An Improved Semantic Link Based Cyber Community Discovery Model on Social Network

Weiran Liu, Qiyu Ruan, Liang Zhang, and Wei Ren(✉)

Southwest University, Chongqing 400715, China
adventure@email.swu.edu.cn, oicq@swu.edu.cn

Abstract. Online communities emerge as a major way of delivering and sharing resources. Yet communities in social networks cannot be accurately classified due to the randomness of clustering and the insufficient use of semantics of links. In this paper, a semantic inference based community discovery model is proposed to extract multiple layers of semantics from the topological structure of node relationships and semantic connections between nodes to search and discover communities. The ego-Twitter dataset was used, which contains 81306 nodes (accounts) and 1768149 edges, to test the proposed model. Experiments show that our model is suitable for sparse networks and nodes that contain rich semantics. Especially, in terms of modularity, our model outperforms the Latent Factor Model (LFW) and K-means algorithm. Our model outperforms LFW by achieved faster speed when the scale of online community is expanded to more than 1000, which demonstrates that our model has higher efficiency with network that has abundant semantics.

Keywords: Community discovery · Semantic link · Semantic inference

1 Introduction

The study of community discovery is aim to decompose complex network topology into meaningful node clusters [1–3]. The mainstream community discovery methods now are based on cluster calculations. These algorithms conduct unsupervised learning by observing the attributes of network nodes; However, when these algorithms are applied to node classification of social networks, accurate classifications cannot be achieved [4].The reasons may lie in the different attributes of existing network nodes, various definitions of online community, unsatisfied initializations for cluster algorithms, which will affect the calculation accuracy and the clustering results may ends in randomness. Meanwhile, the existing partition algorithms focus solely on nodes' data such as in and out degree but do not pay much attention to the semantics of individual node and links' attribute information [5, 6]. The current community discovery algorithms are mainly used on undirected graph, whereas community nodes often show its directivity. In order to accurately describe the node relationships, directivity is added to the analysis of community network [7]. We can simulate real-world networks such as social networks

G. Memmi et al. (Eds.): KSEM 2022, LNAI 13370, pp. 329–338, 2022.
https://doi.org/10.1007/978-3-031-10989-8_26

with high accuracy by pay attention to different kinds of links and explain the attributes of them.

The algorithm Satuluri [8] proposed of symmetrizing directed graph in 2011, and the LSW-OCD [9] proposed by Haiyan Zhang in 2015, all transformed the directed graph into undirected graph with directional weight according to the vector of nodes. Yet the node semantics and the relationship between semantics haven't been extracted and been studied. Semantics can not only demonstrate the meaning of objects, but also the relationship between objects. Particularly, in social networks, users' behaviors are closely connected to their own characteristics, hobbies and habits; the application of semantics in online community discovery makes it possible of mining non-data information and to better understand the attributes of users (nodes), so as to achieve a more accurate results of communities' discovery. This paper proposes a semantic link based cyber community discovery model for online community discover in social network considering the great potential of semantics to community discovery.

The rest of the paper is structured as follows. Section 2 prepares readers with the basis of semantic network and semantic search. Section 3 presents the semantic link based online community discovery model. Section 4 presents experimental analysis. Finally, the conclusion is given in Sect. 5.

2 Semantic Network and Semantic Search

Semantic network is the extension of current network so that people and machines can understand each other better [10]. The concept nodes in the semantic network are organized in levels, which can represent the plane relationship between individual nodes and the vertical relationship of different nodes in different levels [11]. Semantic search is the core of semantic network. According to the difference of ontology processing principle, semantic search process can be classified into three types: enhanced semantic search, knowledge-based semantic search and rest [12, 13]. The enhanced semantic search is based on the traditional search engine and adds the ontology library in the traditional database. The ontology library supplements the abstract concept of keywords. Therefore, the semantic search can map the keywords used for input to one or a group of entities or concepts in the semantic network, and use the "point" and "edge" in the semantic network to analyze and reason the graphical expression of entities, concepts, values, attributes and relationships, End users will get rich relevant knowledge from semantic search. However, methods based on keywords cannot support formal query. The search based on semantic annotation is only used as the enhancement of search engine, and the accuracy has not been significantly improved. Therefore, we propose a spatial community discovery model based on semantic relationship. By paying attention to the semantic characteristics between nodes (Equal, Similar, Reference, etc.), we use semantic link for enhanced reasoning, hoping to get community discovery results closer to the real situation.

