A Method of Biomedical Knowledge Discovery by Literature Mining Based on SPO Predications: A Case Study of Induced Pluripotent Stem Cells

Zheng-Yin Hu1, Rong-Qiang Zeng1,2, Xiao-Chu Qin3, Ling Wei1, and Zhiqiang Zhang1

1 Chengdu Library and Information Center of Chinese Academy of Sciences, Chengdu 610041, Sichuan, People’s Republic of China
{huzy,zhangzq}@clas.ac.cn, weiling@mail.las.ac.cn

2 School of Mathematics, Southwest Jiaotong University, Chengdu 610031, Sichuan, People’s Republic of China
zrq@swjtu.edu.cn

3 Guangzhou Institutes of Biomedicine and Health, Chinese Academy of Sciences, Guangzhou 510530, Guangdong, People’s Republic of China
qin_xiaochu@gibh.ac.cn

Abstract. A large amount of valuable knowledge is hidden in the vast biomedical literatures, publications, and online contents. In order to identify the previously unknown biomedical knowledge from these resources, we propose a new method of knowledge discovery based on SPO predications, which constructs a three-level SPO-semantic relation network in the considered area. We carry out the experiments in the area of induced pluripotent stem cells, and the experimental results indicate that our proposed method can significantly discover the potential biomedical knowledge in this area, and the performance analysis of this method sheds lights on the ways to further improvements.

Keywords: Biomedical knowledge discovery · SPO
Induced pluripotent stem cells · Semantic relation network
Community detection

1 Introduction

Knowledge Discovery in Text (KDT) is the process of identifying and extracting the new, useful, potential and understandable patterns from the literatures in a credible way. With the rapid growth of biomedical literatures, Knowledge Discovery in Biomedical Literature (KDiBL) has become an important research area [8].

Information extraction plays an important part in KDT, which automatically extracts the specific terms, the corresponding characteristics and the semantic
relations among them from the texts as the basic knowledge unit of knowledge
discovery. Subject-Predication-Object (SPO) represents the semantic relationships among the knowledge units, which is widely used in the fields of knowledge organization, semantic network, knowledge discovery, and so on [3].

In this paper, we propose a new method of knowledge discovery based on SPO predications, which constructs a three-level SPO-based semantic relation network in the considered area. Then, we realize the community detection for the SPO-semantic relation network, so as to find the hidden valuable knowledge. The experimental results indicate that the proposed method can effectively discover the unknown biomedical knowledge. The performance analysis explains the behavior of our proposed method and sheds lights on the ways to further improvements.

The remaining part of this paper is organized as follows. In the next section, we briefly review the previous works related to the biomedical knowledge discovery. In Sect. 3, we investigate a new method of constructing three-level SPO-based semantic relation network to discover the potential unknown knowledge. Section 4 provides the experimental results and the performance analysis of the proposed method in the area of induced pluripotent stem cells. The conclusions are presented in the last section.

2 Literature Reviews

In this section, we present the literature reviews concentrating on the biomedical knowledge discovery.

In [1], with techniques from systems medicine, natural language processing, and graph theory, the authors created a molecular interaction network, which represents neural injury and is composed of relationships automatically extracted from the literature, in order to support the diagnosis of mild traumatic brain injury. Actually, they retrieved the citations related to neurological injury and extract the semantic predications that contain potential biomarkers. The experimental results on 99,437 relevant citations and 26,441 unique relations indicated a set of 17 potential biomarkers, which provides an opportunity to obtain more effective diagnosis than the current methods.

In [5], the authors investigated the use of deep learning methods, which have shown significant promise in identifying hidden patterns from large corpus of text in an unsupervised manner, in order to discover the hidden, interesting or previously unknown biomedical knowledge from free text resources. They used the text corpus from MRDEF file in the Unified Medical Language System (UMLS) dataset as training set to discover potential relationships. Taking a manual evaluation from a sample of the non-overlapping set, their proposed algorithm founded 32% of new relationships not originally represented in the UMLS, which provides provide a promising approach in discovering potential new biomedical knowledge from free text.

In [7], 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 are carried out on a mixed part of 398 disease genes from three disease types and 12001 unlabeled genes, and the results indicate a significant improvement in comparison with the other methods in literature.

