics in cross-social networks

Jing Lu [1] , **Qikai Gai** Corresp. [1]

[1] School of Optical-electrical and Computer Engineering, University of Shanghai for Science and Technology, Shanghai, China

Corresponding Author: Qikai Gai
Email address: gaiqikai@163.com

Cross-social network user alignment is divided into user alignment and group alignment. In the user alignment mode, it is difficult to obtain user features fully due to social network privacy protection. In contrast, in the group alignment mode, the alignment accuracy is low due to the existence of many edge users. In order to solve this problem, this research proposes a multi-grained alignment method (MGA). In MGA, First, stable topics are obtained from user-generated content based on topic time jitter and embedded, and the weight of user edges is updated using vector distances. An improved Louvain algorithm ST-L (Stable Topic-Louvain) is designed to complete multi-level community detection without predetermined tags, obtain fuzzy community topic features, and complete cross-social network community alignment. Iterative alignment is executed from coarse-grained communities to fine-grained sub-communities until user-level alignment. The process can be stopped at any layer to achieve multi-granularity alignment, which solves the problem of low accuracy of edge user alignment at a single granularity and improves the accuracy of user alignment. Experiments on real datasets show the effectiveness of our method.

Multi-grained alignment method based on stable topics in cross-social networks

Jing Lu , Qikai Gai*
School of Optical-electrical and Computer Engineering, University of Shanghai for Science and Technology, Shanghai , China
Correspondence Qikai Gai: gaiqikai@163.com

Abstract: Cross-social network user alignment is divided into user alignment and group alignment. In the user alignment mode, it is difficult to obtain user features fully due to social network privacy protection. In contrast, in the group alignment mode, the alignment accuracy is low due to the existence of many edge users. In order to solve this problem, this research proposes a multi-grained alignment method (MGA). In MGA, First, stable topics are obtained from user-generated content based on topic time jitter and embedded, and the weight of user edges is updated using vector distances. An improved Louvain algorithm ST-L (Stable Topic-Louvain) is designed to complete multi-level community detection without predetermined tags, obtain fuzzy community topic features, and complete cross-social network community alignment. Iterative alignment is executed from coarse-grained communities to fine-grained sub-communities until user-level alignment. The process can be stopped at any layer to achieve multi-granularity alignment, which solves the problem of low accuracy of edge user alignment at a single granularity and improves the accuracy of user alignment. Experiments on real datasets show the effectiveness of our method.

Keywords: social network; stable topic; community detection; multi-grained user alignment; fuzzy characteristic

Citiations are not used appropriately, it seems random article cited such as line 28, [1] is unnecessary , line 33, [2] cites what?

0 Introduction

Social networks play an important role in daily life, such as music sharing, game services, dynamic sharing, etc[1]. Identifying the same users in different social networks can improve social experience, increase social efficiency, and make advertising and business recommendations more accurate. People finding services can be provided by aggregating publicly visible user information from multiple online social networks by linking users' accounts in different social networks and analyzing their behavior on multiple online social networks to detect and find illegal users who are trying to steal users' private[2]At present, cross-social network alignment can be divided into two modes: user alignment and group alignment.

Line 30-34, sentence is very long and incomplete , ambiguous

User alignment is based on user attributes, behaviors, and topology information. The literature[3] is matched based on the user's profile (user name, age, profile, educational background), user-generated content, and location attributes by plugging them into the fusion classifier of machine learning. The literature[4] uses the BP neural network to calculate the similarity of the user name by vector mapping. The alignment method based on user attributes faces the problems of limited data acquisition and high data processing difficulty. A joint framework of behavior analysis and social network alignment is proposed in the literature[5]. The principle is to obtain comprehensive user behavior information through a user behavior (forwarding, comment, location punching) information fusion algorithm based on earth-moving distance (EMD), verify user behavior prediction, and complete user alignment based on behavior characteristics. However, user behavior information is complex and highly sparse, making it difficult to find potential consistent patterns from it. The literature[6] uses anchor users to walk outward to obtain local topological maps around anchor users. It performs common maximum subgraph similarity calculation or probabilistic generating graph calculation to complete alignment[7]However, this kind of matching method for pure topology structure relies too much on the influence of anchor users, and the location of anchor points in the network is particularly important. If it is too dense, the local matching calculation will converge quickly, fall into local

Line 35, do you mean article [3] has studied ....

Line 38, BP is never defined

optimal, and cannot be spread to distant locations; if it is too sparse, the connection between local and local areas will be weakened, the matching calculation amount will be larger, and the accuracy will be reduced. Line 51, statement needs a reference

Group alignment is usually combined with user interest. The literature[8] analyzes the topics of users' published content through the Dirichlet model. It combines the method of hypertext topic search (HITS)[9]The literature[10] proposes that the interest should be divided into temporary interest and core interest, reduce the interference of temporary interest and effectively improve the alignment accuracy; the literature[11]The literature[12] uses user interest labels for weight scoring and stratified community classification. Still, the label classification is fixed, and the same user is assigned to multiple label classes, resulting in redundancy. The deep learning model (ReinCom) was successfully used in the literature[13] for the first time to map nodes and communities to the hyperbolic space to learn the degree of coupling between nodes and communities. However, because the input and output parameters of a hierarchical community tree are difficult to define as integers and one-hot vectors, the updating and optimization parameters of a community tree are variable, and the effect is not good when facing larger communities.

line 56-60 is ambigious and badily written

To sum up, most scholars focus on one of the two matching modes of groups or users, and data acquisition is often limited in the user alignment mode. In contrast, in the group alignment mode with determined granularity, there are too many edge users and only reflect the overall characteristics of a group, and accurate user alignment cannot be completed by using this result. At the same time, the single-layer community detection cannot accurately reflect the user interest relationship, and the fixed-label hierarchical detection will cause the same user to assign multiple communities, resulting in redundancy and low efficiency. Line 67-73, how these conclusions are made, please use references

Contributions are as follows: line 74 - please connect the following section to previous, if it is literature review then it is not sufficient include some latest studies and provide analysis table

(a) A stable topic acquisition method is proposed. The original keywords of the content published by users are mined and the original theme sequence is formed by clustering. Calculate the jitter degree of each theme according to the time slice and sort it, eliminate the short-time theme, obtain the stable theme sequence, embed the user's theme sequence into the Word2Vec model vector space, and calculate the user's theme similarity.

(b) Proposed multi-level community detection method ST-L. The topic similarity is used to update the user edge weights, and the Louvain algorithm is improved based on the weights to optimize the aggregation and re-optimize to the upper layer. The multi-level community detection without presetting fixed hierarchical labels is completed.

(c) An improved fuzzy center clustering method is proposed to obtain the community theme features. By updating and optimizing the correlation degree matrix between users and theme feature points, the theme features of communities and user groups are obtained. Starting from coarse-grained communities, the theme feature similarity across social networks of the same granularity is calculated to complete alignment. Vertical iteration is performed until user matching is performed. The alignment results of each layer can be output to achieve multi-granularity alignment.