3 Semantic Link Based Online Community Discovery Model

This paper proposes a spatial model of semantic reasoning for network community discovery and resource tracking. The spatial model divides the complex network topology

into a multi-layer model with semantic relations between nodes. The proposed model can be divided into plan topology and vertical tree topology. The vertical elevation of the model is a forest with tree structure, which reflects the hierarchical relationship of the model and represents all the relationships between the nodes of the *n-th* layer and the semantics (data, resources and services) of each node on the $n + 1$ and *n-th* planes. The semantic relationships of layer n, $n + 1$ are shown in Fig. 1.

(a)The *n-th* layer, *n*+1layers have tree semantic relations (b) *n-th* plane semantic relationship

Fig. 1. The semantic relationship of layers

In order to reduce semantic fuzziness, nodes supporting similar concepts are calculated by similarity and replaced by subclasses or equivalence relations respectively. The semantic query and community discovery can be quickly forwarded to the appropriate semantic level, which can improve the search speed.

3.1 Model Description

Six semantic link types (Equal, Similar, Reference, etc.) are divided into two categories according to their characteristics.

Equality, Reference and Sequence. On the horizontal plane, the relationship between language nodes is represented by a directed graph. In the definition space model, the graph on the *n-th* plane is $G_n = <V_n, Equ_n, Ref_n, Seq_n>$, where V_n is the set of all points in the *n-th* plane, Equ_n represents the set of nodes with "equal" semantic link type in the *n-th* plane, the same goes for Ref_n and Seq_n.

This paper defines similarity as: suppose the semantic similarity of two nodes A and B is α. When α is greater than 0.6, the type of semantic link is classified as equal, if α is less than 0.6, it is considered that the semantics of A and B are irrelevant.

Reference is defined as: *Ref* refers to the node relationship set of all semantic links in the figure with the type of "reference", $Ref_n = <r_{n1}, r_{n2} \ldots r_{nm}>$, and r_{ni} refers to the *i-th* "reference" semantic link in the figure on the *n-th* plane. *Seq* is the set of all nodes whose semantic link type is "sequence", $Seq_n = <s_{n1}, s_{n2} \ldots r_{nm}>$, and s_{ni} is the *i-th* sequence.

Implication and Subclass. The structure on the vertical plane can be expressed as a set of semantic nodes and "subclass" semantic links and "implication" semantic links, $T = (V_n, Sub, Imp)$. The semantic link of "subclass" is expressed as $sub = \cap < v_i, \ldots . v_n >$, where each $<v_i, \ldots . v_n>$ is a "subclass" semantic link from node v_i to v_n. By definition, the "subclass" semantic link has a transitive relationship. "Implication" usually does not simply refer to a simple and clear inclusion relationship.

In order to measure and correct the similarity between node objects, a semantic reasoning table is proposed according to the definition of semantic link relationship in literature [11] and the needs of community discovery, as shown in Table 1.

