MEDINFO 2021: One World, One Health – Global Partnership for Digital Innovation
P. Otero et al. (Eds.)

doi:10.3233/SHTI220080

Hybrid Methods of Bibliographic Coupling and Text Similarity Measurement for Biomedical Paper Recommendation

Hongmei Guo[a], Zhesi Shen[b], Jianxun Zeng[a], Na Hong[c]

[a] *Institute of Scientific and Technical Information of China, Beijing, China*
[b] *National Science Library, Chinese Academy of Sciences, Beijing, China*
[c] *Digital China Health Co., Ltd., Beijing, China*

Abstract

The amount of available scientific literature is increasing, and studies have proposed various methods for evaluating document-document similarity in order to cluster or classify documents for science mapping and knowledge discovery. In this paper, we propose hybrid methods for bibliographic coupling (BC) and linear evaluation of text or content similarity: We combined BC with BM25, Cosine, and PMRA to compare their performances with single methods in paper recommendation tasks using TREC Genomics Track 2005datasets. For paper recommendation, BC and text-based methods complement each other, and hybrid methods were better than single methods. The combinations of BC with BM25 and BC with Cosine performed better than BC with PMRA. The performances were best when the weights of BM25, Cosine, and PMRA were 0.025, 0.2, and 0.2, respectively, in hybrid methods. For paper recommendation, the combinations of BC with text-based methods were better than BC or text-based methods used alone. The choice of method should depend on the actual data and research needs. In the future, the underlying reasons for the differences in performance and the specific part or type of information they complement in text clustering or recommendation need to be examined.

Keywords:
Citation-based methods; text-based methods; hybrid methods.

Introduction

Since the development and growth of the internet, the volume of electronic publication has been increasing at an exponential rate. Specifically, the publication of scientific literature in the biomedical domain has been prolific, and huge amounts of information and knowledge are found in endless unstructured electronic texts. Therefore, it is difficult for users to quickly find the relevant information that they require. Many knowledge discovery and retrieval techniques have been proposed by repositories and scientists to help users better retrieve the information they need. Among these techniques, document-document similarity measurements are the most fundamentals. Accurate similarity measurement can improve retrieval performance and help users manage the tremendous volume of relevant documents returned from search strategies. Researchers have explored various ways to effectively measure document-document similarity, which have been widely used in text clustering, information retrieval, scientific mapping, topic detection, and tracking. On the basis of the information used for calculating document-document similarity, these methods can be divided into 3groups: citation-based methods, text-based methods, and hybrid methods, which combine citation relationship with content analysis.

Citations are ubiquitous in scientific articles and they can reveal topic evolution and inheritable knowledge. Citation measurements are used in many tasks, including text clustering and classification[1], science mapping[2],research front detection[3,4], and information retrieval. Citation-based similarity mainly considers direct citation, co-citation, and bibliographic coupling (BC). If paper A is shown in the reference list of paper B, we know that A was cited by B and a direct citation relationship exists between papers A and B. If paper A and paper B are both cited by paper C, the connection between paper A and paper B is a co-citation, as proposed by Small[5]. If paper A is shown in the reference lists of both paper B and paper C, the connection between paper B and paper C is BC, as proposed by Kessler[6]. Simply, co-citation considers the in-links, while BC considers the common out-links between 2 papers. Amsler [7] combined BC with co-citation to classify papers, which considered the in-links and out-links simultaneously. The similarity measured by direct citation, co-citation, and BC differ in that they are suitable for different datasets and research objects [8]. Klavans et al compared the accuracy of the 3citation-based approaches in science mapping [2] and knowledge generation [9]. Subelj et al presented a systematic comparison of the performances of a large number of clustering methods based on citation relationships from 4criteria to compare the clustering methods, cluster sizes, small clusters, clustering stability, and computing time [10]. Citation-based methods can reveal the objective relationships between articles with discriminating power; however, they may not provide enough information regarding the context between articles.