(d) MGA's multi-granularity alignment method based on the combination of the two alignments on the real data set has better performance indicators.

1 Basic Definition

1.1 Symbol definition

Symbolic Definition 1 Social network. Define the social network as $G=(V,E,P)$ V for the collection of user nodes to be aligned $V=(u_1,u_2,\ldots u_i,\ldots u_R)$, and i for the user number. E stands for powerless edge set $E=\{e_{ij}\mid u_i,u_j\in V\}$ and e_{ij} stands for users u_i and edges; that is, there is a following or being followed relationship between users; that is, there is a connected edge. u_j P stands for the content published by the user, P=(user, published content, published time), indicating the content set published by the user. P_i u_i For the convenience of subsequent modeling and calculation, the source social network and target social network are selected and denoted as and respectively. $G=(V,E,P)$ $G'=(V',E',P')$

Define 2 Thematic hierarchical communities. Define a topic layered community as where $C=(c_1^{00},c_2^{01},\ldots c_1^{10},\ldots c_i^{jh},\ldots c_R^{L})$ i represents the community to which the user belongs, u_i j represents the layer of the community, h represents the number of the community, and when j=0, then the community is called a user group.

Define 3 User topics. Define the user topic as and the topic weight as, where the topic weight subscript corresponds to the topic subscript. $topic = (topic_1, topic_2, \cdots topic_R)$ $W = (w_1, w_2, \cdots w_R)$

Define 4 Anchor users. An anchor user is someone who has an account on both the source social network and the target social network and has the account aligned, authenticating high-impact users by name, etc. G G'

Define five thematic characteristics. The fusion topic feature of the user group and community is defined as representing the overall topic feature of the user group or the internal members of the community in layer j numbered h, where represents the feature point of the middle topic, z is the number of the feature point. The dimension of the feature point should be the same as the dimension of the user topic vector. $V_{c^{jh}}$ $CTopic_z^{V_{c^{jh}}}$ $V_{c^{jh}}$

1.2 Problem definition

For the source social network, find the target social network and use the stable theme to detect and find the user group and the high-level community and obtain the theme feature sum from the coarse-grained high-level community to the fine-grained community to the user-level alignment.

$u \in G$ $u' \in G'$ u u' c_u^{jh} $c_u^{jh}{}'$ $V_{c^{jh}}$ $V_{c^{jh}}{}'$ u u'

2 MGA: Stable topic multi-granularity alignment method

2.1 Method Framework

The process of the MGA method is shown in Figure 1:

Step 1: For source network and target network users, calculate the time jitter degree of the theme from the original topic, filter the unstable hot topic, obtain the stable topic, embed it in the same vector space, and calculate the similarity of the user topic.

Step 2: Update the user side weights according to the user theme similarity, improve the Louvain algorithm without weights, conduct community detection, iterate from low level to high level, and form a hierarchical community structure without preset fixed labels.

Step 3: Improve the fuzzy center clustering algorithm to extract the theme features of user groups or communities by optimizing the correlation degree matrix and theme feature points. According to the similarity of theme features, cross-social network alignment is carried out from coarse-grained to fine-grained, iterative to user-level alignment, and the alignment results of each layer can be used as output.

Figure 1: Flow chart of MGA

2.2 Stable topic extraction

2.2.1 Original theme extraction

For the blog sets published by users, the TF-IDF model is adopted to extract the sequence of users' interest keywords, and K-means clustering[14]

$$A_{u_i} = \begin{pmatrix} w_1 \cdot topic_1^{u_i} \\ w_2 \cdot topic_2^{u_i} \\ \vdots \\ w_k \cdot topic_k^{u_i} \end{pmatrix} \qquad (1)$$

2.2.2 Filter short-time hot topics

Users are extremely vulnerable to the impact of hot events, and their attention to hot events is relatively concentrated[10]Therefore, this paper will find and filter short-time hot topics with large jitter degrees according to the time slice.

First of all, the weights of the medium are decomposed based on the total number of time slices A_{u_i} $topic_j$ w_j T and the length of each slice t, and the weights of the user topic in each time slice are obtained. The time jitter *St* of the user topic is defined as follows:

$$St_{topic_j}^{u_i} = \frac{1}{T} \times \sum_{t=1}^{T} \left(w_{j,t}^{u_i} - \frac{1}{T} \times \sum_{t=1}^{T} w_{j,t}^{u_i} \right)^2 \qquad (2)$$

Where $w_{j,t}^{u_i}$ represents the weight of the user's topic in the u_i t time slice, T represents the total number of time slices, and t represents the number of time slices.

According to formula (2), the time jitter of all the *user topic topics* is obtained, which is arranged in ascending order. The higher the degree of the topic is, the lower it is, and it is the short-term hot interest that needs to be cut. Select the top *kt* topics and cut the topics with a large jitter degree. For the convenience of statistics, the weight of the reduced unstable topics is redistributed in proportion. Finally, the user topic matrix is obtained as shown in formula (3), and the filtering process is shown in algorithm 1:

$$A_{u_i}' = \begin{pmatrix} w_1' \cdot topic_1^{u_i} \\ w_2' \cdot topic_2^{u_i} \\ \vdots \\ w_k' \cdot topic_{kt}^{u_i} \end{pmatrix} \tag{3}$$

Algorithm 1 Filters the user's short-time topic algorithm

Enter: A_{u_i} , $w_{j,t}^{u_i}$

Output: A_{u_i}'

a) St(j)← $St_{topic_j}^{u_i}$,topic[]← A_{u_i}

b) For topic[j] in topic[] do

c) For t in T do

d) St(j)←St(j)+()// Calculate jitter $w_{j,t}^{u_i} - avg(w_{j,t}^{u_i})$

e) St(j)←St(j)/T

f) Order by St asc

g) For j in (kt,k] do

h) Then for q in [0,kt] do

i) w_q ←*(1+) // Assign cut edge weight w_q $w_q * w_j$

j) Update $w_{q'}$ ← w_q

k) Return A_{u_i}'

2.2.3 User topic similarity calculation

Extract the user u_i topic matrix and embed the A_{u_i}' Word2Vec[15] u_i V_{u_i}

$$V_{u_i} = w_1' \cdot \overrightarrow{topic_1^{u_i}} + w_2' \cdot \overrightarrow{topic_2^{u_i}} + \cdots w_k' \cdot \overrightarrow{topic_k^{u_i}} \tag{4}$$

According to the theme vector sum of user and (,) obtained u_i from formula (4) u_j, the similarity of the theme vector is calculated to obtain the user theme similarity, as shown in formula (5) : u_i $u_j \in U$ V_{u_i} V_{u_j} $Sim\_Topic(u_i, u_j)$

$$Sim\_Topic(u_i,u_j) = \frac{V_{u_i} \cdot V_{u_j}}{\left|V_{u_i}\right| \cdot \left|V_{u_j}\right|} \tag{5}$$

2.3 ST-L: User-stable topic hierarchical detection algorithm

2.3.1 Data preprocessing

For social network *G*, there is an edge between users and users in the original network diagram, which only means that there is a concerned relationship between two users and cannot reflect the degree of correlation between the two subjects. u_i u_j In order to better detect the hierarchical community of user topics, this paper converts user edge *E* into weighted edge, takes the user topic similarity calculated by formula (5) as the edge weight of users and users, and then transforms social network ΔE $Sim\_Topic(u_i,u_j)$ u_i u_j G into a weighted undirected network graph of user topics.

