Discovering Emerging Research Topics based on SPO Predications

Zhengyin Hu[1*], Rong-Qiang Zeng[1,2], Lin Peng[1], Hongshen Pang[3], Xiaochu Qin[4], and Cheng Guo[4]

[1] Chengdu Library and Information Center, Chinese Academy of Sciences, Sichuan 610041, P. R. China
{huzy,pengl}@clas.ac.cn
[2] School of Mathematics, Southwest Jiaotong University, Chengdu, Sichuan 610031, P. R. China
zrq@swjtu.edu.cn
[3] Shenzhen University, Shenzhen, Guangdong 518060, P. R. China
phs@szu.edu.cn
[4] Guangzhou Institutes of Biomedicine and Health, Chinese Academy of Sciences, Guangzhou, Guangdong 510530, P. R. China
{qin_xiaochu,guo_chen}@gibh.ac.cn

Abstract. With the rapid growth of scientific literatures, it is very important to discover the implicit knowledge from the vast information accurately and efficiently. To achieve this goal, we propose a percolation approach to discovering emerging research topics by combining text mining and scientometrics methods based on Subject-Predication-Object (SPO) predications, which consist of a subject argument, an object argument, and the relation that binds them. Firstly, SPO predications are extracted and cleaned from content of literatures to construct SPO semantic networks. Then, community detection is conducted in the SPO semantic networks. Afterwards, two indicators of Research Topic Age (RTA) and Research Topic Authors Number (RTAN) combined by hypervolume-based selection algorithm (HBS) are chosen to identify potential emerging research topics from communities. Finally, scientific literatures of stem cells are selected as a case study, and the result indicates that the approach can effectively and accurately discover the emerging research topics.

Keywords: emerging research topics, Subject-Predication-Object, community detection, hypervolume-based selection, stem cell.

1 Introduction

Emerging research topics represent the new areas of science and technology (S&T) in which the scientists are highly concerned. Actually, it is of great significance to mine these topics through S&T literatures for scientific research and policy making [1]. With the rapid growth of S&T literatures, it has become a big challenge to efficiently and accurately discover implicit knowledge from the vast literatures in a credible way.

Then, Knowledge Discovery in Literature (KDiL) has become an important research area. Indeed, it is very interesting and useful to combine text mining with scientometrics methods for KDiL.

SPO predication represents the semantic relationships among knowledge units, which consists of a subject argument (noun phrase), an object argument (noun phrase) and the relation that binds them (verb phrase) [2]. In fact, the SPO predication can be considered as a kind of semantic network widely used in KDiL, which can reflect research topics of literatures with semantic information and represent S&T information with more details.

In this paper, we propose a percolation approach to discovering emerging research topics based on SPO predications combining text mining and scientometrics methods. Firstly, SemRep [2] which is a Unified Medical Language System (UMLS)-based information extraction tool and Semantic MEDLINE [3] which is a SPO database generated by SemRep based on PubMed, are used to get SPO predications from biomedical text. The subject and object arguments of each SPO are the concepts from the UMLS Metathesaurus, and the Predicate is a relation from the UMLS Semantic Network [2, 3]. For example, from the sentence "We used hemofiltration to treat a patient with digoxin overdose that was complicated by refractory hyperkalemia", SemRep extracts four predications as follows: "Hemofiltration-TREATS-Patients, Digoxin overdose-PROCESS_OF-Patients, hyperkalemia-COMPLICATES-Digoxin overdose, Hemofiltration-TREATS (INFER)-Digoxin overdose" [2, 3].

Then, community detection is conducted in the SPO semantic networks, and a community containing SPO predications can be considered as a research topic. Afterwards, two scientometrics indicators of RTA and RTAN combined by HBS algorithm are chosen to find potential emerging research topics from communities. Finally, S&T literatures of stem cells are selected as a case study. The result indicates that the approach can effectively and accurately discover emerging research topics.

The rest of this paper is organized as follows. Section 2 briefly describes the previous works related to the discovery of emerging research topics. In Section 3, we present a percolation approach to discovering research topics based on SPO predications. Afterwards, we conduct a case study in Section 4. The conclusion and discussion about further research are given in the last section.

