# Technology Evolution Analysis Based on SPO using patent documents: a Case Study of Induced Pluripotent Stem Cells

Zhengyin Hu huzy@clas.ac.cn
Chengdu Library and Information Center, Chinese Academy of Sciences
Ling Wei weiling@mail.las.ac.cn
School of Information and management, Shanxi University of Finance and Economics
Xiaochu Qin qin_xiaochu@gibh.ac.cn
Guangzhou Institutes of Biomedicine and Health, Chinese Academy of Sciences
Yi Wen wenyi@clas.ac.cn
Chengdu Document and Information Center, Chinese Academy of Sciences
Chunjiang Liu liucj@clas.ac.cn
Chengdu Library and Information Center, Chinese Academy of Sciences

## Introduction

SPO predications consist of a Subject argument (noun phrase), an object argument (noun phrase), and the relation that binds them (verb phrase), which can represent science and technology (S&T) information with more details in a simple manner and have been widely applied in Knowledge Discovery in Biomedical Literature (KDiBL) (Reeve, Han & Brooks ,2007; Workman, Fiszman, Hurdle, et al., 2010; Min, Zhang, et al., 2013). The SPO predications are extracted from literature and cleaned. The technology is stated by SPO predications. Young et al. (2008) approached a method that can be used to draw technology evolution map of keywords by calculating the distributions of keywords over the documents cluster groups. This paper follows Young's research using SPO predications instead of keywords. Induced Pluripotent Stem Cells (IPSC) patent documents are selected as a case study.

## Methodology

*1. Extracting SPO Structures*

After collecting scientific literatures, some national language processing (NLP) tools are used to extract SPO predications from the text fields, such as "Title" and "Abstract" which are precise and meaningful for NLP. SemRep is a UMLS-based program that extracts SPO from sentences in biomedical text, and the subject and object arguments of each SPO are concepts from the UMLS Metathesaurus and their binding relationship (Predicate) is a relation from the UMLS Semantic Network (Rindflesch & Fiszman, 2003). For example, from the sentence "We used hemofiltration to treat a patient with digoxin overdose that was complicated by refractory hyperkalaemia", SemRep extracts four predications: "Hemofiltration-*TREATS*-*Patients*, *Digoxin overdose-ROCESS_OF-Patients*, *hyperkalemia-COMPLICATES-Digoxin overdose*, *Hemofiltration-TREATS (INFER)-Digoxin overdose*" (Rindflesch & Fiszman 2003). These SPO predications extracted by SemRep are cleaner and more formal and can be directly used as the basis of technology evolution analysis.

*2. Clustering the patent documents*

The patent documents are clustered by technological relevancies. Considering the specificity of patent documents, the patent classifications are chosen for patent documents clustering. The fuzzy similarity matrix of patent classifications-patent documents is the basis of clustering. The distribution of patent classifications in patent documents and the semantic relations between patent classifications are considered at the same time for matrix assignment. The formula 2 (Shi & Yang, 2007) is used to calculate the similarity values. Based on the similarity matrix of patent document, we use hierarchical clustering for patent documents clustering.

*3. Forming semantic network of SPO Predications*

After documents clustering, each cluster group can be represented as one node and SPO predications are counted in each node. If a SPO appears in more nodes, it will be moved to a new node with higher level and draw directed line segments between the new node and the others. Then, the semantic network of all SPO predications is constructed. The frequency of SPO predications appearing in the nodes is marked in the network.

*4. Drawing technology evolution map*

The earliest filling date and country in each node of semantic network is added. The earliest filling date is the earliest priority date or application date of patent documents in which the SPO appears. The corresponding country is the application country of the patent document. Then, a technology evolution map with horizontal axis of timeline and vertical axis of frequency can be drawn. Normally, the technologies in the upper left corner of the map appear in many different technology groups and were applied for patents in early time. They can be considered as the basic technologies. The technologies in the lower right corner of the map appear in fewer technology groups and were applied for patents lately. They can be considered as the latest technologies or emerging technologies.

