Measuring Patent Similarity Based on SAO Semantic Analysis

<table>
<thead>
<tr>
<th>Yun</th>
<th>Chen</th>
<th><a href="mailto:chenyun_helloworld@foxmail.com">chenyun_helloworld@foxmail.com</a></th>
<th>China</th>
<th>Beijing Institute of Technology</th>
</tr>
</thead>
<tbody>
<tr>
<td>Xuefeng</td>
<td>Wang</td>
<td><a href="mailto:wxf5122@bit.edu.cn">wxf5122@bit.edu.cn</a></td>
<td>China</td>
<td>Beijing Institute of Technology</td>
</tr>
<tr>
<td>Huichao</td>
<td>Ren</td>
<td><a href="mailto:maplerenhuichao@sina.com">maplerenhuichao@sina.com</a></td>
<td>China</td>
<td>Beijing Institute of Technology</td>
</tr>
<tr>
<td>Ying</td>
<td>Wang</td>
<td><a href="mailto:wangying318@gmail.com">wangying318@gmail.com</a></td>
<td>China</td>
<td>Beijing Institute of Technology</td>
</tr>
<tr>
<td>Zhinan</td>
<td>Wang</td>
<td><a href="mailto:wzn6768@163.com">wzn6768@163.com</a></td>
<td>China</td>
<td>Beijing Institute of Technology</td>
</tr>
</tbody>
</table>

**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,. 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. 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\textsubscript{i} (one of the SAO structure of Patent\textsubscript{t}) and SAO\textsubscript{j} (one of the SAO structure of Patent\textsubscript{k}). Patent\textsubscript{k} is one patent in the data set except Patent\textsubscript{t}.
   \[
   S = \alpha \left[ \text{Sim}(S_{i\alpha}, S_{j\alpha}) + \text{Sim}(S_{i\beta}, O_{j\beta}) + \text{Sim}(O_{i\gamma}, S_{j\gamma}) \right] + \beta \text{Sim}(A_{i\delta}, A_{j\delta})
   \] (\(\alpha, \beta\) are coefficients. S is the subject of SAO structure. O is the Object of SAO structure. A is the action of SAO structure)

   ⑥ If S \(\geq\) W (W is a threshold value), the f (f is the document frequency of the SAO\textsubscript{i}) plus 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 Patent\textsubscript{k}), j plus 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\textsubscript{t}), k plus 1 and the process turns to ④, else the process turns to ⑨;
   ⑨ Calculate the DW of SAO\textsubscript{i}:
   \[
   DWSAO\textsubscript{i}^f = 1 - \frac{f}{N}
   \] (2)
   ⑩ If i \(<\) M, i plus 1 and the process turns to ②, else the process finishes;
   4) Sim(P, Patent\textsubscript{k}) is the similarity between Patent\textsubscript{t} and Patent\textsubscript{k}. NumSAO\textsubscript{Patent\textsubscript{k}} is the number of SAO structures of Patent\textsubscript{k}:

\textbf{Figure 1} Process to get patent similarity between Patent\textsubscript{k} and Patent\textsubscript{t}.
This paper uses a patent dataset of the 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 results.

Reference: