Research on Topic Recognition Based on Multilayer Relation Fusion

Haiyun Xu[1,2] Rui Luo[1,3] Chunjiang Liu[1,3] Kun Dong[1] Yan Qi[4]

1.Chengdu Library and Information Center, Chinese Academy of Sciences, Chengdu, 610041 (China);
2.Institute of Scientific and Technical Information of China (ISTIC), Beijing 100038 (China);
3. University of Chinese Academy of Sciences, Beijing 100190 (China);
4. Institute Of Medical Information/Medical Library, CAMS & PUMC, Beijing 100020, (China).

Introduction

Background

- A single relationship usually provides the researchers with partial and unbalanced characteristics of one research field.
- Therefore, it is necessary and helpful for researchers to fully understand a research field from different perspectives from the enormous amount of scientific literature.

Purpose

- Methods to enhance the data relationship strength by acquiring complementary information.

Table 1 The Advantages Of Multivlayer Relationship

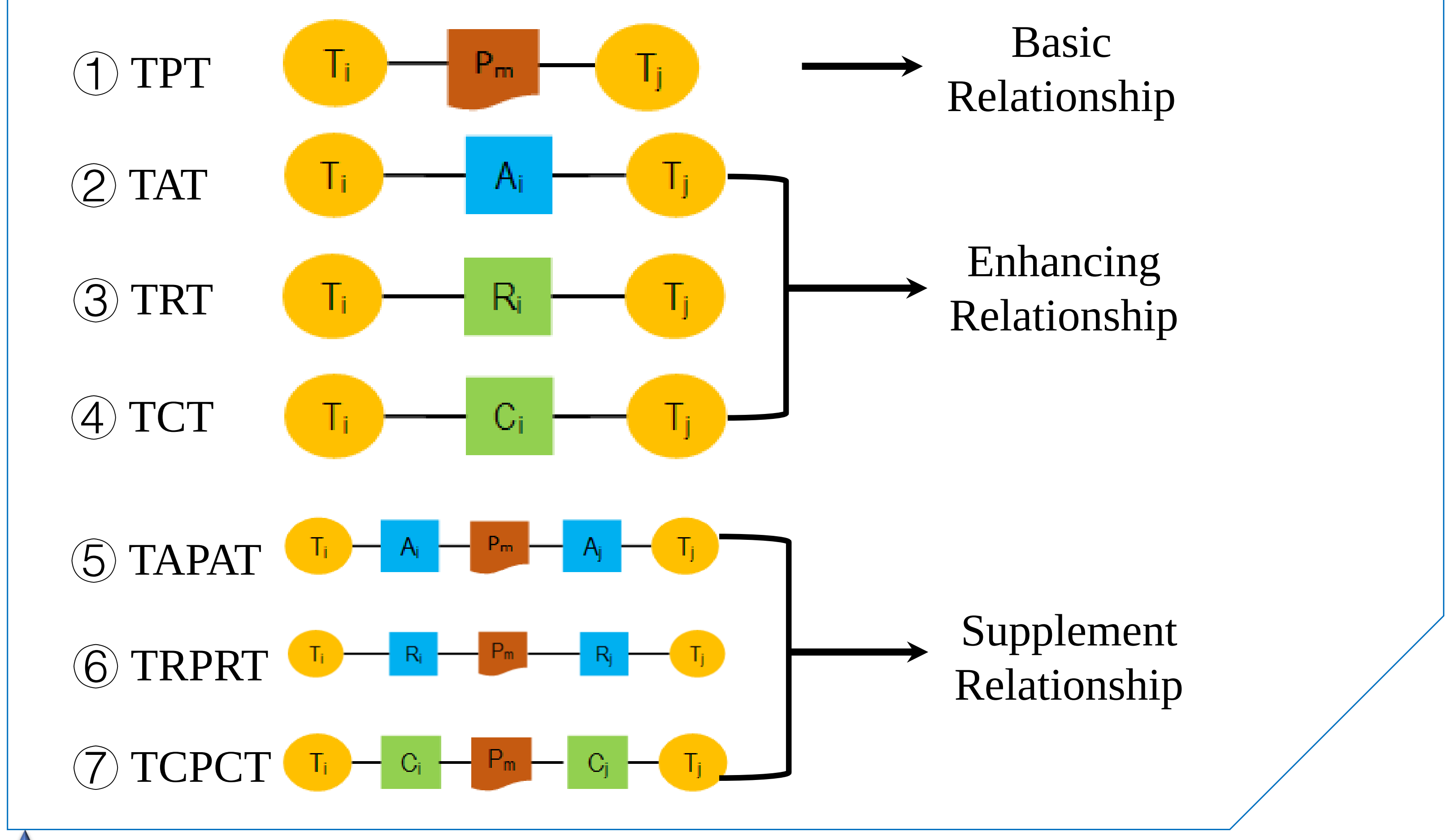
✗	Univariate relationship	→	Incomprehensive
☺	Multilayer relationship	→	More information relationships