**Case Study**

IPSC patents were selected as a case study, Derwent Innovations Index (DII) as data source and 1,282 patent documents are obtained from 2008 to 2017. Following the above methodology, the technology evolution map of IPSC patents is drawn. A part of the map is shown in Figure 1.

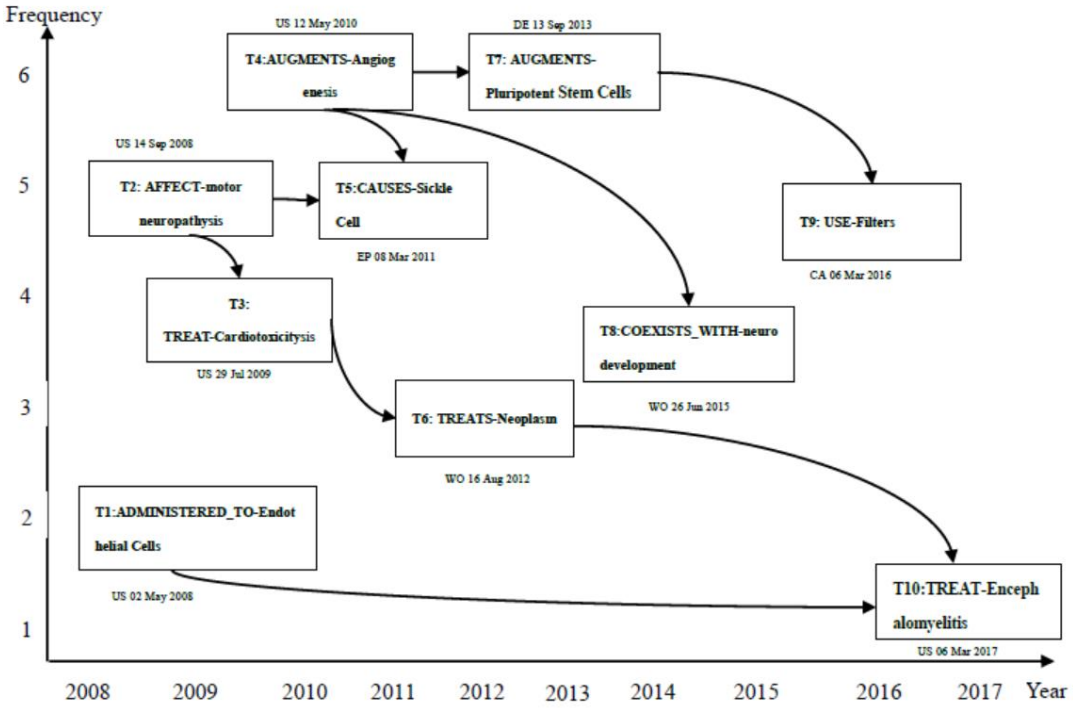


**Figure 1. A part of technology evolution map of IPSC**

## Conclusions

The result indicates that SPO predications which contain more semantic information are more suitable for technology evolution analysis than keywords. But the nodes need to be attaching understandable labels. SPO predications can be used to generate topics and draw more comprehensive technology evolution map.

## References


Min, B., Zhang, L., Zhang, Y., et al. (2017). A study of the application of publication dates on biomedical literature-based knowledge discovery. *Journal of the China Society for Scientific & Technical Information*, 36(6):574-577.

Reeve L. H., Han H., & Brooks A.D. (2007) .The use of domain-specific concepts in biomedical text summarization. *Information Processing & Management*, 43 (6):1765–1776.

Rindflesch, T.C. & Fiszman, M. (2003). 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.

Shi N. Y., Chen Y. (2007). Towards domain ontology-based semantic annotation research. *Computer Engineering and Design*, 28(24), 5985-5987.

Workman T. E., Fiszman M., Hurdle J. F., et al. (2010) .Biomedical text summarization to support genetic database curation: using Semantic MEDLINE to create a secondary database of genetic information. *J Med Libr Assoc*. 2010; 98(4):273–281.

Young Gil K., Jong Hwan S. & Sang Chan P. (2008).Visualization of patent analysis for emerging technology. *Expert Systems with Applications*, 34, 1804–1812.