In [9], the authors presented a set of knowledge discovery framework to identify the unknown knowledge from the biomedical literatures based on subject-predication-object predications. Actually, they extracted the SPO predications from the biomedical literature by using UMLS corpus and SemRep. Then, they constructed the corresponding semantic network diagrams with NetMiner [10], which is applied to the field of induced pluripotent stem cells. The experimental results showed that can effectively reveal the knowledge content from the biomedical literatures.

In [11], the authors proposed a novel Sequence-based Fusion Method (SFM) is proposed to identify disease genes from human genome, which is of great importance to improve diagnosis and treatment of disease. In this method, the amino acid sequence of the proteins has been carried out to present the genes into four different feature vectors, instead of using a noisy and incomplete prior-knowledge. Then, the intersection set of four negative sets generated by distance approach is used to select more likely negative data from candidate genes, and the decision tree has been applied as a fusion method to combine the results of four independent state-of the-art predictors based on support vector machine (SVM) algorithm for the final decision. The experimental results confirm the efficiency and validity of the proposed method.

In [12], the authors proposed a method based on degree centrality that measures connectedness in a graph, in order to automatically summarize the semantic predications representing assertions in MEDLINE citations in the large graph with more than 500 citations. The experiment was carried out on the four categories of clinical concepts related to treatment of disease, the results showed that their proposed method are very competitive, in comparison with the reference standard produced manually by two physicians.

In [13], the authors presented a hybrid model for the extraction of biomedical relations that combines Recurrent Neural Networks (RNNs) and Convolutional Neural Networks (CNNs), in order to extract high-quality biomedical relations from biomedical texts. In this model, RNNs and CNNs are employed to automatically learn the features from the sentence sequence and the dependency sequences to generate the shortest dependency path for the biomedical relation extraction. The experiments are carried out on five public (protein-protein interaction) PPI corpora and a (drug-drug interaction) DDI corpus, and the experimental results indicate the proposed model can effectively boost biomedical relation extraction performance.
3 Methodology

In our work, we propose a new method of knowledge discovery based on SPO predicates, which constructs a three-level SPO-based semantic relation network in a certain area. First, we present an introduction to the SPO-based semantic relation network. Then, we construct a three-level graph, which is different from the graph generated by the NetMiner. Afterwards, we investigate the method of detecting the community in the three-level graph.

3.1 Network Construction

Generally, we construct the SPO-based semantic relation network, according to four basic principles proposed by M. Fiszman et al., which are relevancy, connectivity, novelty and saliency, more information about these principles can be found in [2].

Two examples of directed semantic networks of induced pluripotent stem cells are respectively illustrated in Figs. 1 and 2, which are both depicted by NetMiner. In these two figures, the node in the network represents the semantic concepts, and the corresponding color represents the type of semantic concept. In addition, the edge represents the semantic relationship with the direction from the subject to the object, the color and the width of the line represent the semantic type and the frequency of semantic description respectively.
In Fig. 1, the Induced Pluripotent Stem Cells (IPSC) is the subject, the other nodes are the objects. Whereas, in Fig. 2, the Induced Pluripotent Stem Cells (IPSC) is the object, the other nodes are the subjects. Actually, these two figures present the semantic relation between the IPSC and the other nodes from two different angles.

However, we can only obtain the local semantic relation between one subject and the other objects (or between one object and the other subjects) in Both Figs. 1 and 2. In fact, it is very difficult to illustrate the global semantic relation between different subjects and different objects in one figure with NetMiner, which is the disadvantage of discovering the hidden biomedical knowledge. Then, it is essential that we construct the global semantic relation network for knowledge discovery.

An example of global SPO-based semantic relation network is illustrated in Fig. 3, which is composed of thousands of nodes and edges. In this figure, the grey node denotes the subject, the orange node denotes the object, and the green node denotes the semantic relation. Actually, many nodes can be both the subjects and the objects, which makes the whole network very complicated. That’s to say, it is very difficult for the experts to recognize the valuable biomedical knowledge from the network.
3.2 Community Detection

In order to clearly recognize the valuable biomedical knowledge from global SPO-based semantic relation network, it is essential to detect the community structure, which is the intrinsic properties of networks. In our work, we take the widely accepted modularity function proposed by Newman and Girvan, which is defined as follows [6]:

\[
Q = \frac{1}{2m} \sum_{vw} [A_{vw} - \frac{k_v k_w}{2m}] \delta(C_v, C_w),
\]

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

In addition, the modularity function can be represented in a simple way, which is formulated below [6]:

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

(2)

where $i$ runs over all communities in graph, $e_{ij}$ and $a_i^2$ are respectively defined as follows [6]:

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

(3)

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

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

(4)

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

Then, we employ the local search procedure to effectively detect the community structure, which is presented in the Algorithm 1 [4].