2 Literature and review

In this section, we investigate the literature reviews concentrating on discovering emerging research topics. Generally, they are divided into scientometrics methods and text mining methods.

The scientometrics methods usually use indicators analysis to discover emerging research topics based on citation or co-occurrence relationship. In [4], the authors proposed a multi-level structural variation approach, which is motivated by an explanatory and computational theory of transformative discovery. With the novel structural variation metrics derived from the theory, they integrated the theoretical framework with a visual analytic process, which enables an analyst to study the literature of

a scientific field across multiple levels of aggregation and decomposition, including the field as a whole, specialties, topics and predicates.

In [5], according to the co-cited networks of regenerative medicine literatures based on a combined dataset of 71,393 relevant papers published between 2000 and 2014, the authors presented a snapshot of the fast-growing fields and identified the emerging trends with new developments. Actually, the structural and temporal dynamics are identified in terms of most active research topics and cited references. New developments are identified in terms of newly emerged clusters and research areas, while disciplinary-level patterns are visualized in dual-map overlays.

In [6], the authors proposed a method of discovering research fronts, which compares the structures of citation networks of scientific publications with those of patents by citation analysis and measures the similarity between sets of academic papers and sets of patents by natural language processing. In order to discover research fronts that do not correspond to any patents, they performed a comparative study to measure the semantic similarity between academic papers and patents. As a result, cosine similarity of term frequency-inverse document frequency (tfidf) vector was found to be a preferable way of discovering corresponding relationships.

The text mining methods usually conduct content-analysis using domain ontology, semantic network, and community detection etc. to mine emerging research topics from the contents of literatures. In [7], based on Human Phenotype Ontology (HPO), the authors presented a method named RelativeBestPair to measure similarity from the query terms to hereditary diseases and rank the candidate diseases. In order to evaluate the performance, they carried out the experiments on a set of patients based on 44 complex diseases by adding noise and imprecision to be closer to real clinical conditions. In comparison with seven existing semantic similarity measures, RelativeBestPair significantly outperformed all other seven methods in the simulated dataset with both noise and imprecision, which might be of great help in clinical setting.

In [8], the authors proposed a multi-phase gold standard annotation approach, which was used to annotate 500 sentences randomly selected from MEDLINE abstracts on a wide range of biomedical topics with 1371 semantic predications. According to the UMLS Metathesaurus for concepts and the UMLS Semantic Network for relations, they measured inter-annotator agreement and analyzed the annotations, so as to identify some of the challenges in annotating biomedical text with relations based on ontology or terminology.

In [9], according to some semi-supervised learning methods named Positive-Unlabeled Learning (PU-Learning), the authors proposed a novel method to predict the disease candidate genes from human genome, which is an important part of nowadays biomedical research. Since the diseases with the same phenotype have the similar biological characteristics and genes associated with these same diseases tend to share common functional properties, the proposed method detects the disease candidate genes through gene expression profiles by learning hidden Markov models. The experiments were carried out on a mixed part of 398 disease genes from 3 disease types and 12001 unlabeled genes, and the results indicated a significant improvement in comparison with the other methods in literatures.

In [10], based on Formal Concept Analysis (FCA), the authors proposed a method named FCA-Map to incrementally generate five types of formal contexts and extract mappings from the derived lattices, which is used to identify and validate mappings across ontologies, including one-to-one mappings, complex mappings and correspondences between object properties. Compared with other FCA-based systems, their proposed method is more comprehensive as an attempt to push the envelope of the FCA formalism in ontology matching tasks. The experiments on large, real-world domain ontologies show promising results and reveal the power of FCA.

Both of the above-mentioned methods face specific challenges. Scientometrics methods are mature, but require that there are complete citation networks or high co-occurrence. If the citation networks are incomplete or not available, for example, the citation networks between papers and patents usually are very weak, the scientometrics methods cannot produce reasonable results. Text mining methods, which analysis fine-grained knowledge units such as keywords, SPO predications, and topics, do not need complete citation networks. However, usually the number of knowledge units is large and it is hard to clean and select the right ones without scientometrics methods.

3 Methodology

In our research, we propose a percolation approach to discovering emerging research topics based on SPO predications, which constructs a three-level SPO-based semantic network. First, we present an introduction to the construction of SPO-based semantic network. Then, we investigate a percolation approach to detecting communities in the network. Afterwards, we take the HBS algorithm on two indicators, RTA and RTAN, to identify potential emerging research topics from the communities.