Among text-based methods, co-word methods are the most popular for assessing content similarity. Keywords represent the core points and research focus of literature, so many researchers use co-word analysis to measure document similarity or label topics. Although co-word analysis has a relatively short history, it has become one of the most popular text-based methods and has been used in text clustering [11], the exploration of research topics and the evolution of psychiatry [12] and Hepatitis B [13], and the mapping of the knowledge structure and theme trends in the fields of patient adherence [14], disaster medicine [15], and choroidal neovascularization [16]. In previous studies, researchers used textual statistical features based on term frequency-inverse document frequency(TF-IDF) to measure similarity between documents. However, using term frequency to explain the importance of terms was not sufficient [17]: some high-frequency words extracted from texts were often common words that did not represent the specific research topics. As a result, using common high-frequency co-words to measure similarity between documents was not comprehensive.

Citation links and text similarity measure 2aspects of documents: they each offer their own advantages and can complement each other for many applications [18]. The combination

of citation-based and text-based methods can be used to cluster all papers, even poorly cited or uncited papers [19]. Recently, scientometricians have been studying hybrid methods that combine citation-based and text-based methods to improve the performance of scientific mapping and clustering. Liu et al presented the hybrid method of lexical and citation metrics to cluster large-scale journal data and chose InCites Essential Science Indicators (ESI) from Web of Science Database categorizations as standard [20, 21]. Glänzel and Thijs used the cosine angles of the linear combination of BC and TF-IDF similarities between documents to identify "core documents" and then applied them to label clusters or emerging topics [22, 23]. Yu et al extended the hybrid clustering model of Glänzel and Thijs and proposed a hybrid self-optimized clustering model, which considered both BC and co-citation links in the Amsler network to calculate the citation-based similarity of both TF-IDF and topological features to measure text-based similarity and then applied the Louvain method to cluster papers and detect the research topic in the data envelopment analysis (DEA) field [24]. Janssens et al merged textual contents with citations to cluster papers based on Fisher's inverse chi-square test to reveal the concept structure and dynamics in the field of bioinformatics [25]. Liu integrated BC with context information from paragraphs to measure the similarity between biomedical articles [26]. Leydesdorff et al considered both cited references and MeSH terms as attributes of articles and combined the 2knowledge representations to cluster and map papers in Alzheimer's disease [27].Kolchinsky et al classified protein-protein interactions based on text features and citation networks [28]. Meng et al proposed a simple linear combination of cosine similarity based on the TF-IDF of the terms with citation-based similarity based on citation relationships to cluster journals [29].

Previous studies tried to evaluate the performance of different document-document similarity methods in text clustering or classification. Since methods with different parameter settings yield different results, a benchmark is needed to represent the gold standard of clustering or classification results. Methods for building a benchmark can be classified as expert evaluation or text-based classification. Ahlgren et al extracted the title and abstract from 43 articles published from 2004 to 2006 using a journal information retrieval system: an information retrieval expert performed a subject classification for each article and assigned a label to each class. The classification results of the 43 articles were established as a standard of classification [30, 31] when comparing nine document-document similarity methods. Couto et al trained classifiers of 3 collections in the use of k-Nearest Neighbor and Support Vector Machine methods and used the classifiers as a benchmark [18]. Yu et al followed this benchmark in their hybrid self-optimized clustering model research [32]. Simply, no benchmark (or at least only a small benchmark) was available in previous studies.

In this study, we focused on verifying hybrid methods of BC and text similarity measurement and comparing the effect of different document-document similarity methods in paper recommendation. We used large data sets that had been evaluated and labeled by experts to verify two aspects of topic level: whether BC and text-based methods can complement each other and which hybrid methods are best for paper recommendation.

Methods

Datasets

We used the TREC Genomics Track 2005, which has been evaluated by experts [33], as the dataset in our study. This TREC dataset contained 34,633 unique documents indexed in the PubMed database and grouped into 50 topics corresponding to different information needs. The 50 topics were numbered from 100 to 149 and all generally followed a semantic template, with 10 topics in each of the 5 templates [34].For example, topic groups 100 to 109 focused on standard methods or protocols for doing some sort of experiment or procedure. Each topic corresponded to a different subset of documents ranging in size from 290 to 1356 documents. In each group, documents were marked as definitely relevant, possibly relevant, or non-relevant to the group's topic by experts.