Table 1. Reasoning rules of semantic link network

No	Rules	Summarization
Rule1	$P_i - equ \rightarrow P_i$	-
Rule2	$P_i - equ \rightarrow P_j$	$P_j - equ \rightarrow P_j$
Rule3	$P_i - equ \rightarrow P_j, P_j - equ \rightarrow P_k$	$P_i - equ \rightarrow P_k$
Rule4	$P_i - imp \rightarrow P_j, P_j - imp \rightarrow P_k$	$P_j - imp \rightarrow P_k$
Rule5	$P_i - st \rightarrow P_j, P_j - st \rightarrow P_k$	$P_j - st \rightarrow P_k$
Rule6	$P_i - imp \rightarrow P_j, P_j - st \rightarrow P_k$	$P_j - imp \rightarrow P_k$
Rule7	$P_i - imp \rightarrow P_j, P_j - ref \rightarrow P_k$	$P_j - ref \rightarrow P_k$
Rule8	$P_i - st \rightarrow P_j, P_j - ref \rightarrow P_k$	$P_j - ref \rightarrow P_k$
Rule9	$P_i - N \rightarrow P_j, P_j - sim \rightarrow P_k$	$P_j - N \rightarrow P_k$
Rule10	$P_i - \varnothing \rightarrow P_j, P_j - sim \rightarrow P_k$	$P_j - N \rightarrow P_k$

Among them, $imp = \cap < v_i, \ldots . v_n >$ represents the set of semantic nodes that may generate reasoning. *equ* stands for equality relation, *sim* stands for similarity relation, *ref* stands for reference relation and *st* stands for subclass relation.

Similarity Calculation. In this paper, each entity e defined is represented as a *vector(e)* of a word space, and each dimension corresponds to a word. The value of the dimension indicates the relative importance or representativeness of the word in describing *e*. A keyword query *q* is also expressed as a *vector(q)* in a word space. Finally, the correlation between *e* and *q* can be expressed as the cosine of the angle between *vector(e)* and *vector(q)*.

The similarity of semantic structure depends on two basic elements: 1. Semantic nodes constitute the leaf nodes of the community, and the query is also composed of leaf nodes, so it is related to the semantic structure of the community. 2. The similarity degree of the ancestor node of the node.

Table 2 defines the functions called to implement the similarity algorithm and their explanations.

The similarity of N_i and N_j semantic structures of different nodes in the semantic network is calculated as follows:

Table 2. Notations and explanations

Notation	Explanation
Peer (N_i)	Semantic mapping node of semantic node N_i
Length (N_i, N_j)	Number of nodes on the path from semantic node N_i to N_j
Max-Semantic-Clique (N_i)	Maximum semantic group including semantic node N_i
Min-Common-Sub-Tree (N_i)	Minimum common subtree including semantic node N_i
Semantic-Node-Similarity-Degrees (N_i, N_j)	Similarity between semantic nodes N_i and N_j

Algorithm 1.Semantic-Structure-Similarity-Degrees (N_i, N_j)

1: IF $N_i \in$ One of the largest semantic groups in the community
THEN T = Max-Semantic-Clique (N_i)
ELSE T = Min-Common-Sub-Tree (N_i)
END IF

2 : Root (N_i) = T
IF Length(N_i, N_j) = 1
THEN Semantic-Structure-Similarity-Degrees(N_i, N_j)=Semantic-Node-Similarity-Degrees (N_i, N_j)
ELSE NodeSet = {N_i, ..., Root (N_i)} //All nodes from node N_i to the root node of N_i

$$\overrightarrow{FV} = (fv_{N_i}, \ldots, fv_{Root(N_i)}) \text{ //Contains the similarity vector from root node to node } N_i$$

$$fv_{N_k} = \begin{cases} 0 \text{//The mapping node corresponding to } N_k \text{ does not belong to the mapping path of } N_k \\ Semantic - Node - Similarity - Degrees\left(N_k, Peer(N_k)\right) \text{ //other} \end{cases} \quad (1)$$

// $\vec{w}$ is the weight vector, which indicates the importance of nodes in the NodeSet, and W_{N_k} indicates the importance between nodes.

$$W_{N_k} = \begin{cases} \frac{1}{2}, N_k = N_i \\ \frac{1}{2}^k, k = length\,(N_i, N_k), \text{and } N_k \neq N_i \\ 1 - \sum_{l=1}^{n-1} w_{N_L} = \frac{1}{2}^{n-1}, n = length\left(N_i, Root(N_i)\right), N_k = Root\,(N_i) \end{cases} \quad (2)$$

$$Semantic - Structure - Similarity - Degrees\left(N_i, N_j\right) = \frac{\overrightarrow{W} \cdot \overrightarrow{FV}}{|\overrightarrow{W}||\overrightarrow{FV}|} \quad (3)$$