3.1 SPO-based Semantic Network Construction

After getting required literatures set, SPO predications can be extracted from content of literatures by SemRep. Then, these SPO predications need to be cleaned by Term Clumping which includes general cleaning and pruning processes [11]. General cleaning will remove some common academic/scientific subjects or objects such as “cells,” “organ.” Some predicates such as “LOCATION_OF,” “PART_OF” that reflect hierarchy or position relationship and are meaningless for mining emerging research topics will also be removed. The pruning process helps with further cleaning by discarding the very low frequency and the meaningless subjects, predicates, objects or SPO. After that, each literature is represented as an exchangeable bag-of-SPO.

Based on four basic principles proposed by M. Fiszman et al., which are relevancy, connectivity, novelty and saliency, a SPO-based semantic network is constructed to detect the communities [12]. An example of SPO-based semantic network is illustrated in Fig. 1, which is composed of thousands of nodes and edges. In Fig. 1, the vertices with different colors denote the different SPO predications, and the size of the vertex denotes the frequency of SPO predication. Actually, many vertices can be both the subjects and the objects so that it makes the whole network become very compli-

cated [13]. Therefore, it is very difficult for the experts to recognize the valuable topics from a SPO-based semantic network directly.

Fig. 1. An example of SPO-based semantic network

3.2 Community Detection

In order to effectively find the communities from a SPO-based semantic network, we propose a percolation approach to achieve the community detection, which employs the widely used modularity function defined as follows [14]:

$$Q = \frac{1}{2m}\sum_{v\,w}\left[A_{vw} - \frac{k_v k_w}{2m}\right]\delta(C_v, C_w) \tag{1}$$

Suppose that the vertices are divided into different communities such that the vertex v belongs to the community C denoted by C_v. In Formula 1, A is the adjacency matrix of the network G. $A_{vw} = 1$ if one vertex v is connected to another vertex w, otherwise $A_{vw} = 0$. The δ function $\delta(i, j)$ is equal to 1 if $i = j$ and 0 otherwise. The degree k_v of a vertex v is defined to be $k_v = \sum_v A_{vw}$, and the number of edges in the network is $m = \sum_{wv} A_{wv}/2$.

Furthermore, the modularity function can be presented in a simple way, which is formulated below [14]:

$$Q = \sum_i (e_{ii} - a_i^2) \tag{2}$$

where i runs over all communities in the network, e_{ii} and a_i^2 are respectively defined as follows [14]:

$$e_{ij} = \frac{1}{2m}\sum_{vw} A_{vw}\delta(C_v, i)\delta(C_w, j) \tag{3}$$

which is the fraction of edges that join vertices in community i to vertices in community j, and

$$a_i = \frac{1}{2m}\sum_v k_v \delta(C_v, i) \tag{4}$$

which is the fraction of the ends of edges that are attached to vertices in community i.

Based on the modularity function optimization [15], the percolation approach is a heuristic method to extract the community structure of large networks, which is presented in Algorithm 1.

Table 1. The percolation algorithm to extract community structure

Algorithm 1. Percolation Algorithm
1: **Input**: SPO-based semantic relation network *G*
2: **Output**: Sequence of communities
3: Initialization (*G*)
4: Calculating the average weighted degree *d* of *G*
5: Calculating the weight of each edge of *G* multiplies by a random number with the probability $1 - 1/d$
6: $P = \{x_1, \ldots, x_p\} \leftarrow$ Random Initialization (*P*)
7: **repeat**
8: $x_i \leftarrow$ Local Search (x_i)
9: **until** a stop criterion is met

In this algorithm, we initialize the network into a directed weighted graph according to the SPO-based semantic relation. Then, we calculate the average weighted degree d of this graph. Afterwards, the weight of each edge multiplies by a random number with the probability$1 - 1/d$. Based on the modularity function value, the local search procedure is executed until the modularity does not improve any more. Then, we obtain the communities of the considered network. An example of the communities in a semantic network detected by the algorithm is illustrated in Fig. 2, in which the communities are represented in different colors.

