Scientometrics (2019) 121:1–23
https://doi.org/10.1007/s11192-019-03191-z

Measuring patent similarity with SAO semantic analysis

Xuefeng Wang[1] · Huichao Ren[1] · Yun Chen[1] · Yuqin Liu[2] · Yali Qiao[1] · Ying Huang[1]

Received: 26 April 2018 / Published online: 20 July 2019

Abstract
Patents are not only an important aspect of intellectual property rights, but they are also one of the only ways to protect technological inventions. However, in recent years, the number of patents has been increasing dramatically and, as a result, both patent applicants and patent examiners are finding it more difficult to conduct the due diligence step of the patent registration process. Therefore, the lack of a quick and easy way to accurately measure patent similarity has become a significant obstacle to protecting intellectual property. Currently, there are three main ways to measure patent similarity: IPC code analysis, citation analysis, and keyword analysis. None of these approaches are able to fully reflect the semantics in a patent's content. As an emerging methodology, subject–action–object (SAO) semantic analysis does reflect semantics, but most approaches treat each identified relationship as equally important, which does not necessarily provide an accurate measure of patent similarity. To offer this power to SAO analysis, this article introduces a new indicator called DWSAO as a reflection of the weight of each SAO semantic structure. Further, we present a semantic analysis framework that incorporates the DWSAO index for finding similar patents based on the weight of each SAO structure in the patent. A case study on the similarity of patents in the field of robotics was used to verify the reliability of the method. The results highlight the detailed meanings derived from the method, the accuracy of the outcomes, and the practical significance of using this approach.

Keywords Patent similarity measurement · Text mining · Subject–action–object (SAO) · Semantic analysis · Robot docking stations

Introduction

In the face of economic globalization and fierce competition, technological innovation has become a decisive factor in the success of an enterprise. And, for nearly 500 years, patents have been one of the most important and effective ways to protect technological achievements. At present, almost all countries, regions, and international organizations have patent

Xuefeng Wang
wxf5122@bit.edu.cn

[1] School of Management and Economics, Beijing Institute of Technology, Beijing 100081, China

[2] School of Journalism and Publication, Beijing Institute of Graphic Communication, Beijing 102600, China

offices—for example, the State Intellectual Property Office of China (SIPO), the United States Patent and Trademark Office (USPTO), the Japan Patent Office (JPO), and so on.

To ultimately succeed in being issued a patent, technological innovation usually involves three stages: the invention process, the patent application process, and the patent examination process. During the patent application phase, applicants are usually required to ascertain whether any similar patents exist to avoid the risk of infringing on any other patent holder's rights, i.e., they must perform proper due diligence. Similarly, in the process of considering whether a patent should be granted, patent examiners need to search for similar patents to evaluate whether the invention meets the criteria of novelty and innovation (Adams 2006).

However, with the rapid increase in the number of patents in recent years, proper due diligence has become much more difficult for both applicants and examiners. Finding similar patents is harder, as is assessing an invention's novelty and innovation because the three main methods currently used to measure patent similarity each have shortcomings. A new more accurate method is needed.

Of these three methods, co-classification analysis assesses the similarity between patents through patent classification codes—IPC codes being one of the most common. Yet IPC classifications tend not to represent specific technologies and, therefore, do not necessarily summarize the full content of a patent. As a result, these methods are prone to inaccuracy (Zhang et al. 2016). Citation analysis relies on the references in patents to map the relationships between technologies and draw conclusions about their similarity (Yoon and Park 2004). However, not all patent databases provide citation information. Keyword-based analysis is probably the most widely used method for measuring similarity (Yoon 2008), but keywords do not reflect the relationships between concepts. Sternitzke and Bergmann (2009) compared various methodologies of similarity measures such as co-word analysis, Subject–Action–Object (SAO) structures, bibliographic coupling, co-citation analysis and self-citation links, and found that the two former ones tend to describe rather semantic similarities that differ from knowledge flows as expressed by the citation-based methodologies.

As an extension to keyword analysis, SAO semantic analysis not only emphasizes keywords but also captures the semantic functions between keywords to overcome the disadvantages of this more basic approach. However, previous studies on SAO semantic analysis have been based on the assumption that every SAO structure in a patent is equally important (Park et al. 2012, 2013a, b; Yoon 2012).