2.3.2 User topic hierarchical community detection method

Louvain[16] algorithm is an excellent hierarchical community detection method, which has excellent performance in large-scale complex networks, but the user without rights affects its effect in special application scenarios. In order to better detect the hierarchical community of topics, this paper updates the user edge weights according to the similarity of users' topics. It proposes the Louvain algorithm ST-L based on stable topics. The algorithm is divided into two parts: community optimization and community aggregation. The algorithm judges which community is more suitable for the nodes in the network by the community modularity. The definition of community modularity Q is shown in formula (6) :

$$Q = \frac{1}{2m} \cdot \sum_{u_i,u_j \in U} \left[e'_{ij} - \frac{d_i \cdot d_j}{2m} \right] \delta\left(c_i,c_j\right) \tag{6}$$

Where, otherwise, is the topic similarity between users and users, and represents the sum of the theme similarity of all users connected with and respectively, represents the user group and belongs to, $c_i = c_j$ $\delta\left(c_i,c_j\right)=1$ $\delta\left(c_i,c_j\right)=0$ e'_{ij} u_i u_j d_i d_j u_i u_j c_i,c_j u_i u_j m represents the sum of the similarity between all users in the network, the value is in, the larger the value is, the better the structure. $[-1,1]$

The core of the algorithm is to compare the modularity of a user before and after joining a community, and the modularity increment is derived from formula (6), as shown in formula (7) : u_i c_i ΔQ

$$\begin{aligned}\Delta Q &= \left[\frac{d_{in}+d_{i,in}}{2m}-\left(\frac{d_{ci}+d_i}{2m}\right)^2\right]- \\ &\left[\frac{d_{in}}{2m}-\left(\frac{d_{ci}}{2m}\right)^2-\left(\frac{d_i}{2m}\right)^2\right] \\ &=\frac{1}{2m}\left(d_{i,in}-\frac{d_{in}\cdot d_{i,in}}{m}\right)\end{aligned} \quad (7)$$

Where is the sum of the similarity of users in the community, the sum of the similarity of users connected to other users in the community, and the sum of the similarity of all users connected to each? d_{in} c_i $d_{i,in}$ u_i c_i d_{ci} c_i

After obtaining the modularity increment, community optimization can be carried out; that is, a different user group is initialized for each user of the network. ΔQ When initializing partitions, the number of user groups is as large as the number of users. Then, for each user and its neighbor users, the modularity increment is calculated by placing it in the user group that will be removed from its own user group and placed in the user group with the largest modularity increment. u_i u_j u_i u_j u_i If so, stay in the original user group. $\Delta Q < 0$ Repeat and apply this process sequentially for all users (the ordering of users has no significant effect[16] on the obtained modularity increments) until complete.

As shown in Figure 2, suppose seven user groups c^{0x} are obtained from the first part of the algorithm (x is the group serial number) c^{01}: {1,3,9,10,15}, c^{02}: {2, 6, 12}, c^{03}: {4,5,11,13,17}, c^{04}: {7,8,14,16}, each number represents a user, the edge indicates that there is a relationship between users to be followed or to be followed. The users in the same dotted box are considered to be divided into the same group, representing that the user theme is similar. The structure in groups c^{05}, c^{06} and c^{07} and the connection relationship with other groups are omitted for the sake of simplification.

Figure 2: Examples of single-layer community detection

The next step is group aggregation; that is, the user group obtained in the first part is iterated upward, and the weight of the edge between them is the sum of the weight of the edge between the nodes in the corresponding two groups. Then, the algorithm of the first part is iterated on this new graph. After the algorithm of the first part is completed, the second part can be used to generate a higher-level structure graph. MGA adopts a three-layer structure, and the results of the three-layer top are sparse**Error! Reference source not found.**

As shown in Figure 3, after the completion of the first part of the algorithm, the second part of the algorithm is iterated to the upper layer. The first iteration is as follows: user group c^{0201},c constitutes community c^{11}; group c^{03},c,c^{04} constitutes community c^{05}; group c^{12},c constitutes community c^{06}; and group c^{07},c constitutes community C^{13}. The second iteration of community c^{11}, c^{12}, c^{13} constitutes community c^{21}.

Figure 3: Examples of multi-layer community detection

2.4 Multi-granular community alignment is aligned with users

2.4.1 Topic Feature Extraction

For the communities in the source social network and the target social network, the two communities can be regarded as the anchor community G G' [17] G G'

After completing the first round of community detection, users form a user group and continue to conduct community detection and aggregate into the upper community. In this process, each group will produce a group theme feature because it is close to the group member theme vector, the user theme vector is also very close to the community theme feature points, and there are a large number of edge points, which are in the middle of the two feature points. If it is directly forced to be classified into one class and the feature points are updated at this distance, the accuracy will be reduced. $V_{c^{jh}}$ The fuzzy Center-Based Clustering (FCM) algorithm[18]Therefore, this paper introduces the correlation degree matrix between the user theme vector and the community feature points, as shown in formula (8) :

$$S^{\beta}=\begin{pmatrix} d\left(V_{u_1},CTopic_1^{V_1^{jh}}\right),\cdots\ d\left(V_{u_1},CTopic_z^{V_1^{jh}}\right) \\ d\left(V_{u_2},CTopic_1^{V_2^{jh}}\right),\cdots\ d\left(V_{u_1},CTopic_z^{V_2^{jh}}\right) \\ \vdots \\ d\left(V_{u_a},CTopic_1^{V_a^{jh}}\right),\cdots\ d\left(V_{u_a},CTopic_z^{V_a^{jh}}\right) \end{pmatrix} \tag{8}$$

Optimize and update the correlation matrix by gradient descent algorithm to obtain the best features. $V_{c^{jh}}$ It is a multi-dimensional topic vector, which can represent the theme features of a whole group or community to the greatest extent. $V_{c^{jh}}$ The definition is shown in formula (9) :

$$V_{c^{jh}}=\begin{pmatrix} CTopic_1^{c_i^{jh}} \\ CTopic_2^{c_i^{jh}} \\ \vdots \\ CTopic_z^{c_i^{jh}} \end{pmatrix} \tag{9}$$

In order to achieve the optimal, construct the objective function as shown in formula (10) :

$V_{c^{jh}}$

$$f(\beta)=\begin{cases} \sum_{a=1}^{N}\sum_{b=1}^{Z} S_{ab}^{\beta}\left\|V_{u_a}-CTopic_b^{c^{jh}}\right\|^2 \\ \sum_{b=1}^{Z} S_{ab}^{\beta}\geq 1 \end{cases} \tag{10}$$