**Algorithm 1.** Community Detection Algorithm

1: **Input:** network adjacency matrix $A$
2: **Output:** the best value of the modularity function
3: $P = \{x^1, \ldots, x^p\} \leftarrow$ Random Initialization ($P$)
4: **repeat**
5: $x^i \leftarrow$ Local Search ($x^i$)
6: **until** a stop criterion is met

In this algorithm, we randomly divide the whole network into two communities, and each smaller community is further divided into two smaller communities. Let $C_i$ and $C_j$ be two communities, $w$ be a vertex from $C_i$ or $C_j$, we assume that $w \in C_i$ and the corresponding change by moving vertex $w$ from $C_i$ to $C_j$ can be computed as follows [4]:

$$\Delta Q(w, C_i, C_j) = \frac{k^j_w - k^i_w}{m} + \frac{k_w(a_i - a_j)}{m} - \frac{k^2_w}{2m^2},$$

(5)

where $k^i_w$ and $k^j_w$ are respectively the number of edges connecting vertex $w$ and the other vertices in communities $C_i$ and $C_j$. While, for any vertex $v$ in
community $C_i$, we can also obtain the updated $\Delta Q$ value $\Delta Q'(v, C_i, C_j)$ with the formula below [4]:

$$\Delta Q'(v, C_i, C_j) = \Delta Q(v, C_i, C_j) - \left( \frac{k_{vw}^2}{m^2} - 2\frac{A_{vw}}{m} \right).$$  \hspace{1cm} (6)

With the incremental value of the modularity function in Formulas 5 and 6, the local search procedure can choose the best move at each step until the modularity does not improve any more. Then, we obtain the communities of the considered network.

4 Case Study

Induced pluripotent stem cells technology is one of the most important emerging frontier technologies in the biomedical field, which can nurture new stem cells with similar differentiation potential as embryonic stem cells by reprogramming the mature cells [9]. Then, it is of great significance to realize the knowledge discovery in the SPO-based semantic network. In this section, we apply our proposed method in the area of IPSC and present the experimental results with performance analysis.

4.1 Data Information

In order to carry out the experiments, we obtain the data from the PubMed Database by inputting the key words “Regenerative Medicine” from 2000 to 2014. Then, we select the literatures retrieved by the Semantic Medline Database, the type of literature is “Journal Article”. The exact data information is presented in Table 1.

| SPO       |          |
|-----------|----------|
| Number of subjects | 2055     |
| Number of objects   | 1821     |
| Number of actions   | 45       |

Table 1. The information of SPO-based semantic relation network.

In this table, we have retrieved 10,687 papers and obtained 65,042 SPO-based semantic relations, which consists of 2055 subjects, 1821 objects and 45 actions. With these information, we can construct a three-level SPO-based semantic relation network.
Fig. 4. The global SPO-based semantic relation network. (Color figure online)
4.2 Experimental Results

In this subsection, we present the experimental results in the area of IPSC, which are classified into different communities according to the corresponding actions. The global SPO-based semantic relation network of IPSC is illustrated in Fig. 4 below.

In this figure, the upper level and the lower level respectively represent the subjects and the objects, which are linked by the edges with different colors. The middle level represents the actions, in which the frequency is proportional to the size of the circle.

Furthermore, different communities are represented in different colors, which are composed of the subjects, the objects and the corresponding actions. For example, there is a community colored in green with the action “LOCATION_OF” in Fig. 4.