3.3 Hypervolume-Based Selection

After finding communities in the SPO-based semantic network, we aim to select emerging research topics from these communities based on two scientometrics indicators, which are RTA and RTAN proposed in [16]. Specifically, RTA refers to time span of research topics, the larger RTA value is, the wider the time span of distribution of topics is. While RTAN refers to academic attentiveness, the larger RTAN value is, the hotter the topics are. Therefore, we prefer to select the topics with smaller values of RTA and larger values of RTAN as candidates of emerging research topics. RTA and RTAN are defined by the formulas below:

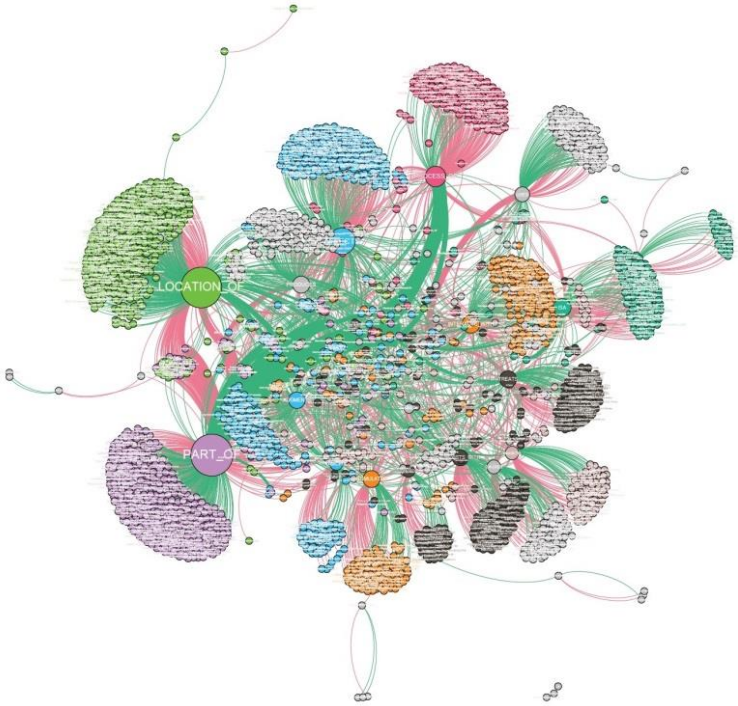

Fig. 2. An example of the communities in SPO-based semantic network

$$f_1 = RTA(topic_i) = \sum_{i=1}^{n} Y_{kw} \frac{n_i}{N} \quad (5)$$

where n_i refers to the number of terms in topic of the time span, *N* refers to the total number of terms in all topics of the time span and Y_{kw} refers to age of each term.

$$f_2 = RTAN(topic_i) = \frac{n_i}{N} \times 100\% \quad (6)$$

where n_i refers to the number of authors in topic$_i$ of the time span, and *N* refers to the total number of authors in all topics of the time span.

$$Y_{kw} = \sum_{i=1}^{n} (Year_{cur} - Year_i) \times tfidf_i / (\sum_{j=1}^{n} tfidf_j) \quad (7)$$

where $Year_{cur}$ refers to the last year of the time span, $Year_i$ refers to the year of the time span in all topics and $tfidf_i$ refers to the TF/IDF value of the i^{th} term.

According to the two objective values of f_1 and f_2, we select topics among the communities with the HBS algorithm, which is presented in Algorithm 2 below [17].

Table 2. The hypervolume-based selection algorithm

Algorithm 2. Hypervolume-Based Selection Algorithm
1: calculate two objective values of *topic*$_i$

2: calculate the fitness value of $topic_i$ with the *HC* indicator
3: select the topics based on the fitness values

In Algorithm 2, $topic_i$ denotes the i^{th} topic in the semantic relation network. First, we calculate the two objective values of $topic_i$. Then, we calculate the fitness value of $topic_i$ with the HC indicator defined as follows:

$$HC(x) = \big(f_1(y_1) - f_1(x)\big) \times \big(f_2(y_0) - f_2(x)\big) \tag{8}$$

As is shown in Fig. 3, the fitness value of $topic_i$ denoted by x corresponds to the size of the red area, where y_0 and y_1 refer to other topics, which are the neighbours of $topic_i$ in the objective space. Thus, we can select a designated number of research topics with high fitness values.