It is well known that the frequency with which different SAO structures appear in a given technical domain varies widely. It is natural then to consider whether distinguishing between the SAO structures that only appear in a few patents versus those that appear in many patents in a given domain could more accurately identify similar patents. To test this theory, we developed a new indicator, called DWSAO, that assigns a weight to each SAO structure in a patent, placing emphasis on the most important semantic structures for assessing similarity.

Traditional SAO analysis methods transform patent documents into subject–verb–object structures, i.e., SAO structures, which are canonical expressions of meaning (Park et al. 2012). Once collected, these SAO structures represent the technological content of a patent. The similarity between two sets of SAO structures can then be used to assess the similarity between two patents. Our procedure also begins by extracting SAO structures, but the DWSAO indicator is used in an interim step to measure the importance of each structure before evaluating the similarity between two patents. For brevity, we have only conducted one analysis from the perspective of comparing a 'target' patent to a set of relevant patents. However, it is important to note that this approach could be used to assess the similarities between any two patents in a set. The specific steps follow: (1) extract SAO structures from the patent using natural language processing and preprocess them; (2) measure the technological similarity between all the extracted SAO structures using traditional semantic analysis techniques; (3) assign a weight to each SAO structure using the DWSAO indicator; (4) assess the similarity between the target patent and a set of related patents.

The rest of this article is organized as follows. The relevant work section reviews previous studies, including SAO semantic analysis and text similarity measurement. In the Methodology section, we propose a detailed framework for discovering similar patents based on SAO semantic analysis and measuring the weight of each SAO structure. The feasibility and effectiveness of the method are then assessed through an empirical case study of 220 robot docking station patents in the DWPI database. Finally, we conclude the paper with a summary of this research, its limitations, and our potential directions for future work.

Related works

SAO semantic analysis

SAO structures are syntactic structures that express the semantic relationship between things, i.e., how the subject (*S*) of a sentence relates to the object (*O*) of a sentence through an action (*A*). For example, in the sentence "Electricity creates light." "Electricity" is the subject, "creates" is the action, and "light" is the object. When combined, a subject, an action, and an object can convey a complete picture of how two things are related to or affect each other. Similar to SAO structure, SPO (Subject–Predication–Object) which consists of a subject argument, an object argument, and the relation that binds them can be considered as a kind of semantic network and are widely used in Knowledge Discovery in Biomedical Literature (KDiBL) (Ahlers et al. 2007; Keselman et al. 2010), and SAO usually is used in text mining in patent documents.

Scholars have discovered that subjects can represent "solutions", actions can represent either the "effect" or the "influence" of the solution, and objects can represent the "invention problem" (Verbitsky 2004; Moehrle 2005; Kim et al. 2009; Zhang et al. 2014). The use of SAO structures to characterize the technological content of patents has significant advantages over traditional patent features (Angeli et al. 2015). For instance, analyzing a collection of SAO structures, as opposed to a plain reading of the patent, can lead to a better and more concise representation of the patent's content. Further, comparisons between patents can be transformed into comparisons between sets of SAO structures to better represent their similarities and differences, as shown in Fig. 1.

Hence, SAO analysis and SAO comparisons have become relatively common research methodologies. Scholars have used SAO analysis to: identify opportunities in technology (Wang et al. 2017); explore trends in technology competition (Wang et al. 2015; Yoon et al. 2013); and identify patent infringements (Park and Yoon 2014). In terms of SAO comparisons, Sternitzke and Bergmann (2009) used SAO structures to measure patent similarity by combining SAO structures with the Inclusion index, the Jaccard index, and the Cosine index, while Yufeng et al. (2016) combined SAO structures with a VSM model. Methods of calculating patent similarity based on SAO structures have also been used to identify patent infringements (Bergmann et al. 2008; Park et al. 2012); to identify and evaluate