Among that

$$\overrightarrow{W} \cdot \overrightarrow{FV} = W_{N_i} fv_{N_i} + \cdots + W_{Root_{(i)}} fv_{Root_{(i)}} \quad (4)$$

$$\left|\vec{X}\right| = \left\|\vec{X}\right\|_2 = \sqrt{x_1^2 + \cdots + x_k^2} \quad (5)$$

END IF

Algorithm 1 describes the similarity of semantic structure between different nodes.

Reconstructing the Semantic Link Network. Algorithm 2 describes how to build a spatial network model by constructing the community spatial structure in the community space, and then divide the community structure by calculating and modifying the semantic similarity.

Algorithm 2. Constructing the Spatial model of semantic link network

1 : The data captured by the web page is filtered through the ontology set, and the output set *S* is initialized.

2 : Traverse the whole community set and determine *Vn*, *Eqa*, *Ref*, *Seq*, *Sub*, *Imp* of some nodes by using the relationship between ontologies,the reasoning table is used to traverse the constructed map to accelerate the convergence of the relationship between nodes.

3 : Take the id of the user's speech as the subject primary key, and construct the spatial model $M = (G, T)$ through semantic group mapping, semantic node mapping and semantic path mapping, where $G = <Vn, Eqa, Ref, Seq>$, $T = (Vn, Sub, Imp)$;

4 : Calculate the word *vector(V)* entered by the user and the semantic structure similarity (*Sim*) between each node.

5 : Normalize the semantic structure similarity of each node (N)$N_{Sim_i} = \frac{Sim_i - Sim_{min}}{Sim_{max} - Sim_{min}}$. (7)

6 : Sort them.Sort(NodeSet$\{N_{Sim_1}, \ldots, N_{Sim_n}\}$). (8)

7 : Let the matrix Mark = 0 //Mark nodes that have been calculated

```
IF N_Sim_i == 1
   THEN FOR Ni in G.Eqa set or T.Sub set ∈ N_Sim_i
      Ni⊂S
      Mark[i] = 1
IF N_Sim_i = <0.2
   THEN FOR Ni in T.Sub set or G.Eqa set ∈ N_sss
      Ni ⊄ S
      Mark[i] = 1
```

8 : FOR Mark[i] == 0 and SimNi > 0.6 //Correction of similarity by correlation degree between nodes

```
ImportantNi = 1
ImportantNj = 0.7
FOR Nj in Imp set
   ImportantNj = ImportantNj + SimNi *(1-α) ImportantNi
FOR Nj in Seq set
   ImportantNj = ImportantNj + SimNi * (1-β) ImportantNi
FOR Nj in Ref set
   ImportantNj = ImportantNj + SimNi * (1-γ) ImportantNi
FOR Mark[j] == 0
   SimNj = Important * SimNj
   IF SimNj > 0.6
   Nj ⊂ S
   Mark[j] = 1
Return S
```

The algorithm is divided into three parts:

First the algorithm adopts the idea of P2P to construct mutual mapping of community space nodes. Then the algorithm normalizes the semantic structure similarity in order to preprocess the similarity correction, and also to obtain standardized output and avoid the influence of extreme outliers. The algorithm quickly forwards to the Equal set and Null/Unknown set of the node. Finally the importance of vertices can be propagated to adjacent vertices along the associated edge. During initialization, the importance of entities with high query relevance is set to 1 and the importance of other entities is set to 0.7. Continuously select the most important entity from the unprocessed entities (Mark == 0) for processing, and add its importance to the importance of implication, sequence and reference entities in the graph.

