Measuring Patent Similarity Based on SAO Semantic Analysis

Yun	Chen	chenyun_helloworld@foxmail.com	China	Beijing Institute of Technology
Xuefeng	Wang	wxf5122@bit.edu.cn	China	Beijing Institute of Technology
Huichao	Ren	maplerenhuichao@sina.com	China	Beijing Institute of Technology
Ying	Wang	wangying318@gmail.com	China	Beijing Institute of Technology
Zhinan	Wang	wzn6768@163.com	China	Beijing Institute of Technology

Keywords: SAO (subject-action-object) similarity; SAO semantic analysis; Similarity patent; Robot technology

Nowadays, patent as the most important pattern of intellectual properties is the way to protect achievements of technology researches. As a result, the number of patents is increasing rapidly. That makes it more difficult for examiners to find the similarity patents quickly and more challengeable for applicants to evaluate the risk of patent infringement. How to measure the similarity between patents accurately and quickly has become an advanced research hotspot. There are three main methods to measure patent similarity. The first is based on co-classification analysis like IPC codes analysis (Zhang and Shang et al., 2016). However the IPC system is a "vague" classification and cannot express the specific technology information of patents. The second is based on the citation analysis (Yoon and Park, 2004). But not all databases provide citation information. The third is keywords-based analysis which has been widely adopted to measure the similarity of patents (Yoon, 2008). Nevertheless, keywords cannot express the semantic technology information. SAO (subject-action-object) structure analysis which not only emphasizes the keywords but also expresses the semantic relevant of components in patent avoids the disadvantage of keyword-based analysis. Some researchers have suggested measuring patent similarity based on the SAO semantic analysis (Park and Yoon et al., 2012; Park and Yoon et al., 2013). But in previous study, the researchers just consider that every SAO structure is equally important for the patent (Yoon and Kim, 2012). As we know the same SAO structure may appears in almost all patents when patents are around the same technology topic. It is appropriate to distinguish SAO structures which appear in many other patents from SAO structures which appear in few patents.

This paper proposes a method to get weight of each SAO structure called DW (distinguishing weight) extracted from the patent. What's more, this paper shows a framework (Figure 1) to discover similar patents in a same topic patent dataset. Figure 1 shows the process to discover the patents similar to $Patent_t$. The specific procedure is below:

1) Extract the SAO structures from the patents;

2) Clean the SAO structures;
3) Calculate the DW of each SAO structure of $Patent_t$. The steps of the program are below:

① i = 1 (Give i an initial value of 1);
② f = 1 (Give f an initial value of 1);
③ k = 1 (Give k an initial value of 1);
④ j = 1 (Give j an initial value of 1);
⑤ Calculate the similarity S between SAO_i^P (one of the SAO structure of $Patent_t$) and SAO_j (one of the SAO structure of P_k). P_k is one patent in the data set except $Patent_t$.

$$S = \alpha[Sim(S_{(i)}, S_{(j)}) + Sim(S_{(i)}, O_{(j)}) + Sim(O_{(i)}, S_{(j)}) + Sim(O_{(i)}, O_{(j)})] + \beta Sim(A_{(i)}, A_{(j)}) \quad (1)$$

(α、β are coefficients. S is the subject of SAO structure. O is the Object of SAO structure. A is the action of SAO structure)