Among the 34,633documents, 4191 unique papers were judged as possibly relevant or definitely relevant. We downloaded the 4191 papers from the PubMed database using the PMID identifier in the MEDLINE format. We gathered citation and reference information from Web of Science (WoS). In all, 3098 of the 4194 papers were indexed in the WoS database and we downloaded the citation and reference information. Finally, we identified both title-abstract text and references of the 3098documentsforfurther validation. The process is shown in Figure 1.

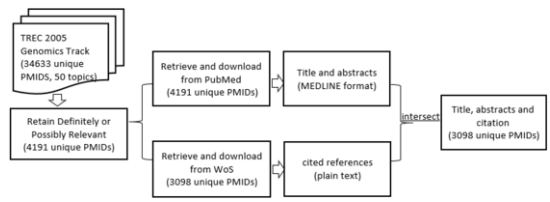

Figure 1-Flowchart for building the datasets

Hybrid methods of evaluating citation links and text similarity

Citation-based methods and text-based methods for assessing document similarity have their own strengths and weaknesses. Citation-based methods can reveal reference-citation links and focus on existing citation relationships between documents, but they neglect the associations of content features between documents. Text-based methods consider content features but neglect the citation relationships between documents. We linearly combined citation similarity with text similarity in this study. For citation links, we chose BC, which is the most common measure of similarity between articles. If 2documents cited 1or more of the same documents, they were bibliographically coupled and the coupling strength was formulated as Eq (1) :

$$BC\_sim(c,d) = N_{\text{share}} \Big/ \sqrt{N_c * N_d} \quad (1)$$

where N_{share} was the number of shared references for document c and document d, N_c was the number of references in document c, and N_d was the number of references in document d.

For text similarity, we selected 3common term-similarity metrics: PubMed Related Article (PMRA) [34, 35], BM25 [36], and Cosine [37]. PMRA was formulated as Eq (2), BM25 was formulated as Eq (3), and Cosine was formulated as Eq (4):

$$PMRA\_sim(c,d) = \sum_{t=1}^{n} wt_{t,c}.wt_{t,d} \quad (2)$$

$$(wt = [1+\eta(\frac{\mu}{\lambda})^{\kappa} e^{-(\mu-\lambda)l}]^{-1}.\sqrt{idf_t}$$

$\kappa = tf(t,a) - 1)$

where l was the length of the document in words, was a term's weight with respect to a particular document, idf_t was the inverse document frequency of term t, $tf(t,a)$ was the frequency of term t in the document a, μ and λ were used as the optimal values.

$$BM25(t,a) = \sum_{t=1}^{n} idf(t).\frac{tf(t,a).(\kappa_1+1)}{tf(t,a)+\kappa_1.(1-b+b.\frac{|a|}{avgdl})} \quad (3)$$

$$idf(t) = \log\frac{N-n(t)+0.5}{n(t)+0.5}$$

where $idf(t)$ was the inverse document frequency weight of term t, was the term frequency of t in document a, $|a|$ was the length of the document a in words, $avgdl$ was the average document length in the text collection, κ_1 a and b were free parameters, N was the total number of documents in the collection, and $n(t)$ was the number of documents containing term t.

$$cosine\_sim(c,d) = \cos(\Theta) = \frac{c.d}{\|c\|\|d\|} = \frac{\sum_{t=1}^{n} c_{(t)} \times b_{(t)}}{\sqrt{\sum_{t=1}^{n}(c_{(t)})^2} \times \sqrt{\sum_{t=1}^{n}(d_{(t)})^2}}$$

$$cosine\_sim(c,d) = \frac{\sum_{t=1}^{n} tf(t,c)idf(t)tf(t,d)idf(t)}{\sqrt{\sum_{t=1}^{n} tf(t,c)idf(t)^2}.\sqrt{\sum_{t=1}^{n} tf(t,d)idf(t)^2}} \quad (4)$$

where $tf(t,c)$ was the term frequency of t in document c and $idf(t)$ was the inverse document frequency weight of term t.