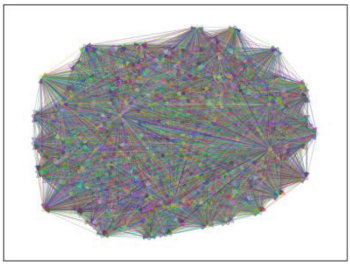
4 Experiment and Analysis

In experiment, we use the ego-Twitter data set provided by Stanford University. The density of its social network graph is 0.00053494, which can well represent the friend relationship in the real world. The experimental environment of this paper is running on a computer with AMD Ryzen 7 4800HS CPU, 16 GB memory and Win10 system.

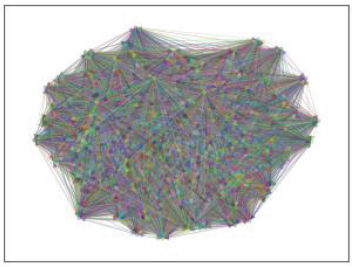
4.1 Experimental Results and Analysis of Community Discovery Model

Effectiveness of Community Discovery Model. First we randomly selects a user in ego-Twitter, then established a two-tier social network relationship diagram, which contains 274 nodes and a total of 5183 sides. We constructs an original directed graph without weight, and the dense matrix is obtained by reasoning. The number of edges is 12476, which contains all semantic relations. The dense matrix is used as the input of similarity calculation to obtain the undirected graph. We divides the community nodes in the undirected graph and obtains the community discovery results. The results of friend directed graph (original), friend directed graph (transformation), friend undirected graph, and friend community discovery are shown in Fig. 2.

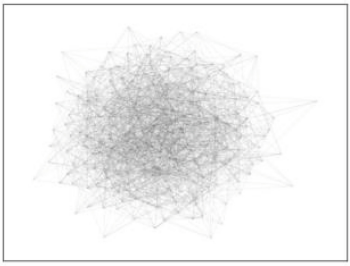
In Fig. 2, (a) Friend directed graph (original data) is inferred based on semantic relationship, and (b) Friend directed graph (transformation) is obtained through six semantic link types (Equal, Similar, Reference, etc.). Then, the directed graph is transformed into undirected graph (c) by calculating node semantic similarity. Finally, reasoning based on semantics to obtain the final discovery result.



(a) Friend directed graph (original data)



(b) Friend directed graph (transformed)



(c) Friend undirected graph

Fig. 2. Results of community discovery

Algorithm Performance Comparison. In order to compare the relationship between community size and modularity faced by different algorithms, we choose *K*-means and Late Factor Model (LFM), which are the mainstream algorithms of community discovery. In the experiment, network starts from 0 nodes, 40 new Twitter user nodes were randomly selected from the database each time and added to the original network. Through simulation, it is found that all algorithms grow linearly under fitting. With the increase of network nodes, the result of K-means algorithm is unstable and fluctuates greatly, so it is not suitable for community discovery of directed graph. The comparison results of the three algorithms are shown in Fig. 3.

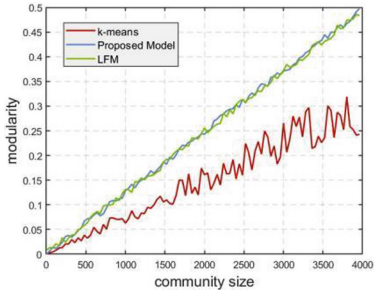


Fig. 3. Results of algorithm comparisons

The results of LFM and the proposed model are similar in terms of community division. Yet, when the community size increases to 1000, the modularity of our pro-posed model shows its advantages.

