STI 2018 Conference Proceedings

Proceedings of the 23rd International Conference on Science and Technology Indicators

All papers published in this conference proceedings have been peer reviewed through a peer review process administered by the proceedings Editors. Reviews were conducted by expert referees to the professional and scientific standards expected of a conference proceedings.

Chair of the Conference

Paul Wouters

Scientific Editors

Rodrigo Costas
Thomas Franssen
Alfredo Yegros-Yegros

Layout

Andrea Reyes Elizondo
Suze van der Luijt-Jansen

The articles of this collection can be accessed at https://hdl.handle.net/1887/64521

ISBN: 978-90-9031204-0

23rd International Conference on Science and Technology Indicators (STI 2018)
"Science, Technology and Innovation indicators in transition"
12 - 14 September 2018 | Leiden, The Netherlands
#STI18LDN

Research on Topic Recognition Based on Multilayer Relation Fusion[1]

Haiyun Xu[*], Kun Dong,[**] Rui Luo,[**] and Chao Wang[**]

[*]*xuhy@clas.ac.cn*
Chengdu Land Information Center, Chinese Academy of Sciences, Chengdu, 610041 (China)
Institute of Scientific and Technical Information of China (ISTIC), Beijing 100038
[**]*dongkun@mail.las.ac.cn; luorui@mail.las.ac.cn; wangchao2015@mail.las.ac.cn*
Chengdu Documentation and Information Center, Chinese Academy of Sciences, Chengdu, 610041 (China);
University of Chinese Academy of Sciences, Beijing 100190

Introduction

Text-based automatic topic recognition forms the basis of numerous types of intelligence analysis, including literature classification, literature retrieval, research frontiers, and research focus identification. With the amount of scientific literature increasing exponentially, the different types of literature are becoming abundant; therefore, developing methods that apply big data to scientific text is important in text-based topic recognition. One of the typical characteristics of intelligence analysis based on big data is analyzing the relations among multilayer data. A research literature has multiple measurement entities, including both multiple direct and indirect relationships among different entities and knowledge units. However, current methods of text-based topic identification rely mainly on univariate relation analysis. These methods cannot comprehensively analyze the available information, making it difficult to accurately recognize the real innovative scientific topics (Xu, Dong, Liu, Wang, & Wang, 2017). Therefore, analyzing the relationships among entity topics and further fusing multilayer relationships is the key for topic recognition based on the enormous amount of scientific literature.

In this paper, we first review the research status of multilayer relation fusion and then summarize the existing methods of extraction and fusion of multiple relationships in topic recognition. Then, considering gene-engineered vaccine (GEV) as an empirical analysis, we use the PathSelClus algorithm to cluster the topics based on multilayer relation fusion. Finally, we processe a comparative analysis and give the research conclution.

Literature Review

Concept of multilayer relation and its application in topic recognition

Multilayer relation fusion refers to comprehensively analyzing different types of relational data through specific methods and using all the information to reveal the characteristics of the research objects. Thus, it remedies the deficiencies of the single-type relation in revealing the inter-entity relations to obtain a more comprehensive and objective measurement result (H.-Y. Xu et al., 2017). The three main types of entity relationships in textual topic recognition are

[1] This work was supported by National Natural Science Foundation of China (Grant No. 71704170), the China Postdoctoral Science Foundation funded project (Grant No. 2016M590124), and the Youth Innovation Promotion Association, CAS (Grant No. 2016159).

citation-based analysis, text-based analysis, and a combination of these two types (Janssens, Zhang, De Moor, & Glänzel, 2009). At present, the majority of research focuses on the use of the first two methods, which are primarily based on a single relationship analysis.

Besselaar et al. (Van Den Besselaar & Heimeriks, 2006) proposed using word-reference co-occurrences to represent research topics and defined research fronts as the literature sets derived from the clusters of co-occurrences maps. Wen et al. (Wen, Horlings, Marie lle, & Peter, 2016) studied the cognitive structure of scientific fields from three types of mapping approaches—journal-journal citation relations, shared author keywords, and title word-cited reference co-occurrence. Then, they used cross-tabular analysis to perform cross-validation. The empirical analysis proved that the three methods can be combined to recognize the cognitive structure and the hybrid method can identify the interdisciplinary fields more precisely.