1. What are the types of multi-relationships ?

According to the semantic distance between subject terms, this research divided the relationships for topic recognition into three types. The specific meanings are as follows:

Types of text entities

- T: Subject term
- R: Reference
- A: Author
- C: Citation literature



2. How to acquire and measure these relationships?

What is meta-path?

The meta-path P of A is defined as: $A_1 \overset{R_1}{\rightarrow} A_2 \overset{R_2}{\rightarrow} \ldots \overset{R_L}{\rightarrow} A_{L+1}$. It represents a combination of relationships between different node A1 and AL+1, as is shown in the figure above.

How is meta-path useful?

The meta-path contains rich semantic information, and different objects can represent different relationships through different link paths.

How can we use meta-path?

Constructing meta-paths of different lengths and calculating the similarity of keywords based on multiple meta-paths are the focus of this study.

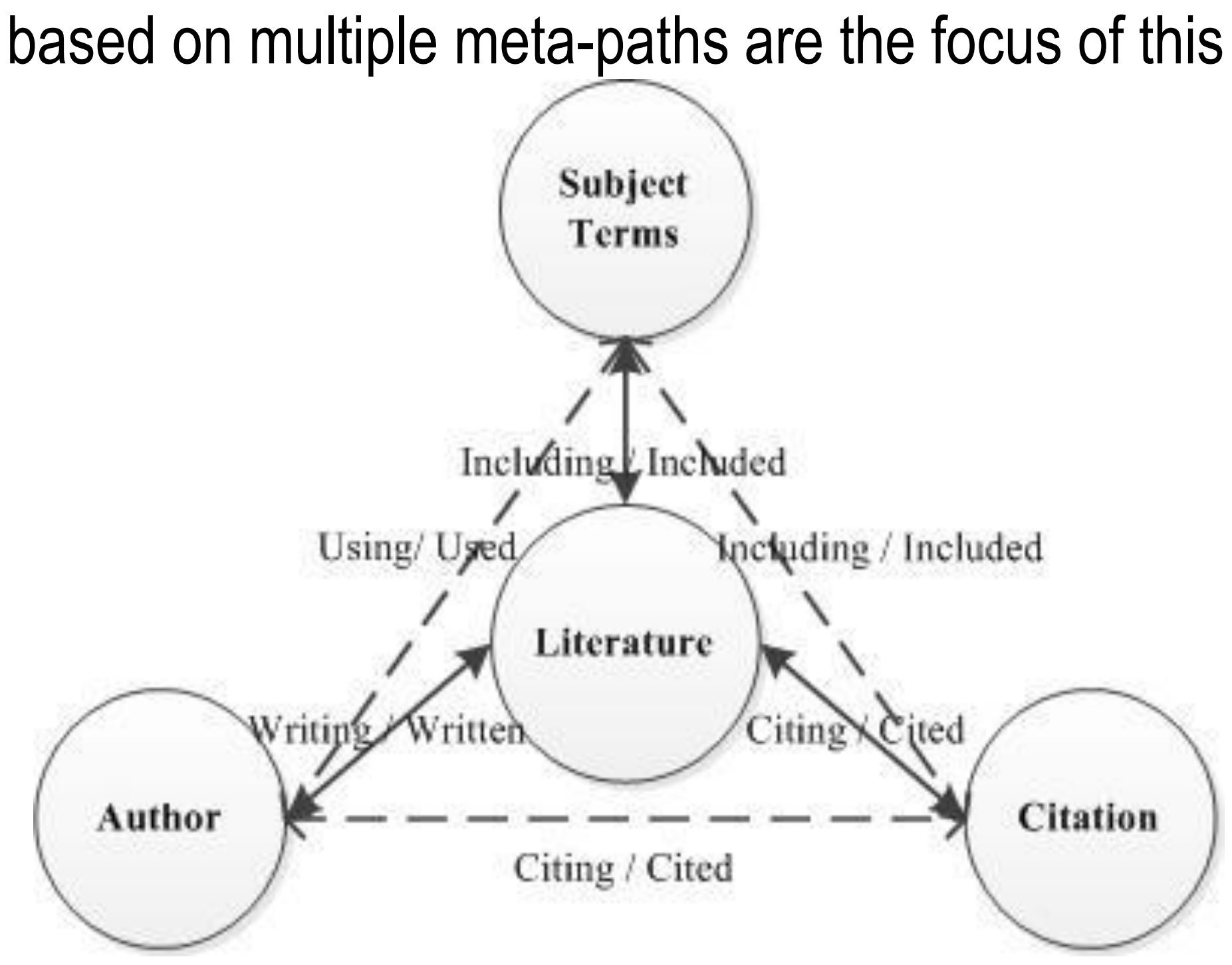


Figure 1 Meta path schematic

3. How to fuse multilayer relation?

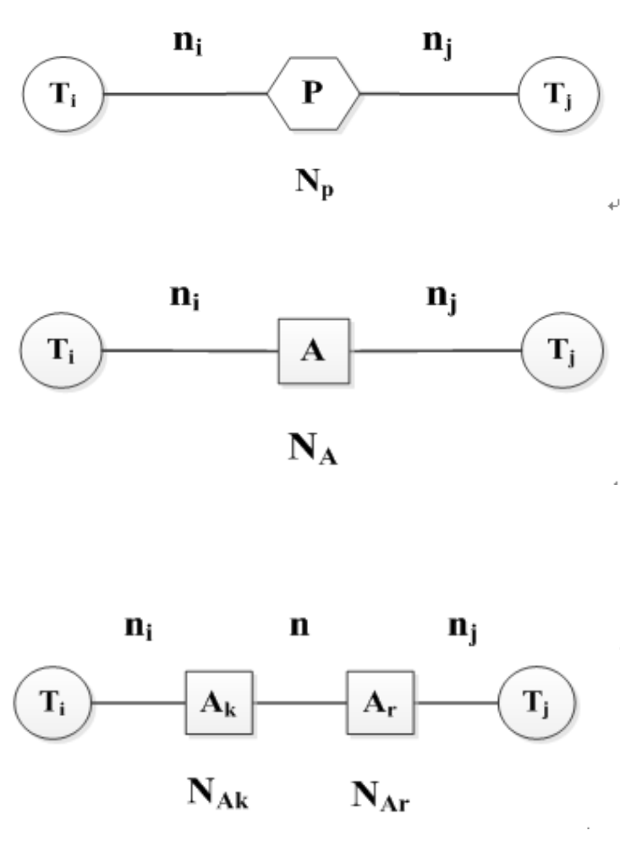
It is based on Morris's definition on the weight of links of multiple relationships of measurement entities.

Step 1

① TPT → Weight of links of basic relationship

② TAT
③ TRT
④ TCT → Weight of links of enhancing relationship

⑤ TAPAT
⑥ TRPRT
⑦ TCPCT → Weight of links of supplement relationship



This study uses the PathSelClus algorithm to fuse 7 types of relation matrices and calculate the comprehensive similarity of the comprehensive subject terms.