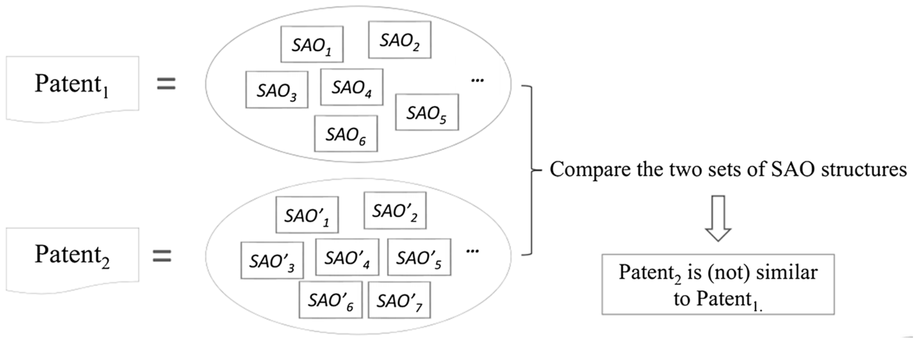

Fig. 1 Schematic diagram of a patent comparison based on SAO structures

corporations for merger and acquisition strategies (Park et al. 2013b); and, again, to identify new opportunities in technology (Yoon 2012).

Manually extracting SAO structures from patents is the most accurate way to assemble the structures for analysis, but it is also the most inefficient and is practically infeasible with a large volume of patents. However, advancements in natural language processing techniques now make it possible to extract SAO structures using text mining tools, such as GoldFire (former name Knowledgist2.5™), Stanford OpenIE, and OLLIE.

Text similarity measurement

Text similarity measurement assesses the extent to which the information in two texts is either the same or semantically the same. The indicators used are typically in line with the general idea of co-word analysis, in which patents are seen as similar if they share a high number of common textual elements (Moehrle 2010). The main measurement methods include: string-based methods (e.g., LCS, Jaccard similarity, overlap coefficient, weighted word overlap, and sentence vector) (Braam et al. 1988; Saric et al. 2012); corpus-based methods (distributional meanings of words and latent semantic analysis) (Boyack et al. 2005; Magerman et al. 2010); and syntactic-based methods (Manning et al. 2014). Traditional text similarity measures simply use the frequency of raw terms to calculate the similarity between records. However, in recent years, researchers have paid more attention to measuring semantic similarity as a result of the advancements in natural language processing techniques (Bär et al. 2012; Zarrella et al. 2015).

From the perspective of information theory, Lin (1998) asserts that the greater the commonality of two texts, the higher the degree of similarity. After comparing and analyzing previous methods for measuring the similarity between concepts, Lin developed a method that is more aligned with natural language laws.

In addition to similarity calculation methods, there are some commonly-used weight indicators. Term frequency–inverse document frequency (TF–IDF) is a weighting technique commonly used for searching similar text. The TF–IDF value of a word is derived by multiplying the frequency of the word in a given document with the inverse value of the word's frequency across a set of documents. The main idea behind this weighting is that a word may frequently appear in one article but rarely appear in other articles. If so, the word has a strong ability to distinguish between topics. Most researchers use TF–IDF to filter out common words, while preserving the more important and meaningful ones. This weight has been widely used in information retrieval and data mining analysis.

Patents tend to contain a specific and uniform set of text elements to describe technological innovations—for example, a title, an abstract, a detailed description, claims, and so on. The abstract summarizes the invention. A detailed description follows, which is the longest part and comprises: background information (prior art), a summary of the invention, precise details about the invention (including experimental details, drawings, and tables), and why it is claimed to be superior over the prior art. Next are the claims, which is the most important part of the document. Here, the goal is to explicitly and distinctly highlight the subject matter regarded as the invention(s). However, due to the necessity of protecting one's inventions, patents are often written with complex sentences, synonyms, and rare words to prevent easy retrieval by competitors. While this creates difficulties for other analysis techniques, it is good news for measuring patent similarity based on concepts and semantics rather than specific terms.

In addition, even though each patent contains a unique technological innovation, different patents in the same field are likely to contain some of the same technical information. Therefore, considering the distinctions between the common technical information and unique technical information, by assigning each with a different weight, can improve the accuracy of similarity measurements.

Methodology

This paper focuses on identifying patents that are similar to a target patent from a set of related patents using SAO semantic analysis. Previous studies on SAO semantic analysis have assumed that every SAO structure in a patent is of equal importance. The DWSAO indicator introduced in this paper, has been designed to weight the relative importance of each SAO structure as a more accurate measure of the similarity between patents. The overall procedure for accomplishing this goal is shown in Fig. 2.