According to whether the fusion relationship is filterable or compatible, multilayer relation fusion can be divided into serial and parallel fusion. Serial fusion refers to considering other relationships under the constraint of one type of specific relationship. In essence, it is the sequential superposition of restricted relationships. Parallel fusion implies considering multiple relationships simultaneously (Xu, Dong, Wang, Wei, & Yue, 2018). Serial fusion has a more important role in the cross-correction of information, whereas parallel fusion has an influence on information enhancement.

Multilayer relationship fusion enhances accuracy of text recognition
There are two preferred combination methods: one is to use references as the restriction of the relationship between words, the other is to construct the relationship between references and words through citation.

Calero-Medina et al. (Calero-Medina & Noyons, 2008) combined word co-occurrence and citation network analysis to investigate the process of knowledge creation and transfer through scientific publications. He et al. (He, Zha, Ding, & Simon, 2002) proposed a method of web text clustering that combined textual hyperlink structure, co-citation relations, and textual content analysis. They used a textual hyperlink structure as the dominant method to measure the textual similarity and textual similarity to modulate the strength of hyperlink. Then, they combined the co-citation method linearly to integrate a weighted adjacent matrix. Janssens (Janssens, 2007; Janssens, Glänzel, & De Moor, 2007; Janssens et al., 2009) used Fisher's inverse chi-square method to integrate textual methods and bibliometrics methods, which effectively improved the accuracy of information fusion.

Extraction and Fusion of Multilayer Relation in Topic Recognition

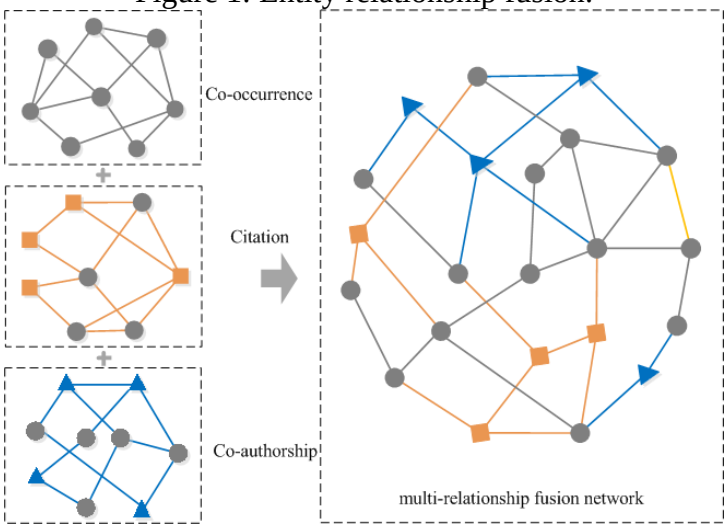
Topic-related relationship types in scientific literature

The multi-relationship of topic recognition refers to the different relationships established by the subject terms based on other measurement entities. The text is composed of subject terms with semantic information according to a certain logical structure. In addition to the physical positional relationship between the subject words, there are implicit semantic associations. Therefore, the semantic representation rules in the text can be determined, and then the text's topic can be identified through mining the intrinsic associations between the terms.

Semantic relations, which are linked by research topic between the subject terms, authors, and citations in the scientific literature (H.-Y. Xu et al., 2017), are present. Relationships, subject

term co-occurrence, citation, and co-authorship all reveal thematic relationships from different perspectives. This study applies the parallel fusion method to the relationships between subject terms, authors, and citations to accommodate for the lack of information in a unitary relationship and obtain a more accurate thematic relationship. As displayed in Figure 1 (Xu et al., 2018), in this study a multi-relationship fusion network can be formed by integrating co-occurrence, citation, and co-authorship relationships.

Figure 1: Entity relationship fusion.



Types of multi-relationships in topic recognition

According to the semantic distance between subject terms, we divided the relationships for topic recognition into three types: basic, enhancing, and supplement.