Step 2

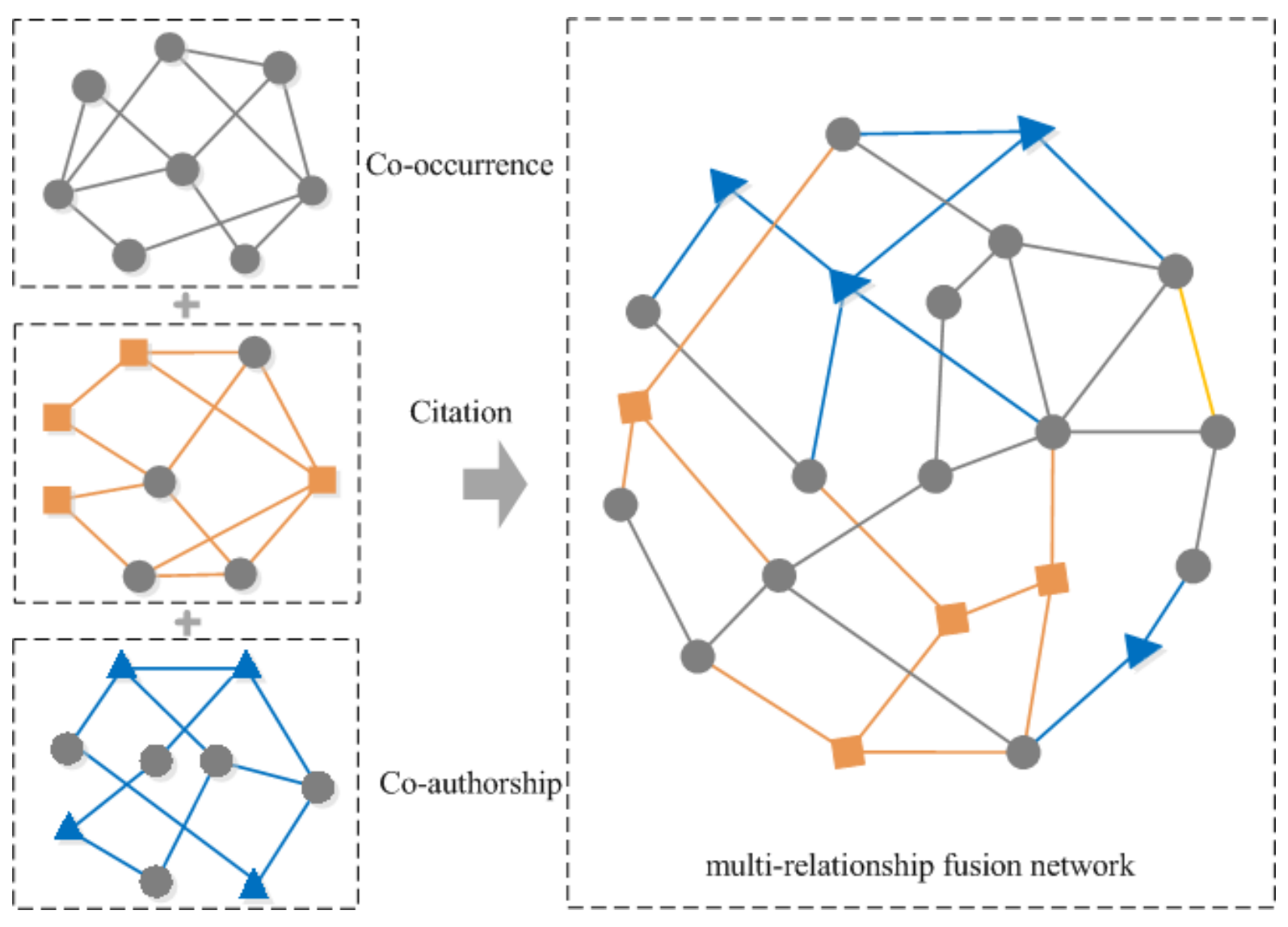


Figure 2 Relationship Fusion

4. Empirical Analysis

Field : Gene-engineered vaccine)

Data Source: China Knowledge Resource Integrated Database

Data Processing: The use of PathSelClus algorithm and nonnegative matrix factorization(NMF) based on multivariate relational clustering of thematic analysis can be used to enhance the topic identification.