Where is the weighted constant, user, is the topic vector of the user, is a topic feature point in the topic feature, β $u_a(1\leq a\leq N)\in c^{jh}$ V_{ua} u_a $CTopic_1^{c_i^{jh}}$ c^{jh} $V_{c^{jh}}$ b is the number of the feature point, indicating the correlation degree between the user's topic vector and the feature point, the problem is transformed into a minimum problem. S_{ab}^{β} $f(\beta)$ Adjust the sum of the correlation degree of each user's theme vector to the condition that it is greater than or equal to 1, and arrange it to get the formula (11) as shown: u_a V_{ua} $CTopic_1^{c_i^{jh}}$

$$f(\beta)=\sum_{a=1}^{N}\sum_{b=1}^{Z}S_{ab}^{\beta}\left\|V_{u_a}-CTopic_b^{c^{jh}}\right\|^2+\sum_{b=1}^{Z}\mu_b\left(1-\sum_{a=1}^{N}S_{ab}^{\beta}\right) \tag{11}$$

Where μ_b is any coefficient, and the partial derivative of formula (10) and respectively is equal to 0, and its iterative equation is obtained as shown in formula (12), formula (13) : S_{ab}^{β} $CTopic_b^{c^{jh}}$

$$S_{ab}^{\beta}=\left(\sum_{k=1}^{Z}\left(\frac{\left\|V_{ua}-CTopic_b^{c^{jh}}\right\|}{\left\|V_{ua}-CTopic_k^{c^{jh}}\right\|}\right)\right)^{-1} \tag{12}$$

$$CTopic_b^{c^{jh}}=\frac{\sum_{a=1}^{N}S_{ab}^{\beta}\cdot V_{ua}}{\sum_{a=1}^{N}S_{ab}^{\beta}} \tag{13}$$

Where k is the number of iterations. When iteration is stopped, it means that the iterative change of the user theme vector and the correlation degree of feature points is small, and no further optimization is required. The overall algorithm is shown in Algorithm 2: $\nabla S_{ab}^{\beta}<\varepsilon$

Algorithm 2 topic feature generation algorithm

Input: V_{ua} (User $u_a\left(1\le a\le N\right)\in c^{jh}$)

Output: $V_{c^{jh}}$

a) Initialize V=[$V_{c^{jh}}$] matrix ,V(0)

b) Calculate S=[S_{ab}^{β}] matrix,S(0) with V_{ua}

c) k-step :calculate the vectors CTopic(k)=[] with S(k) by formula(13) // Update the topic point $CTopic_b^{c^{jh}}$

d) Update S(k),S(k+1) by formula(12) // Update the correlation matrix

e) If $\left\|S(k+1)-S(k)\right\|<\varepsilon$

f) Then stop

g) Else return to step c

h) Return $V_{c^{jh}}$

2.4.2 Multi-granular community Alignment method

For the communities and groups in source social network G and target social network, the topic features are obtained in the same vector space based on the above algorithm, and pair-to-pair alignment is carried out at the same level, as shown in formula (14) : G'

$$Sim\_V_c = \frac{\sum_{b=1}^{z}\left(CTopic_b^{c^{jh}} \times CTopic_b^{c^{jh}}{}'\right)}{\left|\sum_{b=1}^{z} CTopic_b^{c^{jh}}\right| \times \left|\sum_{b=1}^{z} CTopic_b^{c^{jh}}{}'\right|} \qquad (14)$$

Because the actual calculation is the same dimensional matrix, the distance operation with () $V_{c^{jh}}$ $V_{c^{jh}}{}'$ $V_{c^{jh}}$ $V_{c^{jh}}{}'^{T}$ can be performed to obtain a square matrix; the module of the square matrix is the similarity of two communities or two user groups.

In terms of running order, this paper adopts the alignment method of multi-granularity iteration from the upper large community with coarser granularity to the lower level. Figure 3 above shows the community structure, as shown in Figure 4: C^{21} and $C^{21'}$ are the highest level communities and have completed the alignment, then iteration to ^{21}the inside of C $^{21'}$and C, and align their internal sub-communities. In this round, C $^{1212'}$and C are aligned, and the iteration continues to its internal user group. User group C $^{05'05}$is aligned with C and continues iterating into two user groups for user matching.

2.4.3 User alignment

User groups aligned in the source social network G and target social network User members can match users based on user attribute data. For two groups of small users with highly similar themes, the attribute difference is an efficient way to distinguish. G' Similar user names[4] and user profiles (profile, education, unit, etc.)[3]Therefore, based on the document similarity algorithm[19]

$$Sim\_A(u,u') = \min_{P\geq 0} \sum_{i,j=1}^{n} P_{ij} D_{ij} \qquad (15)$$

Where represents the best conversion matrix and represents the word shift distance. P_{ij} D_{ij}

3 Experiment and Analysis

3.1 Data Sets

This paper uses a crawler to obtain the Weibo - Zhihu user data set from the real network for the experiment. In the process of acquisition, users with less than 20 posts per year are filtered in order to avoid the topic being inaccurate due to the rarity of user-generated content. By identifying Weibo accounts or high-impact real-name authentication anchor users in Zhihu accounts, 1468 pairs of positive samples were obtained. In addition, 1468 pairs of non-aligned users were randomly captured in Zhihu and Weibo as negative samples so that the number of positive and negative samples was equal. In the end, a total of 5876 users were obtained, including 61002 user contacts and 138749 user posts. In the experiment, to extract the topic of the user, in order to avoid interference, the "Harbin Institute of Technology stop Word List"[20] assisted filtering.

In order to analyze the performance of the subsequent community detection, this paper also analyzed the external performance indicators of the community by using the open-labeled data set Aminer[13]Cora [21]Cora contains 2708 users, 5278 edges, and seven default categories.

Figure 4: Examples of Multi-grained community alignment

3.2 Evaluation Metrics

3.2.1 Evaluation index of the original and stable topic number of users

The contour coefficient is used to evaluate the clustering effect of user topics, i.e

$$s(i) = \frac{b(i) - a(i)}{\max(a(i), b(i))} \tag{16}$$

Where represents the average distance between $a(i)$ i and other samples of the same cluster and represents the minimum average distance between $b(i)$ i and other clusters.

Topic time jitter is used to evaluate the effect of stable topic selection, as shown in formula (2) above.

3.2.2 Evaluation index of feature points

Feature density[18] is used to evaluate the effect of feature extraction on community themes, i.e

$$\Delta dense = avg(\frac{\left|\{(p,q) \mid p,q \in in_i,(p,q) \in in_i\}\right|}{\left|\{in_i\} + \{out_i\}\right|}) \qquad (17)$$

Where is the sample point, representing the edge collection of feature points within the link cluster and the edge collection of feature points outside the link cluster? p,q in_i out_i

3.2.3 Evaluation index of community detection

normalized mutual information *(NMI)* index and *modularity* index are mainstream indicators[22]

$$NMI = \frac{2 * I(U,V)}{H(U) + H(V)} \qquad (18)$$

U represents real classification, *V* represents algorithmic classification results, *I* represents interactive information, and *H*() represents cross-entropy.

$$modularity = \sum_{i=1}^{n}\left[\frac{L_i}{TL} - \left(\frac{D_i}{TL}\right)^2\right] \qquad (19)$$

L_i Is the total number of edges in community *i*, is the sum of vertex degrees in community i, D_i *TL* is the total number of edges, *NMI* measures the similarity between the algorithm's running result and the preset label. M*odularity* measures the inner tightness of the community.