1. First, the SAO structures are extracted from both the target patent, denoted as Patent$_T$, and a set of related patents, denoted as Patent$_i$ where $1 \leq i \leq n$ and n is the count of related patents. Then, the SAO structures are cleaned using standard data pre-processing

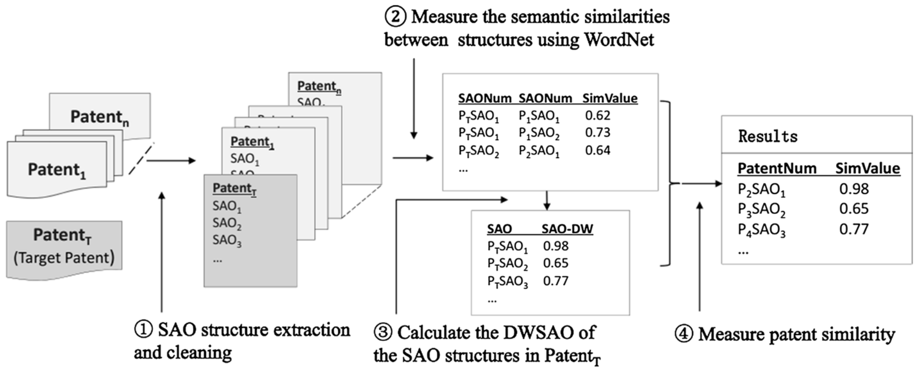

Fig. 2 The overall procedure for identifying similar patents

techniques. Once this process is complete, each patent will correspond to a collection of SAO structures. We used the enhanced abstracts provided by DWPI as our source material. These abstracts comprise up to seven separate sections (fields)—novelty, detailed description, use, advantage, activity, mechanism-of-action, and description of drawing(s) to provide a concise yet detailed summary of the claimed invention. The SAO structures were cleaned using the procedure outlined in the Extraction and Cleaning section below.

2. Second, an initial measure of semantic similarity needs to be calculated to determine the similarity between the SAO structures of $Patent_T$ and the structures in related patents.
3. Third, a DWSAO value is calculated for each SAO structure in $Patent_T$. The DWSAO indicator quantifies how important a SAO structure is as a representative of the technical features in that patent. We chose WordNet as the base source for the semantic relations, and designed our own algorithm for analyzing the similarities between the SAO structures based on the DWSAO weightings.
4. The last step is to calculate the similarity between $Patent_T$ and each related patent.

SAO structure extraction and cleaning

To extract the collection of SAO structures from the DWPI enhanced abstract, we developed a bespoke program based on Stanford Parse. The main steps are listed as below, and Fig. 3 provides more details in a schematic form.

1. Split the abstract into separate sentences.
2. Analyze the sentences syntactically using Stanford Parser.
3. Handle complex sentences and simple sentences separately, and re-analyze the complex subjects and objects. Extract the backbone of every sentence.
4. Extract the SAO structures.

Natural language processing technology continues to develop and improve, but it still has some limitations. The SAO structures extracted from the abstracts contain some noisy data. To generate a set of more accurate and effective SAO structures, the extracted structures must be cleaned. Five cleaning rules were applied, as shown in Table 1.

Calculating the semantic similarities between SAO structures

As previously mentioned, each patent is represented as a collection of SAO structures. Each SAO structure is composed of a subject (S), an action (A), and an object (O). Subjects and objects are nouns; actions are verbs. Additionally, each component may comprise more than one word, i.e., noun or verb phrases (Fig. 4).

Since calculating the similarity between patents must measure the similarities between corresponding elements and also all pairs of elements, the method needs to calculate the similarity between words before calculating the similarity between patents.