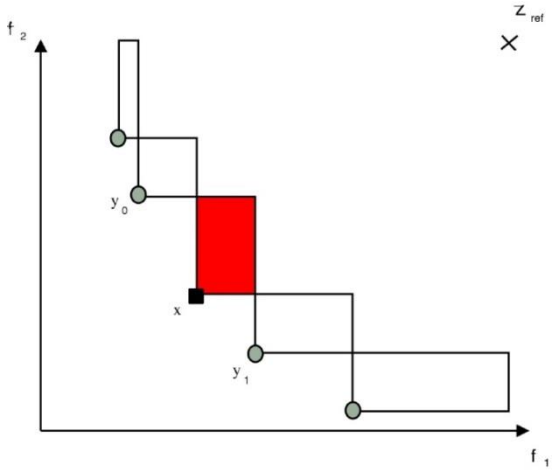

Fig. 3. An example of fitness value calculation

4 Case study

Stem Cells, which are a group of cells that are capable of self-renewal and multidirectional differentiation, are important research objects in biomedical area. Due to their important value and tremendous development prospects in the treatment of diseases and regenerative medicine, stem cells have drawn the worldwide attention and become the hot point of life science and medical research [18]. In this section, stem cells scientific papers are selected as the case study to demonstrate the approach.

4.1 Data Information

Initially, we selected PubMed as the data source, and made retrieval strategy as follow: "(stem cells [MeSH Major Topic]) AND ("2008-01-01"[Date - Publication]: "2017-12-31"[Date - Publication])", and 86,452 records were obtained. After excluding some non-technical papers such as “Clinical Trial,” “Dataset”, SPO predications

were extract from the *title* and *abstract* fields of each paper. Then, the general cleaning and pruning processes were applied to clean these SPO predications. After that, the SPO-based semantic network of stem cell was constructed according to the above method.

4.2 Experimental Results

In this subsection, we presented the experimental result. Some communities from the semantic networks are respectively illustrated in Fig. 4. In the figure, the different communities are represented in different colors, which are composed of subjects, objects and the corresponding predications, and the frequency of the SPO predications is proportional to the size of the vertex. From Fig. 4, we can observe that the predications "AFFECT, STIMULATE, COEXISTS_WITH" etc. are higher frequency. Some communities with these three predications are summarized in Table 3. In this table, we do not present all the found communities in the network but to provide parts of three different communities according to the HBS process, which are considered as emerging research topics in stem cell by experts.

Table 3. Some emerging research topics in stem cell

Predication	Subject	Object
AFFECT	Genes, MicroRNAs, Transcription Factor, …	Gene Expression, Cell Proliferation, stem cell division, …
	Signal Transduction Pathways, Signaling Molecule, …	Neuronal Differentiation, Cell Survival, Signal Transduction, …
	Proteins, Bone Morphogenetic Proteins, Therapeutic procedure, …	Growth, Wound Healing, Chondrogenesis, …
STIMULATE	Fibroblast Growth Factor, Transforming Growth Factor, …	Mitogen-Activated Protein Kinases, Collagen, Reactive Oxygen Species, …
	Erythropoietin, VEGF protein, bone morphogenetic protein, …	Osteogen, NOS3 protein, Granulocyte Colony-Stimulating Factor, …
	Transplantation, agonists, Antibodies, …	Blood flow, Antigens, Recovery, …
COEXISTS_WITH	Functional disorder, Endothelial dysfunction, Injury, …	Cardiovascular Diseases, Hypertensive disease, Diabetes, …
	Pathologic Neovascularization, Malignant Neoplasms, …	Myocardial Ischemia, Neoplasm, Obesity, …
	Down-Regulation, Fibrosis, Application procedure, …	Replacement therapy, Tissue Engineering, Cell Therapy, …