3.2.4 Alignment Performance Evaluation Indicators

The *Accuracy* index is used to detect the alignment accuracy[24]

$$Accuracy = \frac{1}{NC}\sum_0^j\sum_0^h success\left(c^{jh}, c^{jh}{}'\right) \qquad (20)$$

Where NC represents the total number of communities and user groups, if the alignment of users exceeds 20% of the total number, one is returned. Otherwise, 0 is returned. $c^{jh}, c^{jh}{}'$ $success\left(c^{jh}, c^{jh}{}'\right)$

The user alignment accuracy metric uses Precision, Recall, and F1 values.

3.3 Experimental Results and Analysis

3.3.1 Analysis of user original topic number k and user stable topic number kt

In this paper, the TF-IDF model was used to analyze user interests, and then 50 keywords with the highest frequency and at least ten occurrences were selected after the stop words were removed. K-means clustering was performed on the keyword sequence to generate K original topics. Figure 5 shows the change of the average contour coefficient when the value of k number of topics [2,20] changed. As shown in FIG. 6, both Zhihu and Weibo users reach the local optimal average contour coefficient when k takes a value of 12. In subsequent experiments, users of the two networks are processed together to analyze parameter selection, and the original topic takes a value of 12 to carry out a short-time hot spot filtering discussion.

Figure 5: Effect of k value on contour coefficient

When short-term hot topics are filtered, the statistical change trend of user topics tends to be flat[25]According to formula (2), the jitter degree of 12 original topics of users is calculated and arranged in ascending order. Figure 6 shows the average jitter degree of users' original topics. It can be seen that the average jitter degree of the 9th topic is significantly higher than that of the 8th topic, and it can be concluded that kt=8 can achieve the optimal experimental state and the user topic is the most stable.

Figure 6: average dithering degree of topic

3.3.2 Z-analysis of community theme feature points

The number of feature points z of the theme feature is related to the accuracy of describing the community theme feature. If the z value is too small, the feature is too general and not accurate enough; if it is too large, it is too sparse. In this paper, weighted constant and termination constant $\beta = 2$ $\varepsilon = 10^{-7}$ [26]z is 6, the density of community theme features is 0.28 to achieve local optimization. It means that when the number of community theme feature points reaches 6, the aggregation effect of user theme vectors in the community is the best. Therefore, in the subsequent experiment, the number of feature points z takes the value of 6.

Figure 7: Effect of *z* value on density of characteristic

3.3.3 Performance analysis of topic-based hierarchical community detection

In order to evaluate the performance of the ST-L algorithm, the following community detection models are used in this paper for comparison:

GEMSEC[23]: Community single-layer self-clustering algorithm based on machine learning to regularize users' social attributes;

vGraph[21]: Based on the probability generation model, using user information to predict interaction probability, single-layer community detection;

UICD[8]: Based on user-generated content, obtain user interest through LDA algorithm and use interest for single-layer community detection;

HIOC[12]: Based on corpora fixed label classification, hierarchical community detection is carried out by user interest and label similarity;

ReinCom[13]: learning user characteristics based on deep learning model, optimizing community tree and hierarchical community detection algorithm.

Since Weibo and Zhihu data sets have no preset classification, they cannot obtain the *NMI* index, so the *NMI* index is only discussed in Aminer and Cora data sets. The detailed experimental data are shown in Table 1. It can be found that six algorithms have better effects on Aminer and Cora data sets than Weibo and Zhihu data sets. Because Aminer and Cora are public data sets with preset classification, users have high convergence according to classification. At the same time, for the four data sets, the *NMI* and *modularity* performance of the hierarchical community detection method is significantly superior to that of the single-layer community detection algorithm, indicating that the adoption of hierarchical community detection is effective. Among the hierarchical community detection algorithms, ST-L and ReinCom are superior to the HIOC algorithm, with fixed classification labels and multiple user detection in each data set. Compared with ReinCom in Aminer and cora, ST-L has increased the *NMI* index by 14.8% and 2.6%, while the *modularity* index has increased by 1.8% and 2.1%. In Weibo and Zhihu, the *modularity* index improved slightly. In general, ST-L makes full use of subject features and adopts the non-fixed label layering mode to improve the performance of the data set.

Table 1 *NMI* index and *modularity* index of each algorithm

3.3.4 Community Alignment Performance Analysis

In order to better evaluate the MGA community alignment performance, this paper compares it with the following algorithm on the microbo-Zhihu dataset:

CAlign[24]: Based on user attributes, Dirichlet distributed clustering is used to build communities and cross-network community alignment is performed.

PERFECT[17]: Embed the cross-network user information into the hyperbolic space based on the Poincare sphere model, cluster the users within the hyperbolic distance threshold into communities, and align the communities by the hyperbolic distance of each user.

Detailed experimental data are shown in Table 2. It can be found that MGA improves the accuracy of community alignment by 6.1% and 2.5% compared with CAlign and PERFECT algorithms, which proves that MGA adopts multi-granularity fuzzy topic feature alignment effectively reduces the influence of edge users on community alignment accuracy compared with forced clustering method.

Table 2: Performance of community alignment

3.3.5 User Alignment Performance Analysis

In order to better evaluate the MGA user alignment performance, this paper will use the mainstream user alignment algorithm to compare the micro-blog Zhihu dataset:

BSNA[4]: Based on the BP neural network, the user name is uniformly mapped, the problem is transformed into a vector mapping problem, and the user alignment is performed;

PERFECT[17]: Embed cross-network user information into hyperbolic space based on the Poincare ball model and align users by hyperbolic distance between vectors;

UGCLink[27]: Model user-generated content based on a convolutional neural network and align users by the relationship between content and time;

MEgo2Vec[28]: Self-centered network based on graph neural network mining user attributes and structure for user alignment;

MUIUI[3]: Based on the three characteristics of user attributes, generated content and relationship, a fusion classifier carries out user alignment.

The Precision, Recall and F1 values of each comparison algorithm and MGA algorithm are shown in Table 3. It can be found that the performance of the BSNA algorithm for vector mapping alignment based only on user name is inferior to other algorithms, indicating that the accuracy of the alignment method based on a single feature is low. Similarly, the UGCLink algorithm, based on user-generated content combined with time features, and the PERFECT algorithm, based on user topology embeddings combined with a small number of attribute features, has improved performance, but MEgo2Vec and MUIUI algorithms are significantly superior to the previous two algorithms. Compared with MEgo2Vec and MUIUI, the MGA algorithm proposed in this paper has improved the accuracy rate by 4.5% and 3.4%, the recall rate by 2.7% and 2.1%, and the F1 value by 3.4% and 2.5%. Because MGA makes full use of the topic features generated by user-generated content, the community detection before user alignment reduces the interference of user pairs with similar attribute features to a certain extent. MGA has richer granularity choices than the two multi-granularity alignment algorithms.