We chose WordNet as the source of word relationships to calculate the semantic similarity between terms. WordNet is a lexical database for English created by Princeton University (Miller 1995). It contains nouns, verbs, adjectives, and adverbs. The four kinds of terms are grouped into sets of cognitive synonyms, called synsets. Each synset represents a distinct concept and also labels the relations among words by interlinking conceptual and lexical semantic relationships. As a result, WordNet provides an effective combination of

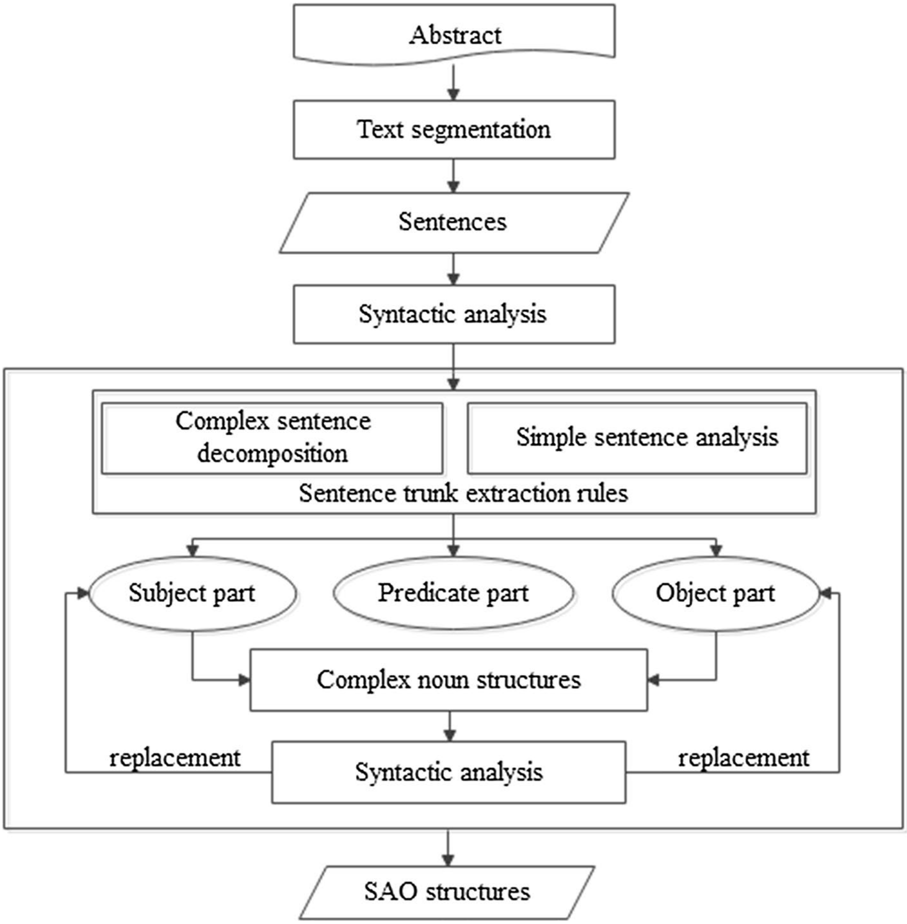

Fig. 3 The process for extracting SAO structures from the abstracts of patents

Table 1 Rules for cleaning the SAO structures

No.	Steps
1	Split the *S* and *O* components of long SAO structures
2	Remove meaningless SAO structures
3	Remove stop words
4	Convert abbreviations into long-form phrases
5	Eliminate or reduce extraneous parts of speech

traditional lexicographic information and modern computing. JWI (the MIT Java WordNet Interface) was chosen as the interface for WordNet (Finlayson 2014). The measure of similarity between two terms is defined as follows (Lin 1998):

$$\mathrm{Sim}\left(\mathrm{Term}_{(i)}, \mathrm{Term}_{(j)}\right) = \frac{2 * \mathrm{IC(Lcs)}}{\mathrm{IC}\left(\mathrm{Term}_{(i)}\right) + \mathrm{IC}\left(\mathrm{Term}_{(j)}\right)} \tag{1}$$