Table 2. The communities detected in global SPO-based semantic relation network.

| Action          | Subject                                             | Object                                      |
|-----------------|-----------------------------------------------------|---------------------------------------------|
| LOCATION_OF     | Embryo, Liver, Epidermis                           | Toxic effect, Purinoceptor                  |
|                 | Body tissue, Basement membrane                      | Injury wounds, Tissue Engineering           |
|                 | Entire bony skeleton, Retina, ···                    | Cell Transformation, Neoplastic, ···         |
| PART_OF         | Mammary gland, Chorionic villi                      | Infraclass Eutheria, Equus caballus         |
|                 | Bone Marrow Cells, Serum                           | Cementoblasts, Entire tendon                |
|                 | Mesenchymal Stem Cells, ···                         | Rattus norvegicus, ···                      |
| ADMINISTERED_TO | Small Interfering, MicroRNAs                        | Cells, Patients, Mus                        |
|                 | RNA, Growth Factor, Adiponectin                     | Human embryonic stem cell                   |
|                 | High Throughput Screening, ···                       | Urothelial Cell, ···                         |

The computational results are summarized in Table 2. In this table, we do not present all the found communities in the network but to provide parts of three different communities, which are colored in green (located in the center of Fig. 4), in pink (located in the center of Fig. 4) and in red (located on the left of Fig. 4).

Moreover, the subjects and the objects of the community in pink are linked by the action “PART_OF”, which is the highest frequency among all the actions. From Fig. 4, we can clearly recognize the different communities, which is very helpful to realize the biomedical knowledge discovery.

5 Conclusions

In this paper, we have investigated a new method of constructing the three-level SPO-based semantic relation network for biomedical knowledge discovery. To achieve this goal, we have carried out the experiments in the area of induced pluripotent stem cells. The experimental results indicate that our proposed
method can significantly discover the potential unknown biomedical knowledge in the considered area.

Acknowledgments. The work in this paper was supported by the key projects of the National Social Science Foundation of China “Theory and Applications Research of Subject-Informatics for Domain Knowledge Discovery” (Grant No: 17ATQ008), 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 supported by the Fundamental Research Funds for the Central Universities (Grant No. A0920502051722-53).

References

1. Cairelli, M.J., Fiszman, M., Zhang, H., Rindflesch, T.C.: Networks of neuroinjury semantic predications to identify biomarkers for mild traumatic brain injury. J. Biomed. Semant. 6(25), 1–14 (2015)
2. Fiszman, M., Rindflesch, T.C., Kilicoglu, H.: Abstraction summarization for managing the biomedical research literature. In: Proceedings of the HLT-NAACL Workshop on Computational Lexical Semantics, pp. 76–83 (2004)
3. Keselman, A., Rosemblat, G., Kilicoglu, H.: Adapting semantic natural language processing technology to address information overload in influenza epidemic management. J. Am. Soc. Info. Sci. Technol. 61(12), 2531–2543 (1990)
4. Lü, Z.P., Huang, W.Q.: Iterated tabu search for identifying community structure in complex networks. Phys. Rev. E 80, 026130 (2009)
5. Rather, N.N., Chintan, O.P., Khan, S.A.: Using deep learning towards biomedical knowledge discovery. I. J. Math. Sci. Comput. 2, 1–10 (2017)
6. Newman, M.E.J., Girvan, M.: Finding and evaluating community structure in networks. Phys. Rev. E 69(2), 026113 (2004)
7. Nikdelfaz, O., Jalili, S.: Disease genes prediction by HMM based PU-learning using gene expression profiles. J. Biomed. Inform. 81, 102–111 (2018)
8. Swanson, D.R.: Medical literature as a potential source of new knowledge. Bull. Med. Libr. Assoc. 78(1), 29–37 (1990)
9. Wei, L., Hu, Z.Y., Pang, H.S., Qin, X.C., Guo, H.M., Fang, S.: Study on knowledge discovery in biomedical literature based on spo predications: a case study of induced pluripotent stem cells. Digit. Libr. Forum 9, 28–34 (2017)
10. Workman, T.E., Fiszman, M., Hurdle, J.F., Rindflesch, T.C.: Biomedical text summarization to support genetic database curation: using semantic medline to create a secondary database of genetic information. J. Med. Libr. Assoc. 98(4), 273–281 (2010)
11. Yousef, A., Charkari, N.M.: SFM: a novel sequence-based fusion method for disease genes identification and prioritization. J. Theor. Biol. 383, 12–19 (2015)
12. Zhang, H., Fiszman, M., Shin, D., Miller, C.M., Rosemblat, G., Rindflesch, T.C.: Degree centrality for semantic abstraction summarization of therapeutic studies. J. Biomed. Inform. 44, 830–838 (2011)
13. Zhang, Y., Lin, H., Yang, Z., Wang, J., Zhang, S., Sun, Y., Yang, L.: A hybrid model based on neural networks for biomedical relation extraction. J. Biomed. Inform. 81, 83–92 (2018)