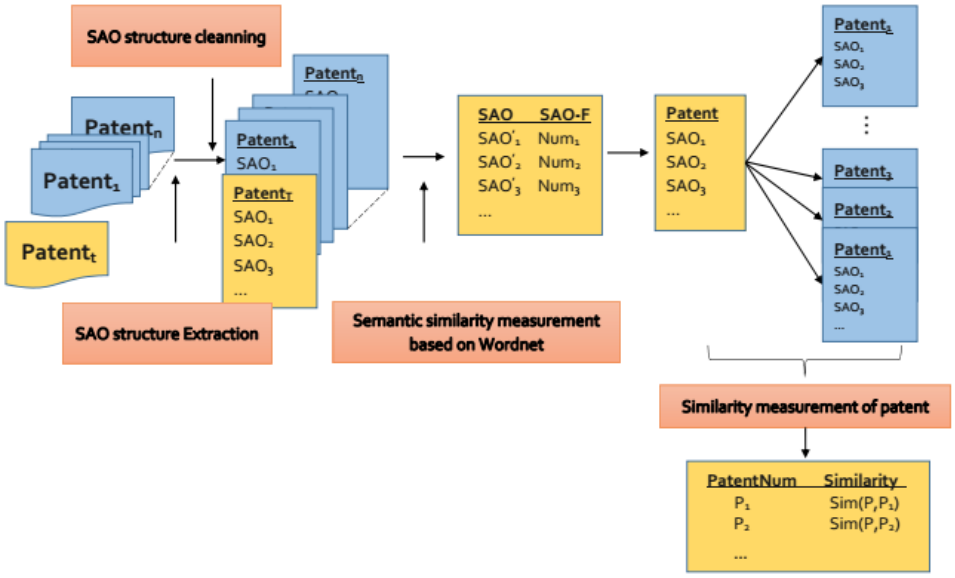

Figure 1 Process to get patent similarity between P_k and $Patent_t$

⑥ If S ≥ W (W is a threshold value), the f (f is the document frequency of the SAO_i^P) pluses 1 and the process turns to ⑧, else the process turns to ⑦;
⑦ If j < Q (Q is the number of SAO structure of P_k), j pluses 1 and the process turns to ⑤, else the process turns to ⑧;
⑧ If k < N-1 (N is the number of the patent data set include $Patent_t$), k pluses 1 and the process turns to ④, else the process turns to ⑨;
⑨ Calculate the DW of SAO_i^P;

$$DWSAO_i^P = 1 - \frac{f}{N} \quad (2)$$

⑩If i < M, i pluses 1 and the process turns to ②, else the process finishes;

4) Sim(P, P_k) is the similarity between $Patent_t$ and P_k. $NumSAO_{Pk}$ is the number of SAO structures of P_k:

$$Sim(P, P_k) = \frac{Match(P, P_k)}{M \times NumSAO_{P_k}} \quad (3)$$

$$Match(P, P_k) = \sum_{i=1}^{m} \sum_{j=1}^{q} (Sim(SAO_i^P, SAO_j^{P_k}) \times DWSAO_i^P) \quad (4)$$

This paper use a patent dataset of robot industry which is a kind of innovation and high technology industry as a case study. The case study to measure the similarities of patents about robot technology demonstrates the reliability of our method and the results indicate the practical meaning of our method to get more accurate result.

Reference:

[1]Miller, G. A. (1995). WordNet: a lexical database for English. Communications of the Acm, **38**(11): 39--41.

[2]Park, H. and J. Yoon, et al. (2012). Identifying patent infringement using SAO based semantic technological similarities. Scientometrics, **90**(2): 515-529.

[3]Park, H. and J. Yoon, et al. (2013). Identification and evaluation of corporations for merger and acquisition strategies using patent information and text mining. Scientometrics, **97**(3): 883-909.

[4]Yoon, B. (2008). On the development of a technology intelligence tool for identifying technology opportunity. Expert Systems with Applications, **35**(1–2): 124-135.

[5]Yoon, B. and Y. Park (2004). A text-mining-based patent network: Analytical tool for high-technology trend. Journal of High Technology Management Research, **15**(1): 37-50.

[6]Yoon, J. and K. Kim (2012). Detecting signals of new technological opportunities using semantic patent analysis and outlier detection. Scientometrics, **90**(2): 445-461.

[7]Zhang, Y. and L. Shang, et al. (2016). A hybrid similarity measure method for patent portfolio analysis. Journal of Informetrics, **10**(4): 1108-1130.