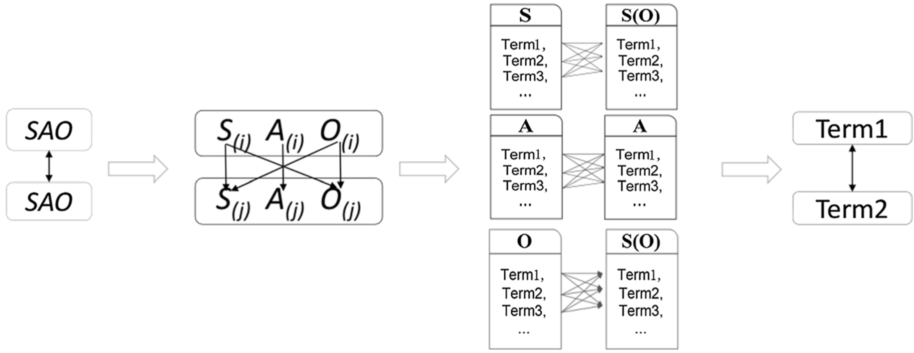

Fig. 4 A exploded view of the SAO structure comparison

where $\mathrm{Sim}\left(\mathrm{Term}_{(i)}, \mathrm{Term}_{(j)}\right)$ is the similarity between $\mathrm{Term}_{(i)}$ and $\mathrm{Term}_{(j)}$, IC(Lcs) is the least common sub-concepts of $\mathrm{Term}_{(i)}$ and $\mathrm{Term}_{(j)}$,and $\mathrm{IC}\left(\mathrm{Term}_{(i)}\right)$ and $\mathrm{IC}\left(\mathrm{Term}_{(j)}\right)$ represents the number of semantic concepts included in $\mathrm{Term}_{(i)}$ and $\mathrm{Term}_{(j)}$, respectively. The similarity between two concepts is measured as $0 \le \mathrm{Sim}\left(\mathrm{Term}_{(i)}, \mathrm{Term}_{(j)}\right) \le 1$. If the similarity of two terms $\mathrm{Sim}\left(\mathrm{Term}_{(i)}, \mathrm{Term}_{(j)}\right)$ is greater than or equal to the threshold R, the two terms are considered to match.

A measure for the semantic similarity between two subjects, two actions, or two objects can be formulated by exploiting their matching average (Park et al. 2013a):

$$\mathrm{Sim}\left(N_{(i)}, N_{(j)}\right) = \frac{2 * \mathrm{Match}\left(N_{(i)}, N_{(j)}\right)}{\mathrm{NumTerm}(N_{(i)}) + \mathrm{NumTerm}(N_{(j)})} \tag{2}$$

where N expresses subject, action, or object which is one component of the SAO, $N_{(i)}$ and $N_{(j)}$ are two components that have the same attributes of two different SAO structures, $\mathrm{Sim}\left(N_{(i)}, N_{(j)}\right)$ is the similarity between $N_{(i)}$ and $N_{(j)}$, $\mathrm{NumTerm}(N_{(i)})$, $\mathrm{NumTerm}(N_{(j)})$ is the number of terms for $N_{(i)}$ and $N_{(j)}$, and $\mathrm{Match}\left(N_{(i)}, N_{(j)}\right)$ is the sum of the number of matching terms between $N_{(i)}$ and $N_{(j)}$.

This measure of similarity between two SAO structures is defined as follows:

$$\begin{aligned}&\mathrm{Sim}\left(\mathrm{SAO}_i, \mathrm{SAO}_j\right)\\ &= \begin{cases} \alpha * \frac{\left[\mathrm{Sim}\left(S_{(i)}, S_{(j)}\right) + \mathrm{Sim}\left(O_{(i)}, O_{(j)}\right)\right]}{2} + \beta \mathrm{Sim}\left(A_{(i)}, A_{(j)}\right), \mathrm{Sim}\left(S_{(i)}, S_{(j)}\right) + \mathrm{Sim}\left(O_{(i)}, O_{(j)}\right) \ge \mathrm{Sim}\left(S_{(i)}, O_{(j)}\right) + \mathrm{Sim}\left(O_{(i)}, S_{(j)}\right) \\ \alpha * \frac{\left[\mathrm{Sim}\left(S_{(i)}, O_{(j)}\right) + \mathrm{Sim}\left(O_{(i)}, S_{(j)}\right)\right]}{2} + \beta \mathrm{Sim}\left(A_{(i)}, A_{(j)}\right), \mathrm{Sim}\left(S_{(i)}, S_{(j)}\right) + \mathrm{Sim}\left(O_{(i)}, O_{(j)}\right) < \mathrm{Sim}\left(S_{(i)}, O_{(j)}\right) + \mathrm{Sim}\left(O_{(i)}, S_{(j)}\right) \end{cases}\end{aligned} \tag{3}$$

where $\mathrm{Sim}\left(\mathrm{SAO}_i, \mathrm{SAO}_j\right)$ is the similarity between SAO_i and SAO_j, α and β are coefficients, and $0 < \alpha < 1$ and $0 < \beta < 1$ and $1 - \alpha = \beta$.