- Basic relationship

It is our opinion that subject terms appearing in the same article have the closest semantic relationship. Therefore, the direct co-occurrence relationship of subject term nodes, based on the same literature, is the foundation for the relationship analysis of the subject terms.

- Enhancing relationship

It is not only subject terms co-occurring in the same literature that can associate with each other; typically, subject terms that do not appear in the same literature can be linked through other text entities. This type of relationship is weaker than the basic relationship; however, it also suggests the semantic relation between subject terms and strengthens the basic relationship.

- Supplement relationship

Apart from the above two relationships, there exists an indirect relationship between subject terms, that is, subject terms cannot be directly associated with each other through text entities. Rather, they require a specific measurement entity as the intermediate bridging point. Then, the subject terms can associate with each other through the intermediate node. This type of indirect relationship is commonly neglected in generic thematic relation analysis as it is weaker than a basic or enhancing relationship.

Multilayer Relation Fusion

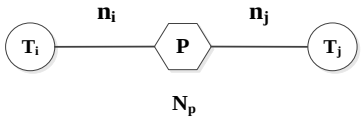
(1) Acquisition of relation matrix and links' weight

This paper proposes the calculation method of direct and indirect relations between measurement entities, which is based on Morris's definition of the weight of links of multiple

relationships of the measurement entities (Steven A. Morris, 2005; Steven. A Morris & Yen, 2005). According to the relationships of the literatures and the three different types of nodes, which incorporate subject terms, author, and citation, this study constructs seven types of relation matrices to represent the basic, enhancing, and supplement relationships between subject terms.

- Basic relationship matrix

Figure 3: Weight of links of basic relationship



Assume that Q is the total literature set, subject terms, T_i and T_j, appear in the same literature P, and the literature set where T_i and T_j satisfy this condition is Q_0. The frequency of T_i in P is n_i, the frequency of T_j in P is n_j, and the frequency of all subject terms in P is N_p (Figure 3). Assume that there are k literatures in Q_0, that is, T_i and T_j co-occur in k literatures. Then, in Q_0, the links' weight of T_i and T_j is given as follows:

$$d_{ij} = \sum_1^k ({}^{n_i}/_{N_P} + {}^{n_j}/_{N_P}) \qquad ①$$

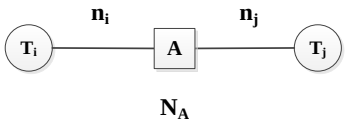
where ${}^{n_i}/_{N_P}$ represents the correlation degree between T_i and literature P. The higher the value, the stronger the correlation between T_i and literature P. Similarly, ${}^{n_j}/_{N_P}$ represents the correlation degree between T_j and literature P. The higher the value, the stronger the correlation between T_j and literature P, while $p \in [1, k]$. The higher the value of d_{ij}, the greater the overall correlation degree between T_i and T_j in the literature set Q_0.

According to ①, the basic relation links' weight between any two subject terms in the total literature set Q can be obtained. The links' weight can be used as a correlation coefficient of the subject terms to form the subject terms relation matrix. This matrix is the basic relation matrix.

- Enhancing relationship matrix

Use the author as an intermediate bridging point as an example.

Figure 4: Weight of links of enhancing relationship



Assume that subject terms, T_i and T_j, do not appear in the same literature yet have a common author A. Q_1 is the literature set where T_i and T_j satisfy the above condition. The frequency of association between T_i and author A (in Q_1, the frequency of $\boldsymbol{T_i}$ written by author A) is n_i, and the frequency of association between T_j and author A (in Q_1, the frequency of T_j written by author A) is n_j. In the total literature set Q, the sum of the frequencies of all subject terms written by author A is N_A (Figure 4). If there are k authors satisfying the condition of author A in

literature set Q_1, that is, any one of the k authors is a co-author of T_i and T_j, then in Q_1, the links' weight of T_i and T_j is as follows:

$$d_{ij} = \sum_1^k ({}^{n_i}/_{N_A} + {}^{n_j}/_{N_A}) \qquad ②$$

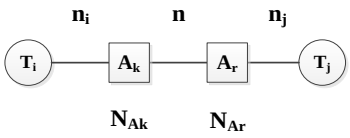
where ${}^{n_i}/_{N_A}$ represents the correlation degree between T_i and author A. The higher the value, the stronger the correlation between T_i and author A. Similarly, ${}^{n_j}/_{N_A}$ represents the correlation degree between T_j and author A. The higher the value, the stronger the connection force between T_j and author A. A$\in [1,k]$, d_{ij} is the correlation degree between T_i and T_j constructed by the author. The higher the value, the stronger the correlation.

According to ②, the enhancing relation links' weight between any two subject terms based on author in the total literature set Q can be obtained. The relation link weight can be used as a correlation coefficient of subject terms to form the subject terms relation matrix, which is the enhancing relation matrix using author A as the intermediate bridging point.

- Supplement relationship matrix

Use the author as an intermediate bridging point as an example.

Figure 5: Weight of links of supplement relationship



Assume that subject terms, T_i and T_j, do not appear in the same literature or share the same author, yet T_i's author, A_k, and T_j's author, A_r, have a co-authorship. Q_2 is the literature set where T_i and T_j satisfy the above conditions. The frequency of association between T_i and author A_k (in Q_2, the frequency of all T_i written by author A_k) is n_i, and the frequency of association between T_j and author A_r (in Q_2, the frequency of all T_j written by author A_r) is n_j. The frequencies of all subject terms where A_r collaborated with A_k is n. In the entire literature set Q, the sum of the frequencies of all the subject terms written by author A_k is N_{Ak}, and the sum of the frequency of all the subject terms written by author A_r is N_{Ar} (Figure 5). If there are t kinds of combinations satisfying the conditions of A_k and A_r in Q_2, that is, T_i and T_j can be related by any one of the t types of A_k and A_r combinations, then in Q_2, the links' weight of the supplement relationship between T_i and T_j taking author A as an intermediate bridging point is as follows:

$$d_{ij} = \sum_1^t ({}^{n_i}/_{N_{Ak}} \times {}^{n}/_{N_{Ak}} + {}^{n_j}/_{N_{Ar}} \times {}^{n}/_{N_{Ar}}) \qquad ③$$

where ${}^{n_i}/_{N_A}$ represents the correlation degree between T_i and author A. The higher the value, the stronger the connection force between T_i and author A. Similarly, ${}^{n_j}/_{N_A}$ represents the correlation degree between T_j and author A. The higher the value, the stronger the connection force between T_j and author A. The combinatorial number C (A_k, A_r) $\in [1,t]$, d_{ij} indicates the correlation degree between T_i and T_j using authors as the intermediate bridging point. The higher the value, the stronger the correlation.

According to ③, the supplement relation link weight between any two subject terms based on the author in the entire literature set Q can be obtained. The relation link weight can be used as a correlation coefficient of subject terms to form the subject terms relation matrix. This matrix is the supplement relation matrix using authors as the intermediate bridging point.

(2) Relation fusion
This study uses the PathSelClus algorithm to fuse seven types of relation matrices and calculate the comprehensive similarity of the comprehensive subject terms (Sun et al., 2013). The model clusters a certain type of target node based on heterogeneous multi-relationship. As the clustering is conducted under the guidance of a user labeling a small amount of the data results, which introduces the user's intent into the model by indexing the seed object, it is more in line with the user's intentions. Simultaneously, different relation matrices combined with their weights can better analyze the source of the semantic relationships of the clustering results; thus, they are more interpretable.

Empirical Analysis

Data acquisition
This study selected the GEV field as an empirical study. The data was collected from the database of China Knowledge Resource Integrated Database. The retrieval was performed in June 2017; 4,315 research papers were obtained.

Data processing
- Conduct Data preprocessing
- Extract the relation matrices
- Perform multilayer relation fusion
- Perform relation clustering and topic naming

Table 1 displays the partial clustering topics.

Table 1: Clustering topics.