Castro et al compared the performance of the three similarity metrics (BM25, PMRA, and Cosine) based on Unified Medical Language System (UMLS) annotations for 4191 unique documents that were considered relevant or partially relevant [38]. In order to simplify the calculation, we used their research results of BM25, PMRA, and Cosine directly in this study.

In this paper, we propose hybrid methods of linearly combined citation-based methods and text-based methods at different λ values, which is shown in Eq (5):

$$Hybrid\_sim(c,d) = (1-\lambda).BC\_sim(c,d) + \lambda.Text\_sim(c,d)$$
$$\forall Text\_sim(c,d) \in (PMRA\_sim(c,d), BM25\_sim(c,d), Cosine\_sim(c,d)) \quad (5)$$

where λ was the coefficient of text-based methods and the range of λ was [0 1].

Evaluation indicator

In order to compare and select the best performing algorithm in the recommended documents, we used the Area Under the Precision-Recall (AUPR)curve to measure the performance of single and hybrid similarity methods. The AUPR curve plotted precision against recall and showed the trade-off between precision and recall for different thresholds. The larger the area under the curve, the higher the recall and precision of the algorithm.

Results

Document recommendation is one of the main tasks of information retrieval. In part, documents in the same topic area should be recommended on the basis of precalculated similarities.

Mutual complementary value of citation-based similarity and text-based similarity assessments

Citation-based and text-based similarities reflect the relationships between 2dimensions of different documents. In Figure 2, we show the correlation between citation-based similarity and text-based similarity for a randomly selected target document against other documents (i.e., BC vs. PMRA, BC vs. Cosine, and BC vs. BM25). The red dots correspond to articles within the same TREC topic as the target document and the blue dots represent articles from other topics. The majority of the red dots had large values in both similarity measures, implying the intrinsic similarity between documents in the same topic group (Figure 2(a)). There were some blue dots with large values in content-based similarity, but these papers had no shared references (i.e., BC was zero). Therefore, combining citation-based and text-based measurements may significantly decrease the similarity values and reduce the effect of these non-related documents in the recommendation. Some documents in the same topic group had low or even zero BC, but their similarity was enhanced by considering content-based similarity. Similar phenomena are shown in Figure 2 (b) and 2 (c).

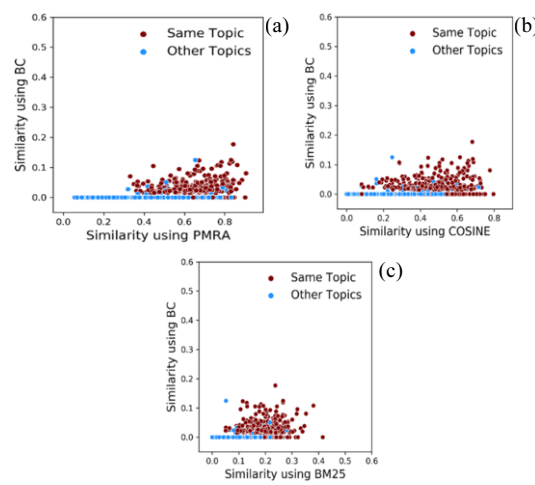

Figure 2-The correlation of similarity between citation-based measures (i.e., BC) and text-based measures (a) PMRA, (b) COSINE, and (c) BM25) for a randomly selected document (PMID:11445269) against documents from the same topic group and other topics

Comparison of different hybrid methods in paper recommendation

From Figure 2, we knew that citation-based methods and text-based methods can complement each other. Specifically, BC and 3text-based methods can be combined at different weighted values. We used the AUPR indicator to measure recall and precision of hybrid methods in paper recommendations. Figure 3 shows the average AUPR of hybrid methods at different λ values. When λ=0, the recommendation was fully based on BC; when λ=1, only the text-based methods were used. For BC alone, the AUPR value was about 0.55. For BM25 and COSINE alone, AUPR values were about 0.62 and 0.66, respectively. When λ increased in value, it signified a decreased weight of BC and an increased weight of BM25 or COSINE similarity. AUPR values first increased and then decreased slightly with increasing λ (Figure 4). For hybrid methods of BC withBM25 and BC with COSINE, AUPR values were higher than those of the similarity methods used alone. For hybrid methods of BC and PMRA, most of the AUPR values were higher than the single PMRA but lower than the single BC.