Table 3: Performance of user alignment

The experimental results show that the MGA algorithm makes full use of the generated content and fully combines the time characteristics to obtain the user stability theme, performs hierarchical community detection and cross-network multi-granularity alignment, solves the problem of excessive users at the edge of the community, and effectively improves the user alignment accuracy, recall rate and F1 value by combining the user attribute characteristics.

4 Conclusion

Based on user-generated content, this paper makes full use of time characteristics, filters short-time topics, obtains user-stable topics, improves the Louvain algorithm, conducts multi-level community detection for users without preset label classification, and proposes a multi-granularity alignment method for two alignment modes of groups and users. The improved fuzzy topic feature acquisition algorithm is adopted to solve the problem of the reduced accuracy of forced clustering of edge users effectively. In the real data set, it is verified that MGA has good performance in the multi-granularity alignment mode, both community alignment and user alignment. In future work, the edge weights based on stable topics can be integrated into the user structure features and attribute features to make the weights more diverse and further improve the performance of hierarchical community detection. In the multi-granularity alignment method it can also be improved into the alignment method integrating structure, attributes and content, reducing the impact of edge users on community alignment and further improving user alignment accuracy.

5 References

[1] Jiang Y, Jiang J C. Diffusion in social networks: A multiagent perspective[J]. IEEE Transactions on Systems, Man, and Cybernetics: Systems, 2014, 45(2): 198-213.

[2] Shi Wei, Wang Yuanyuan, Li Gang, Zhang Xing. Application Research of Computers,.2023-06-26,1-7.(Shi wei,Wang yuanyuan,Li gang,Zhang xing (K, l) weighted anonymous algorithm for social networks based on KFCMSA, Application Research of Computers, 2023-06-26,1-7) L) Weighted Anonymous Algorithm for Social networks based on KFCMSA, Application Research of Computers, 2023-06-26,1-7)

[3] Zeng W , Tang R , Wang H , et al. User Identification Based on Integrating Multiple User Information across Online Social Networks[J]. Security and Communication Networks, 2021, 202:1-14.

[4] Yuan Z, Yan L, Xiaoyu G, et al. User Naming Conventions Mapping Learning for Social

Network Alignment[C]//2021 13th International Conference on Computer and Automation Engineering (ICCAE). IEEE, 2021: 36-42.

[5] Ren F , Zhang Z , Zhang J , et al. BANANA: when Behavior ANAlysis meets social Network Alignment[C]// Twenty-Ninth International Joint Conference on Artificial Intelligence and Seventeenth Pacific Rim International Conference on Artificial Intelligence IJCAI-PRICAI-20. 2020.

[6] Wang Qian, Shen derong, Feng Shuo, et al. Cross-social Network User Identification based on Full perspective feature combined with crowdsourcing [J]. Journal of Software, 2018, 29(3):811-823.(Wang qian,Shen derong,Feng shuo,et al. Identifying Users Across Social Networks Based on Global View Features with Crowdsourcing[J].Journal of Software,2018,29(3):811−823.)

[7] Wang S , Li X , Ye Y , et al. Anchor Link Prediction across Attributed Networks via Network Embedding[J]. Entropy, 2019, 21(3):254.

[8] Jiang L, Shi L, Liu L, et al. User interest community detection on social media using collaborative filtering[J]. Wireless networks, 2019: 1-7.

[9] Shi L , Yan W , Lu L , et al. Event detection and identification of influential spreaders in social media data streams[J]. Big Data Mining and Analytics, 2018, 1(01):34-46.

[10] Jia, Yan, Zhu, et al. Identifying users across social networks based on dynamic core interests[J]. Neurocomputing, 2016210:107-115.

[11] Chen X , Song X , Cui S , et al. User Identity Linkage across Social Media via Attentive Time-aware User Modeling[J]. IEEE Transactions on Multimedia, 2020, PP(99):1-1.

[12] Zheng J , Wang S , Li D , et al. Personalized recommendation based on hierarchical interest overlapping community[J]. Information Sciences,2019, 479:55-75.

[13] Ding D, Zhang M, Wang H, et al. A Deep Learning Framework for Self-evolving Hierarchical Community Detection[C]//Proceedings of the 30th ACM International Conference on Information & Knowledge Management. 2021: 372-381.

[14] Macqueen J. Some Methods for Classification and Analysis of MultiVariate Observations[C].

Proceedings of Berkeley Symposium on Mathematical Statistics and Probability. 1967:281-297.

[15] Mikolov T , Chen K , Corrado G , et al. Efficient Estimation of Word Representations in Vector Space[J]. Computer Science, 2013

[16] Blondel V D , Guillaume J L , Lambiotte R , et al. Fast unfolding of communities in large networks[J]. Journal of Statistical Mechanics Theory & Experiment, 2008, 2008(10): P10008.

[17] Sun L , Zhang Z , Zhang J , et al. Perfect: A Hyperbolic Embedding for Joint User and Community Alignment[C]// 2020 IEEE International Conference on Data Mining (ICDM).IEEE, 2020,501-510.

[18] Naderipour M, Fazel Zarandi M H, Bastani S. Fuzzy community detection on the basis of similarities in structural/attribute in large-scale social networks[J]. Artificial Intelligence Review, 2022: 1-35.

[19] KUSNER M J, Sun Y , KOLKIN N I, et al. From word embeddings to document distances[C]//International Conference on Machine Learning. JMLR.org, 2015, 37: 957-966.

[20] Guan Qin, Deng Sanhong, Wang Hao. Data Analysis and Knowledge Discovery,2017,1(03):72-80.(Guan qin,Deng sanhong,Wang hao. A Comparative Study on Common Discontinued Words in Chinese Text Clustering[J]. Data Analysis and Knowledge Discovery, 2017,1(03):72-80.) Discontinued words in Chinese text Clustering[J]. Data Analysis and Knowledge Discovery, 2017,1(03):72-80.

[21] Sun F Y , Qu M , Hoffmann J , et al. vGraph: A Generative Model for Joint Community Detection and Node Representation Learning[J]. Advances in Neural Information Processing Systems(S), 2019, 32.

[22] Wang xh, Hao cx , Guan xh . Hierarchical and overlapping social circle identification in ego networks based on link clustering - ScienceDirect[J]. Neurocomputing, 2020, 381:322-335.

[23] Rozemberczki B , Davies R , Sarkar R , et al. GEMSEC: Graph Embedding with Self Clustering[C]// 2019 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM). IEEE, 2019,65-72.

[24] Chen Z, Yu X, Song B, et al. Community-based network alignment for large attributed network[C]//Proceedings of the 2017 ACM on Conference on Information and Knowledge Management. 2017: 587-596.

[25] Ni H Y. Design and implementation of News recommendation system based on user interest drift and semantic features [D]. Beijing University of Posts and Telecommunications,2020.(Ni hanyu. Design and Implementation of a News Recommendation System Based on User Interest Drift and Semantic Features[D]. Beijing University Of Posts andTelecommunications,2020.)