Topic 1	Construction and expression of antineoplastic nucleic acid vaccine
Topic 2	Suicidal DNA vaccine and immune response
Topic 3	Construction of a dual-promoter DNA vaccine vector and immune response
Topic 4	Research on nucleic acid vaccine associated with avian infections
Topic 5	Manufacturing of anti-idiotype monoclonal antibody vaccine

Comparative analysis
In order to test the advantages of topic recognition based on multilayer relation fusion compared to a single relationship, we further carried out a comparative analysis. The results of comparative analysis show that there are typically an excessive number of extraction subject words in single co-word clustering and the extraction words contained in each topic are present more than once compared to those in the relation fusion method. 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. Similarly, an excessive number of clustering results are formed in each time window, resulting in insufficient effective clustering. The relation fusion method is better than single co-word clustering, and the degree of difference between the topics is evident.

Conclusion

This paper reviewed the research status of multilayer relationship fusion in topic recognition. According to the difference of semantic distance, the subject terms' relation in topic recognition was divided into basic, enhancing, and supplement relationships. Further, multilayer relationship extraction and relation fusion methods for topic recognition were proposed. The fusion relationship matrix could represent three types of relationship information: co-occurrence of the subject terms, coupling relationship between the authors and the subject terms, and coupling relationship between the citations and the subject terms. In this manner, the limitation of the co-word analysis and the lack of standardization of citation analysis was resolved. Empirical analysis proved that multilayer relationship fusion can effectively improve the effect of topic clustering.

References

Calero-Medina, C., & Noyons, E. C. (2008). Combining mapping and citation network analysis for a better understanding of the scientific development: The case of the absorptive capacity field. *Journal of Informetrics, 2*(4), 272-279.

He, X., Zha, H., Ding, C. H., & Simon, H. D. (2002). Web document clustering using hyperlink structures. *Computational Statistics & Data Analysis, 41*(1), 19-45.

Janssens, F. (2007). Clustering of scientific fields by integrating text mining and bibliometrics.

Janssens, F., Glänzel, W., & De Moor, B. (2007). *Dynamic hybrid clustering of bioinformatics by incorporating text mining and citation analysis.* Paper presented at the Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining.

Janssens, F., Zhang, L., De Moor, B., & Glänzel, W. (2009). Hybrid clustering for validation and improvement of subject-classification schemes. *Information Processing & Management, 45*(6), 683-702.

Morris, S. A. (2005). Unified Mathmatical Treatment of Complex Cascaded Bipartite Networks: The Case of Collections of Journal Papers. *Paperlandia*.

Morris, S. A., & Yen, G. G. (2005). Construction of bipartite and unipartite weighted networks from collections of journal papers. *Physics*.

Sun, Y., Norick, B., Han, J., Yan, X., Yu, P. S., & Yu, X. (2013). PathSelClus: Integrating Meta-Path Selection with User-Guided Object Clustering in Heterogeneous Information Networks. *ACM Transactions on Knowledge Discovery from Data (TKDD), 7*(3), 11.

Van Den Besselaar, P., & Heimeriks, G. (2006). Mapping research topics using word-reference co-occurrences: A method and an exploratory case study. *Scientometrics, 68*(3), 377-393.

Wen, B., Horlings, E., Mariëlle, V. D. Z., & Peter, V. D. B. (2016). Mapping science through bibliometric triangulation: An experimental approach applied to water research. *Journal of the Association for Information Science & Technology*.

Xu, H.-Y., Yue, Z.-H., Wang, C., Dong, K., Pang, H.-S., & Han, Z. (2017). Multi-source data fusion study in scientometrics. *Scientometrics, 111*(2), 773-792. doi:10.1007/s11192-017-2290-5

Xu, H., Dong, K., Liu, C., Wang, C., & Wang, Z. (2017). A review on topic identification to scientific text files. *INFORMATION SCIENCE, V35*(1), 153-160.

Xu, H., Dong, K., Wang, C., Wei, L., & Yue, Z. (2018). Research on Multi-source Data Fusion Method in Scientometrics. *Journal of The China Society for Scientific and Technical Information, in press*.