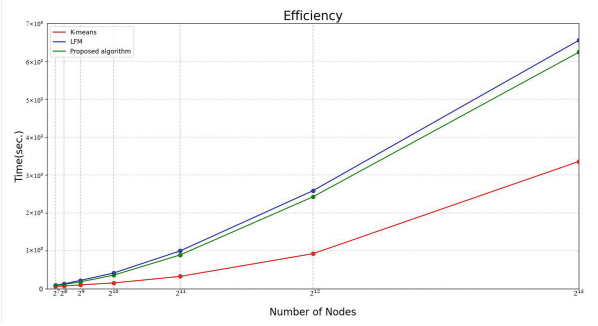


Fig. 4. The proposed model compares with K-means and LFM in efficiency

Figure 4 shows that as the number of nodes increases, our proposed algorithm takes less time than LFM and more time than K-means. The efficiency of our algorithm is better than that of LFM. The efficiency of K-means algorithm is the best. The reason may lies in that K-means uses simple clustering of nodes attributes while the other two algorithms applied reasoning rules in the process.

5 Conclusion

This paper proposed a network community discovery model based on semantic reasoning, and creatively discusses the composition and segmentation of network community from the perspective of semantic network. Based on the horizontal and vertical characteristics of nodes' semantics, the model uses both the tree topology and the spatial network structure to classify nodes into groups based on semantic reasoning. The experiment results show that the proposed model outperforms LFW and k-means algorithms in terms of modularity. Our model per-form even better when the node scale of online

community expands to more than 1000. And our proposed model has higher efficiency in grouping nodes. On the other hand, due to the semantic segmentation and the complexity of reasoning, the speed of this model is not high. It is expected that we improve the speed of community discovery in further research.

Acknowledgement. This paper is funded in part by the Capacity Development Grant of Southwest University (SWU116007).

References

1. Rossetti, G.: ANGEL: efficient, and effective, node-centric community discovery in static and dynamic networks. Appl. Netw. Sci. **5**(1), 1–23 (2020). https://doi.org/10.1007/s41109-020-00270-6
2. Qiu, H., Zheng, Q., Msahli, M., Memmi, G., Qiu, M., Lu, J.: Topological graph convolutional network-based urban traffic flow and density prediction. IEEE Trans. Intell. Transp. Syst. **22**(7), 4560–4569 (2020)
3. Li, Y., Song, Y., Jia, L., Gao, S., Li, Q., Qiu, M.: Intelligent fault diagnosis by fusing domain adversarial training and maximum mean discrepancy via ensemble learning. IEEE Trans. Industr. Inf. **17**(4), 2833–2841 (2020)
4. Coscia, M., Giannotti, F., Pedreschi, D.: A classification for community discovery methods in complex networks. Stat. Anal. Data Mining ASA Data Sci. J. **4**(5), 512–546 (2011)
5. Yang, Z.L., Zhang, W.J., Yuan, F., et al.: Measuring topic network centrality for identifying technology and technological development in online communities. Technol. Forecast. Soc. Chang. **167**, 120673 (2021)
6. Ransa, C.: Research on network sampling and statistical inference method for social network. National University of Defense Technology (2018)
7. Satuluri, V., Parthasarathy, S.: Symmetrizations for clustering directed graphs. In: Proceedings of the 14th International Conference on Extending Database Technology, pp. 343–354. ACM (2011)
8. Zhang, H., Liang, X., Zhou, X.: Overlapping community discovery algorithm for local extension of directed graph. Data Acquisition Process. (003), 683–693 (2015)
9. Berners-Lee, T., Hendler, J., Lassila, O.: The semantic web. Sci. Am. **284**(5), 34–43 (2001)
10. Tan, Y., Zhang, J., Xia, X.: Research on the development process and current situation of semantic network. Libr. Inf. Knowl. (06), 102–110 (2019)
11. Principles of semantic networks: Explorations in the representation of knowledge. Morgan Kaufmann, San Francisco (2014)
12. Weikum, G., Dong, X.L., Razniewski, S., et al.: Machine knowledge: creation and curation of comprehensive knowledge bases. Found. Trends® Databases **10**(2–4), 108–490 (2021)
13. Hu, F., Lakdawala, S., Hao, Q., Qiu, M.: Low-power, intelligent sensor hardware interface for medical data preprocessing. IEEE Trans. Inf. Technol. Biomed. **13**(4), 656–663 (2009)