[26] Chan, Keith, C, et al. Fuzzy Clustering in a Complex Network Based on Content Relevance and Link Structures[J]. IEEE Transactions on Fuzzy Systems A Publication of the IEEE Neural Networks Council, 2016, 24(2): 456-470.

[27] Gao H , Wang Y , Shao J , et al. UGCLink: User Identity Linkage by Modeling User Generated Contents with Knowledge Distillation[C]// 2021 IEEE International Conference on Big Data (Big Data). IEEE, 2021,607-613.

[28] Jing Z , Bo C , Wang X , et al. MEgo2Vec: Embedding Matched Ego Networks for User Alignment Across Social Networks[C]// the 27th ACM International Conference. ACM, 2018,327-336.

Figure 1

Flow chart of MGA

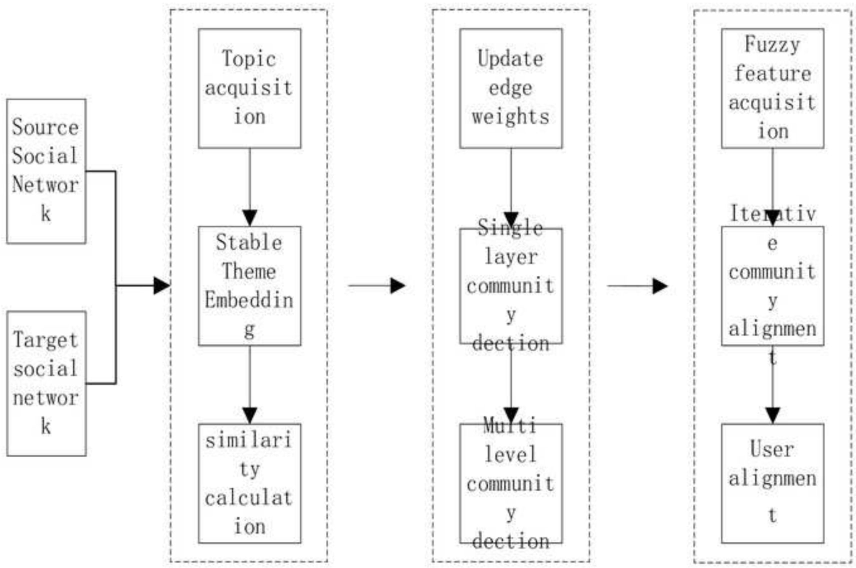

Source Social Networ k
Target social networ k
Topic acquisit ion
Stable Theme Embeddin g
similari ty calculat ion
Update edge weights
Single layer communit y deection
Multi level communit y deection
Fuzzy feature acquisit ion
Iterativ e communit y alignmen t
User alignmen t

Figure 2

Examples of single-layer community detection

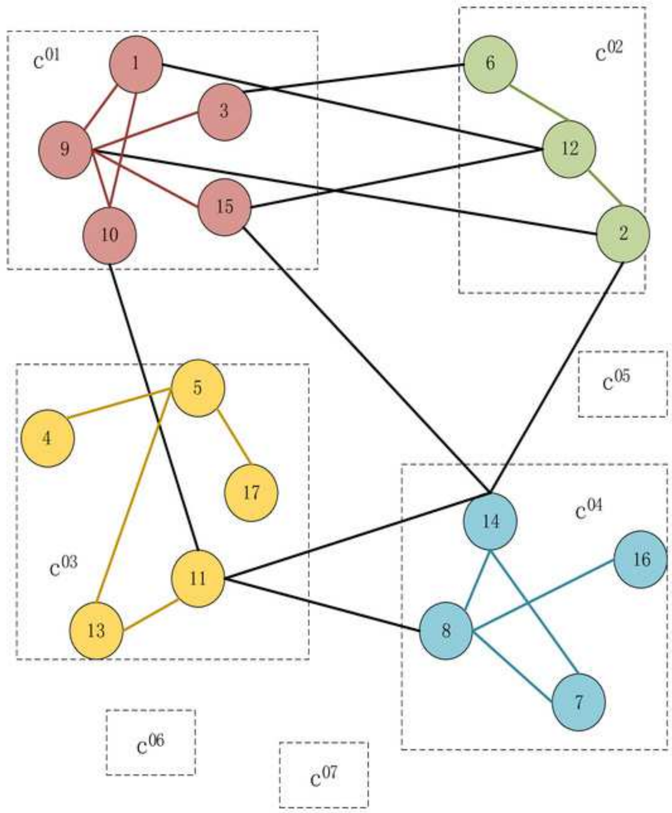

c01
1
3
9
15
10
c02
6
12
2
c05
5
4
17
c03
11
13
14
c04
16
8
7
c06
c07

Figure 3

Examples of multi-layer community detection

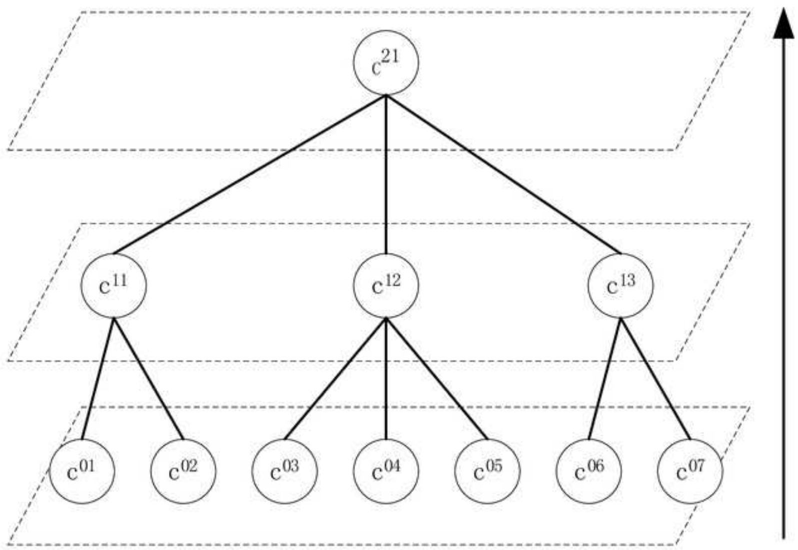
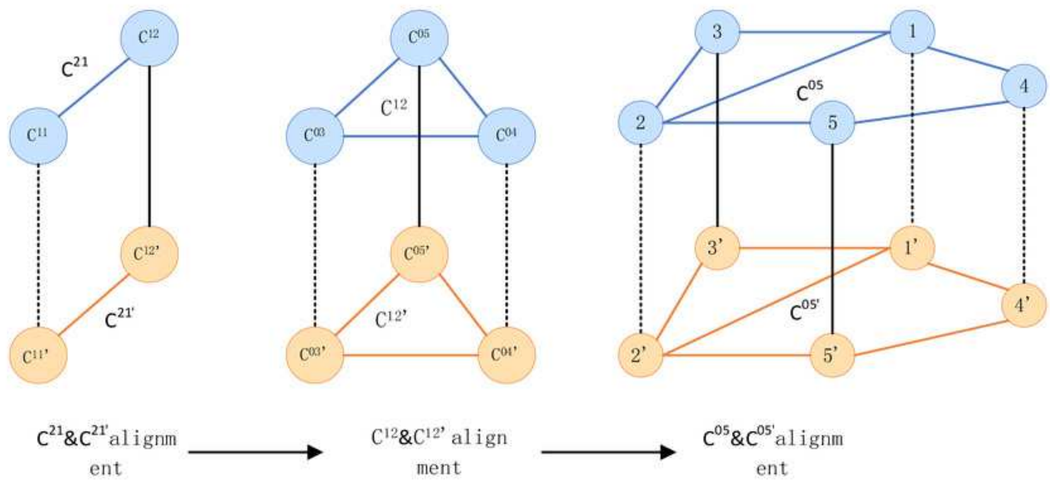