To facilitate subsequent calculations, the similarities between SAO structures are then standardized as follows:

$$\mathrm{stSim}\left(\mathrm{SAO}_i, \mathrm{SAO}_j\right) = \frac{\mathrm{Sim}\left(\mathrm{SAO}_i, \mathrm{SAO}_j\right) - \mathrm{Min}_{x,y \in \{1,n\}}\left(\mathrm{Sim}\left(\mathrm{SAO}_x, \mathrm{SAO}_y\right)\right)}{\mathrm{Max}_{x,y \in \{1,n\}}\left(\mathrm{Sim}\left(\mathrm{SAO}_x, \mathrm{SAO}_y\right)\right) - \mathrm{Min}_{x,y \in \{1,n\}}\left(\mathrm{Sim}\left(\mathrm{SAO}_x, \mathrm{SAO}_y\right)\right)} \tag{4}$$

where $\mathrm{stSim}\left(\mathrm{SAO}_i, \mathrm{SAO}_j\right)$ represents the standardized value of the semantic similarity between SAO_i and SAO_j ranging from 0 to 1, $\mathrm{Sim}\left(\mathrm{SAO}_i, \mathrm{SAO}_j\right)$ represents the semantic similarity between SAO_i and SAO_j, n represents the number of SAO semantic structures, $\mathrm{Max}_{x,y\in\{1,n\}}\left(\mathrm{Sim}\left(\mathrm{SAO}_x, \mathrm{SAO}_y\right)\right)$ represents the maximum SAO semantic similarity, and $\mathrm{Min}_{x,y\in\{1,n\}}\left(\mathrm{Sim}\left(\mathrm{SAO}_x, \mathrm{SAO}_y\right)\right)$ represents the minimum SAO semantic structure similarity.

Calculating the DWSAO for each SAO structure

At present, the most common method of measuring the semantic similarity between two pieces of text is to count the number of common words that appear in both, then represent the similarity as a proportion of the total number of shared words as follows:

$$\mathrm{Sim}\left(T_1, T_2\right) = \frac{2 * \mathrm{Match}\left(T_1, T_2\right)}{\mathrm{Num}(T_1) + \mathrm{Num}(T_2)} \tag{5}$$

where T_1 and T_2 are the two texts, $\mathrm{Sim}\left(T_1, T_2\right)$ indicates the similarity between T_1 and T_2, $\mathrm{Match}\left(T_1, T_2\right)$ indicates the number of the matching words in T_1 and T_2, and $\mathrm{Num}(T_1)$ and $\mathrm{Num}(T_2)$ represent the number of words in T_1 and T_2, respectively. A larger $\mathrm{Sim}\left(T_1, T_2\right)$ means T_1 and T_2 are more similar. Based on the above method for calculating the similarity between texts, some scholars have designed a method for measuring patent similarity using the semantic meanings of SAO structures. Angeli et al. (2015) used:

$$\mathrm{Sim}\left(P_{\mathrm{T}}, P_i\right) = \frac{2 * \mathrm{Match}\left(P_{\mathrm{T}}, P_i\right)}{\mathrm{NumSAO}\left(P_{\mathrm{T}}\right) + \mathrm{NumSAO}\left(P_i\right)} \tag{6}$$

where the $\mathrm{Sim}\left(P_{\mathrm{T}}, P_i\right)$ indicates the similarity between P_{T} and P_i the related patent, $\mathrm{Match}\left(P_{\mathrm{T}}, P_i\right)$ indicates the number of SAO structures that appear in both P_{T} and P_i, $\mathrm{NumSAO}\left(P_{\mathrm{T}}\right)$, and $\mathrm{NumSAO}\left(P_i\right)$ is the number of SAO structures that correspond to P_{T} and P_i. It is worth highlighting that, when this formula is used to measure patent similarity, each SAO structure in a patent holds the same importance by default.