For all three text-based methods, citation-based assessments of similarity improved the recommendation performance, especially for PMRA. Combining text-based similarity with BC could also improve the performance of BC, even in cases in which the text-based approach alone had low performance. The values of AUPR changed with λ values: they initially rose and then fell. We calculated the λ values when AUPR for hybrid methods reached their maximums: for BC with PMRA, the AUPR reached its maximum when λ was 0.025; for BC with BM25 and BC with Cosine, both AUPRs reached their maximums when λ was 0.02.

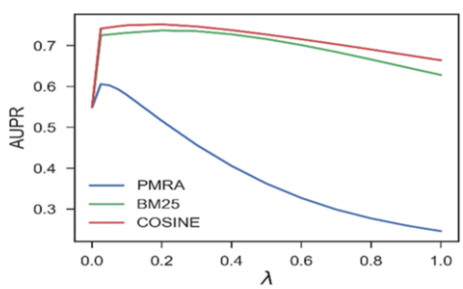

Figure 3-Average AUPR of hybrid methods with different values.

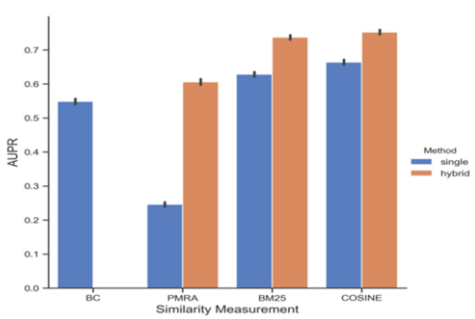

Figure 4-Performance comparison between single and hybrid similarity measurements

Figure 4 showed a comparison of the performance of single and hybrid similarity methods at maximum AUPR. For the hybrid methods, the best λ was chosen (i.e., when AUPR was at its maximum). For PMRA, BM25, and Cosine, the values of were0.025, 0.02, and 0.02, respectively. Blue columns represent single methods and yellow columns represent hybrid methods of BC and text-based measurements. The AUPR values of all hybrid methods were higher than those of any single method. The black line on each column corresponds to its 95% confidence interval (CI): the 95% CIs for each single method and its hybrid method did not intersect, which demonstrates that the difference in AUPR between each single method and its hybrid method was statistically significant. For document recommendation, hybrid methods were better than single methods.

Discussion

Citation-based methods have advantages in the ability to distinguish among papers. However, it is impossible for researchers to cite all relevant documents, as a massive amount of literature is available. Documents that were cited by one paper may be positive or negative references, and references located in different paragraphs play different roles in regards to the citing document. For example, references in the Results or Discussion sections are more important than those in the Introduction or Methods sections. As a result, it is not sufficient to measure the similarity between documents only on the basis of citation links.

In this study, we explored hybrid methods of BC and text similarity in paper recommendation, and we compared the performances of single BC, BM25, Cosine and PMRA methods with hybrid methods of BC linearly combined with BM25, Cosine, and PMRA using TREC Genomics Track 2005 datasets. Overall, BC and text-based methods can complement each other. For paper recommendation, the performances of all hybrid methods were better than any of the single methods alone. Among the hybrid methods, the performance of BC combined withBM25 and BC combined with Cosine were better than BC combined with PMRA. The hybrid methods performed best when the weights of BM25, Cosine, and PMRA were 0.025,0.2, and0.2, respectively.

Conclusions

Citation-based methods and text-based methods may be suitable for different analysis granularities or levels. For topic level, text content is easier to identify, so text-based methods would be better. However, at middle or macro levels, such as journal or subject level, BC may be better than text-based content methods of evaluation. At different levels, citation-based methods and text-based methods play different roles in document recommendation or clustering. In the future, we plan to test the hybrid methods at more levels and discern which method's weight is higher and identify how they can be used together.

In this study, we linearly combined BC with text-based methods, but, perhaps other hybrid methods are better. Document recommendation results depend on actual experimental data and research objective, so hybrid methods in various datasets and under different user demands should be verified. For this study, we only used the titles and abstracts for methods based on text content, but some methods, such as PMRA, perform better when full-text analysis is used. These hybrid methods should be tested in more practical applications and larger data collections.

Acknowledgements

Ethics approval and consent to participate: Not applicable.

Consent for publication: Not applicable.

Competing interests: The authors declare that they have no competing interests.

Funding: The study was supported by the National Social Science Fund of China (17ATQ002).

Authors' contributions: Hongmei Guo designed the method, analyzed the results, and wrote the manuscript. Zhesi Shen performed the experiments, analyzed the results, and modified the manuscript. Jianxun Zeng and Na Hong analyzed the results and modified the manuscript.

The authors would like to acknowledge Per Ahlgren from Uppsala University for assistance and LeylaJael Garcia Castro and colleagues for sharing data about text similarity based on PMRA, BM25, and COSINE.

References

[1] Parthasarathy G, Tomar D C. A novel approach for classification and clustering of biomedical citations. *Biomedical Research-tokyo*.2016; 22-30.

[2] Boyack KW, Klavans R. Co-citation analysis, bibliographic coupling, and direct citation: Which citation approach represents the research front most accurately? *Journal of the Association Information Science and Technology*. 2010; 61(12): 2389-2404.

[3] Jarneving B. Bibliographic coupling and its application to research-front and other core documents. *J Informetric*. 2007;1(4):287–307.

[4] Huang M, Chang CP. A comparative study on detecting research fronts in the organic light-emitting diode (OLED) field using bibliographic coupling and co-citation. *Scientometrics*.2015;102(3): 2041-2057.

[5] Small H. Co-citation in the scientific literature: a new measure of the relationship between two documents. *J Am Soc Inf Sci*. 1973; 24(4):265–269.

[6] Kessler MM. Bibliographic coupling between scientific papers. *Am Doc*. 1963; 14(1):10–25.

[7] Amsler RA. Applications of citation-based automatic classification. 1st ed. Linguistics Research Center, University of Texas at Austin; 1972.

[8] Couto T, Ziviani N, Calado P, Cristo M, GoncËalves M, de Moura ES, Brandão W. Classifying documentswith link-based bibliometric measures. *Inf Retrieval.* 2010;13(4):315-345.

[9] Klavans R, Boyack K W. Which Type of Citation Analysis Generates the Most Accurate Taxonomy of Scientific and Technical Knowledge. *Association for information science and technology*.2017; 68(4): 984-998.

[10] Subelj L, Van Eck N J, Waltman L, et al. Clustering Scientific Publications Based on Citation Relations: A Systematic Comparison of Different Methods. *PLOS ONE*. 2016;11(4).

[11] Rashmi G Dukhi,Pratibha Mishra.A New Hybrid Feature Selection using Natural Language Processing for Text Clustering. *ICRTIET-2014 Conference Proceeding*. 2014;36-39

[12] Wu Y, Jin X, Xue Y, et al. Evaluation of research topic evolution in psychiatry using co-word analysis. *Medicine*.2017;96(25).

[13] W. Zhang, Y.B. Wang, X.Z. Zhang, H.M. The Study of Hot Spots on Hepatitis B Dissertation Based on Co-Word Analysis in China.*MEDINFO 2017: Precision Healthcare through Informatics*.2017; 245: 1293 – 1293

[14] Zhang J, Xie J, Hou W, et al. Mapping the Knowledge Structure of Research on Patient Adherence: Knowledge Domain Visualization Based Co-Word Analysis and Social Network Analysis. *PLOS ONE*.2012;7(4).

[15] Zhong-Yi Wang, Gang Li, Chun-Ya Li, Ang Li. Knowledge Maps of Disaster Medicine in China Based on Co-Word Analysis. *Scientometrics*. 2012; 90(3): 855–875.

[16] Zhao F, Shi B, Liu R, et al. Theme trends and knowledge structure on choroidal neovascularization: a quantitative and co-word analysis[J]. *BMC Ophthalmology*. 2018; 18(1):86-97.

[17] Chen G, Xiao L. Selecting publication keywords for domain analysis in bibliometrics: a comparison of three methods. *J Informetr*ic. 2016; 10(1):212-223.