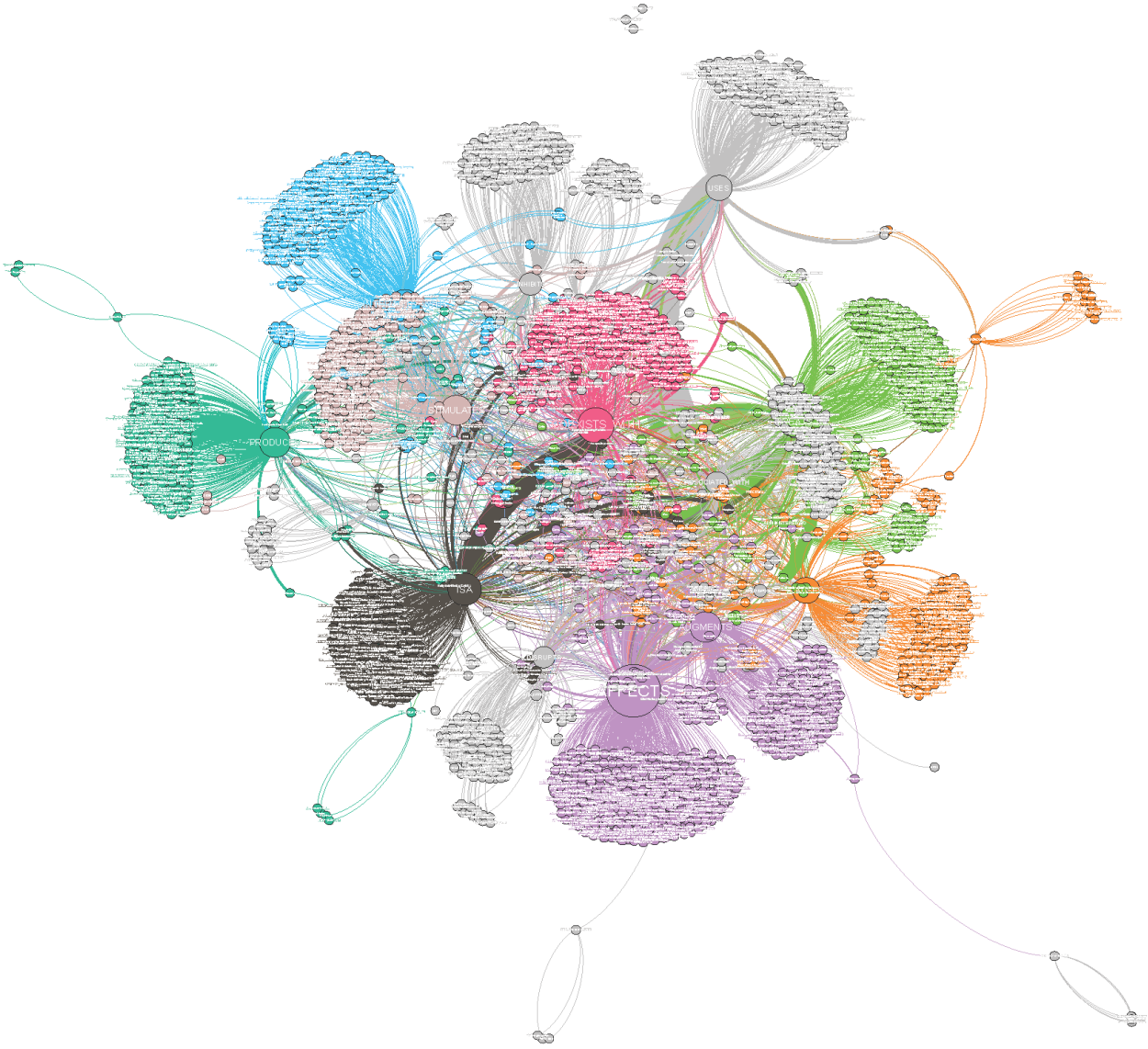

Fig. 4. Some communities from SPO-based semantic network of stem cell

5 Conclusion

In this work, we investigated a percolation approach to discovering emerging research topics from the SPO-based semantic network. Then, we perform the experiments in the research area of stem cells. The results indicate that it can help significantly discover the emerging research topics in the considered area.

However, there are some challenges in the processes. First, there are so many noises in the "Subjects, Objects, Predicates and SPO predications" from the papers and the general cleaning and pruning processing deeply depend on experts' opinions. Secondly, the SPO predications are extracted only from the *title* and *abstract* fields. Maybe it is not enough. Thirdly, the topics should be described in a more understandable way.