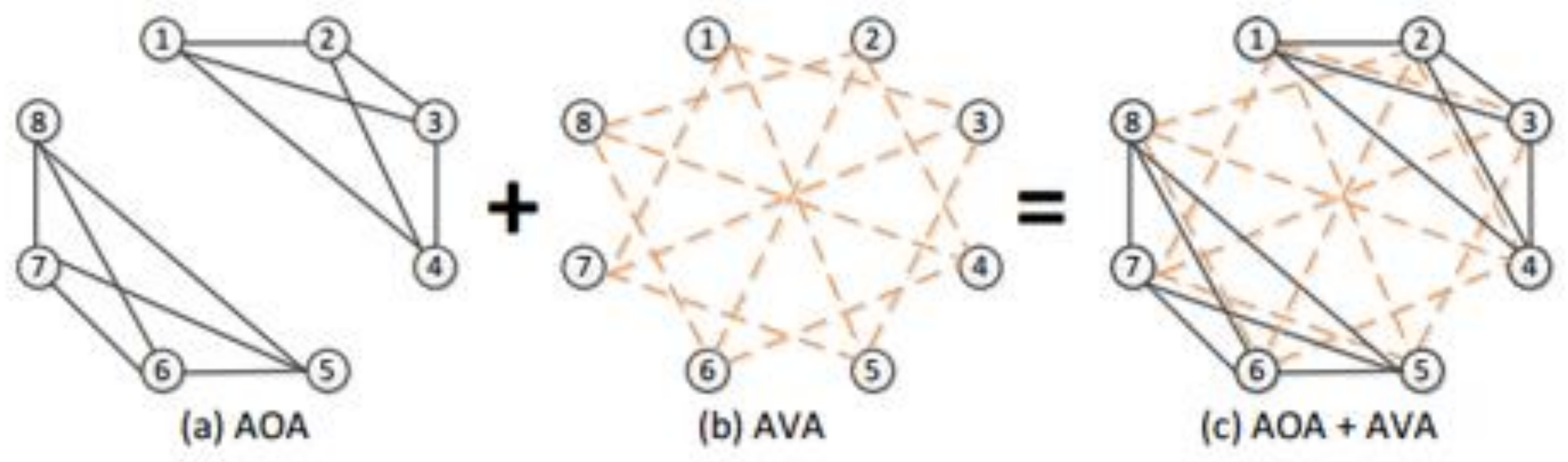


Figure 3 The PathSelClus algorithm algorithm schematic diagram

Partial Clustering Topics

- Construction and expression of antineoplastic nucleic acid vaccine;
- suicidal DNA vaccine and immune response;
- Construction of a dual-promoter DNA vaccine vector and immune response;
- Nucleic acid vaccine associated with avian infections;
- Manufacturing of anti-idiotype monoclonal antibody vaccine.

5. Comparative analysis and Conclusion

Comparative analysis

- This results in a reduction of difference in the interpretation of multiple topics, thus leading to a generalization of a topic's meaning and an increased difficulty in the topic's naming.
- An excessive number of clustering results are formed in each time window, resulting in insufficient effective clustering.

Conclusion

- The relation fusion method is better than single co-word clustering;
- The degree of difference between the topics is evident;
- In different time windows, topic clustering has the following specific differences.

6. Acknowledgement

- National Natural Science Foundation of China (71704170)
- China Postdoctoral Science Foundation Funded Project (2016M590124)
- Youth Innovation Promotion Association, CAS (2016159)