[18] Couto T, Ziviani N, Calado P, et al. Classifying documents with link-based bibliometric measures. *Information Retrieval*.2010;13(4): 315-345.

[19] Ren X, Liu J, Yu X, et al. ClusCite: effective citation recommendation by information network-based clustering. *knowledge discovery and data mining*.2014;821-830.

[20] Liu X, Yu S, Moreau Y, et al. Hybrid Clustering by Integrating Text and Citation Based Graphs in Journal Database Analysis. *International conference on data mining*. 2009; 521-526.

[21] Liu X, Yu S, Janssens F A, et al. Weighted hybrid clustering by combining text mining and bibliometrics on a large-scale journal database. *Journal of the Association for Information Science and Technology*. 2010; 61(6): 1105-1119.

[22] GlaÈnzel W, Thijs B. Using ªcore documentsº for the representation of clusters and topics. *Scientometrics*.2011; 88(1):297-309.

[23] GlaÈnzel W, Thijs B. Using ªcore documentsº for detecting and labelling new emerging topics. *Scientometrics*.2012; 91(2):399-416.

[24] Yu D, Wang W, Zhang S, et al. Hybrid self-optimized clustering model based on citation links and textual features to detect research topics. *PLOS ONE*.2017;12(10).

[25] Janssens F A, Glanzel W, De Moor B, et al. Dynamic hybrid clustering of bioinformatics by incorporating text mining and citation analysis. *knowledge discovery and data mining*. 2007;360-369.

[26] Liu R. Passage-Based Bibliographic Coupling: An Inter-Article Similarity Measure for Biomedical Articles. *PLOS ONE*.2015;10(10).

[27] Leydesdorff L, Comins J A, Sorensen A A, et al. Cited references and Medical Subject Headings (MeSH) as two different knowledge representations: clustering and mappings at the paper level. *Scientometrics*.2016;109(3): 2077-2091.

[28] Kolchinsky A, Abihaidar A, Kaur J, et al. Classification of Protein-Protein Interaction Full-Text Documents Using Text and Citation Network Features. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*.2010; 7(3): 400-411.

[29] Meng X, Liu X, Tong Y, GlaÈnzel W, Tan S. Multi-view clustering with exemplars for scientific mapping. *Scientometrics*.2015; 105(3):1527±1552.

[30] Ahlgren P, Colliander C. Document-document similarity approaches and science mapping: experimental comparison of five approaches. *Journal of Informetrics*.2009;3(1): 49-63.

[31] Ahlgren P, Jarneving B. Bibliographic coupling, common abstract stems and clustering: A comparison of two document–document similarity approaches in the context of science mapping. *Scientometrics*. 2008;76(2):273–290.

[32] Yu D, Wang W, Zhang S, et al. Hybrid self-optimized clustering model based on citation links and textual features to detect research topics. *PLOS ONE*.2017;12(10).

[33] https://dmice.ohsu.edu/trec-gen/2005data.html

[34] Pitigala S, Li C. Extending PubMed Related Article (PMRA) for Multiple Citations[C]. *industrial conference on data mining*.2014;55-69.

[35] Lin J J, Wilbur W J. PubMed related articles: a probabilistic topic-based model for content similarity[J]. *BMC Bioinformatics*.2007;8(1): 423-423.

[36] S.E. Robertson, S. Walker, Some simple effective approximations to the 2-Poisson model for probabilistic weighted retrieval. *Proceedings of the 17thAnnual International ACM SIGIR Conference on Research and Development Information Retrieval*.1994; 232–241.

[37] D. Jannach, M. Zanker, A. Felfernig, G. Friedrich, The Cosine similarity measure in Recommender Systems: An Introduction. *Cambridge University Press*.2010;360-376.

[38] Castro L J, Berlanga R, Garcia A, et al. In the pursuit of a semantic similarity metric based on UMLS annotations for articles in PubMed Central Open Access. *Journal of Biomedical Informatics*.2015;204-218.

Address for correspondence

Hongmei Guo, guohm@istic.ac.cn, No.15 Fuxing Road, Beijing, 100038