c^21
c^11
c^12
c^13
c^01
c^02
c^03
c^04
c^05
c^06
c^07

Figure 4

Examples of Multi-grained community alignment



C21
C12
C11
C12'
C21'
C11'
C05
C12
C03
C04
C05'
C12'
C03'
C04'
3
1
C05
4
2
5
3'
1'
C05'
4'
2'
5'
C21&C21'alignment
C12&C12' alignment
C05&C05'alignment

Figure 5

Effect of k value on contour coefficient

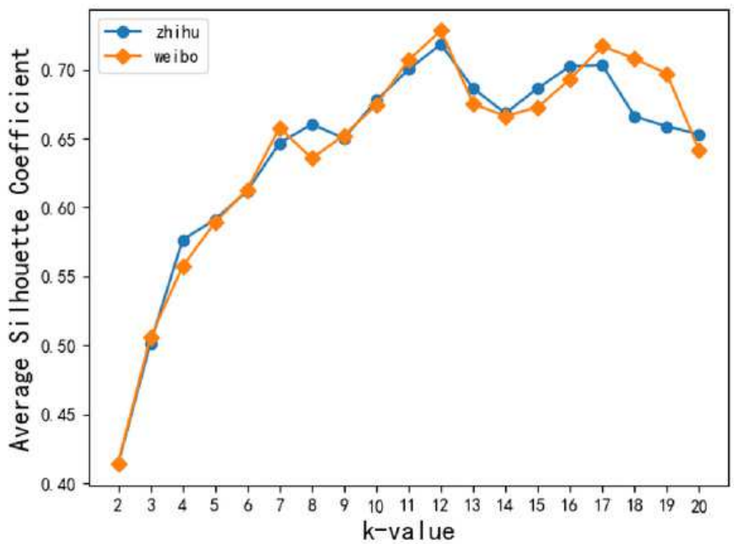

zhihu
weibo
Average Silhouette Coefficient
0.70
0.65
0.60
0.55
0.50
0.45
0.40
2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20
k-value

Figure 6

average dithering degree of topic

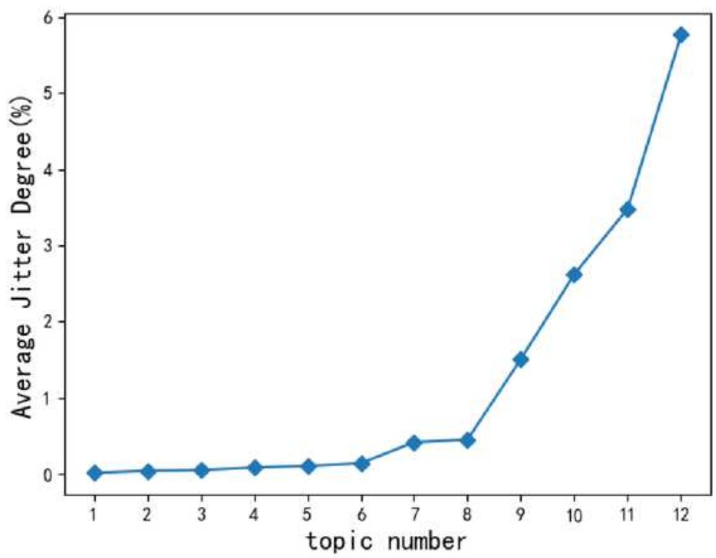

Average Jitter Degree(%)
topic number
0
1
2
3
4
5
6
1
2
3
4
5
6
7
8
9
10
11
12

Figure 7

Effect of *z* value on density of characteristic

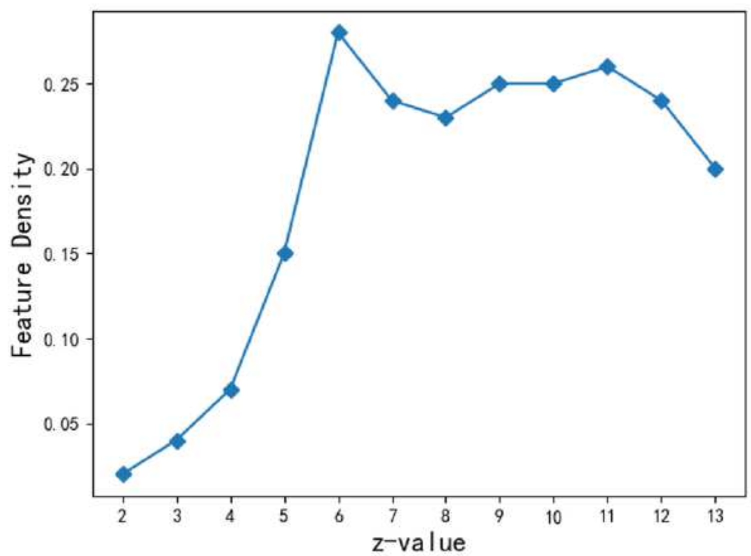

Feature Density
z-value
0.25
0.20
0.15
0.10
0.05
2
3
4
5
6
7
8
9
10
11
12
13

Table 1(on next page)

NMI index and *modularity* index of each algorithm

	Weibo		Zhihu		Aminer		cora	
	NMI	modularity	NMI	modularity	NMI	modularity	NMI	modularity
GEMSEC	/	0.536	/	0.528	0.352	0.653	0.345	0.679
vGraph	/	0.579	/	0.570	0.284	0.710	0.366	0.735
UICD	/	0.498	/	0.503	0.298	0.689	0.315	0.702
HIOC	/	0.617	/	0.611	0.314	0.709	0.442	0.731
ReinCom	/	0.648	/	0.653	0.459	0.742	0.574	0.741
ST-L	/	0.650	/	0.659	0.527	0.756	0.589	0.757

Table 2(on next page)

Performance of community alignment

	Accuracy
CAlign	0.672
PERFECT	0.695
MGA	0.713

Table 3(on next page)

Performance of user alignment

	Weibo - Zhihu		
	Precision	Recall	F1
BSNA	0.572	0.495	0.530
PERFECT	0.646	0.534	0.581
UGCLink	0.652	0.541	0.587
MEgo2Vec	0.667	0.556	0.593
MUIUI	0.674	0.559	0.598
MGA	0.697	0.571	0.613