In the future, we intend to make more specific general cleaning and pruning rules to help conduct more objective data cleaning. In addition, we will directly extract SPO predications from full texts of papers by SemRep to get more SPO predications. Moreover, further research such as attaching understandable labels to topics, mining the linkages between topics, and discovering topics evolution are ongoing.

Acknowledgments

The work in this paper was supported by the Informationization Special Project of Chinese Academy of Sciences “E-Science Application for Knowledge Discovery in Stem Cells” (Grant No: XXH13506-203) and the Fundamental Research Funds for the Central Universities (Grant No. A0920502051815-69).

References

1. D. R. Swanson. Medical literature as a potential source of new knowledge. Bulletin of the Medical Library Association, 78(1):29–37, 1990.
2. T.C. Rindflesch and M. Fiszman. The interaction of domain knowledge and linguistic structure in natural language processing: interpreting hypernymic propositions in biomedical text. Journal of Biomedical Informatics, 36(6):462-477, 2003.
3. T.C. Rindflesch. et al. Semantic MEDLINE: An advanced information management application for biomedicine. Information Services & Use, 31, 15-21, 2011.
4. C. Chen. Hindsight, insight, and foresight: A multi-level structural variation approach to the study of a scientific field. Technology Analysis & Strategic Management, 25(6):619–640, 2013.
5. C. Chen, R. Dubin, and M. C. Kim. Emerging trends and new developments in regenerative medicine: A scientometric update (2000 – 2014). Expert Opinion on Biological Therapy, 14(9):1295–1317, 2014.
6. N. Shibata, Y. Kajikawa, and I. Sakata. Detecting potential technological fronts by comparing scientific papers and patents. Foresight, 13(5):51–60, 2011.
7. X. Gong, J. Jiang, Z. Duan, and H. Lu. A new method to measure the semantic similarity from query phenotypic abnormalities to diseases based on the human phenotype ontology. BMC Bioinformatics, 19(4):111–119, 2018.

8. H. Kilicoglu, G. Rosemblat, M. Fiszman, and T. C. Rindflesch. Constructing a semantic predication gold standard from the biomedical literature. BMC Bioinformatics, 12(1):1–17, 2011.
9. O. Nikdelfaz and S. Jalili. Disease genes prediction by hmm based pu-learning using gene expression profiles. Journal of Biomedical Informatics, 81:102–111, 2018.
10. M. Zhao, S. Zhang, W. Li, and G. Chen. Matching biomedical ontologies based on formal concept analysis. Journal of Biomedical Semantics, 9(11):1–27, 2018.
11. Yi Zhang, Alan L. Porter, Zhengyin Hu, et al. “Term clumping” for technical intelligence: A case study on dye-sensitized solar cells[J]. Technological Forecasting & Social Change, 2014, 85:26-39.
12. M. Fiszman, T. C. Rindflesch, and H. Kilicoglu. Abstraction summarization for managing the biomedical research literature. In Proceedings of the HLT-NAACL Workshop on Computational Lexical Semantics, pages 76–83, 2004.
13. Z. Y. Hu, R.-Q. Zeng, X. C. Qin, et al. A Method of Biomedical Knowledge Discovery by Literature Mining Based on SPO Predications: A Case Study of Induced Pluripotent Stem Cells[C]:Springer, Cham,2018.
14. M. E. J. Newman and M. Girvan. Finding and evaluating community structure in networks. Physical Review E, 69(2):026113, 2004.
15. V. D. Blondel, J. Guillaume, R. Lambiotte, and E. Lefebvre. Fast unfolding of communities in large networks. Journal of Statistical Mechanics: Theory and Experiment, 2008(10):P10008, 2008.
16. X. Y. Xu, Y. N. Zheng, and Z. H. Liu. Study on the method of identifying research fronts based on scientific papers and patents. Library and Information Service, 60(24):97–106, 2016.
17. M. Basseur, R.-Q. Zeng, and J.-K. Hao. Hypervolume-based multi-objective local search. Neural Computing and Applications, 21(8):1917–1929, 2012.
18. L. Wei, Z. Y. Hu, H. S. Pang, et al.. Study on knowledge discovery in biomedical literature based on spo predications: A case study of induced pluripotent stem cells. Digital Library Forum, 9:28–34, 2017.