However, in general, similar patents in the same technical field will contain some common technical information. Thus, many patents contain common terms and information, which does not particularly characterize the salient technical features of the patents. Similarly, patents from the same domain in the same technology category are also likely to contain the same or similar SAO structures. Therefore, each SAO structure in the patent represents the features of the technology to different degrees.

For example, given $\mathrm{Patent}_{\mathrm{T}}$, assume that the most similar patent in the set of related patents is Patent_5 (Fig. 5). Each patent in the figure has a corresponding SAO structure set, but some have some individual SAO structures in common, as represented by the numbered boxes. Past methods of semantic SAO similarity measurement (e.g., Eq. 6) will calculate the similarity between the target patent and every SAO structure in each related patent set. According to this method of measurement, as shown in Table 2, $\mathrm{Patent}_{\mathrm{T}}$ and Patent_5 have the highest similarity, and $\mathrm{Patent}_{\mathrm{T}}$ and Patent_4 have the lowest similarity.

However, a deeper analysis of these results reveals some important observations. SAO Structure 1 appears in five patents, while SAO Structures 2, 3, and 4 appear in all patents. Thus, it is reasonable to conclude that the technical information in SAO

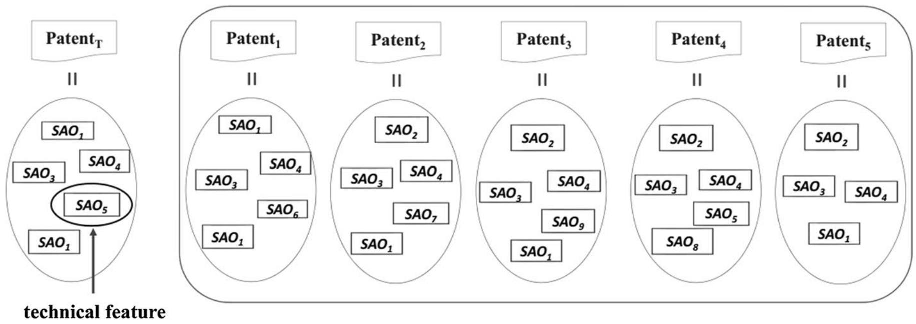

Fig. 5 The sets of SAO structures for representing target patents and related patents

Table 2 The similarity between PatentT and related patents using Eq. (6)

Formula	Patent$_1$	Patent$_2$	Patent$_3$	Patent$_4$	Patent$_5$
$\text{Match}(P_T, P_i)$	4	4	4	3	4
$\text{NumSAO}(P_T) + \text{NumSAO}(P_i)$	10	10	10	10	9
$\text{Sim}(P_T, P_i)$	0.800	0.800	0.800	0.600	0.889

Structures 1–4 are relatively common features of this patent collection, and probably represent basic or commonly-used technologies in the field. It is also highly likely that these four SAO structures do not represent innovative technologies in the target Patent$_T$ and, hence, are not its most representative characteristics for our purposes. However, SAO Structure 5 only appears in Patent$_T$ and Patent$_4$, and may represent a technical feature that is unlike most of the related patents. According to this analysis, the patent most similar to Patent$_T$ should be Patent$_4$, not Patent$_5$.

The above example demonstrates that, beyond a simple count of similar SAO structures, finding similar patents also needs to consider the *dis*similarities between the SAO structures. However, this type of manual analysis would quickly become tedious with a large number of patents. Therefore, this paper presents a novel indicator for assessing the weight of each SAO structure, i.e., DWSAO.

Assume that the number of patents in the relevant patent set is N, the number of SAO structures in the target patent P is m, and that the similarities between the SAO structures in the target patent and the related patents are known.

The relevant patents are numbered from 1 to N, and $P_k (1 \le k \le N)$ represents a related patent. The SAO structures in P are also numbered from 1 to m. SAO_i^P $(1 \le i \le m)$ denotes the SAO structure i in P. F denotes the document frequency of SAO_i^P, and DWSAO_i^P denotes the feature weight of SAO_i^P. DWSAO_i^P is specifically calculated as follows:

$$\text{DWSAO}_i^P = 1 - \frac{F}{N+1} \tag{7}$$

The specific algorithm implementing this procedure is as follows (Fig. 6):

1. F = 1 (Give F an initial value of 1);
2. $k = 1$ (k is the subscript of the related patent; give k an initial value of 1);

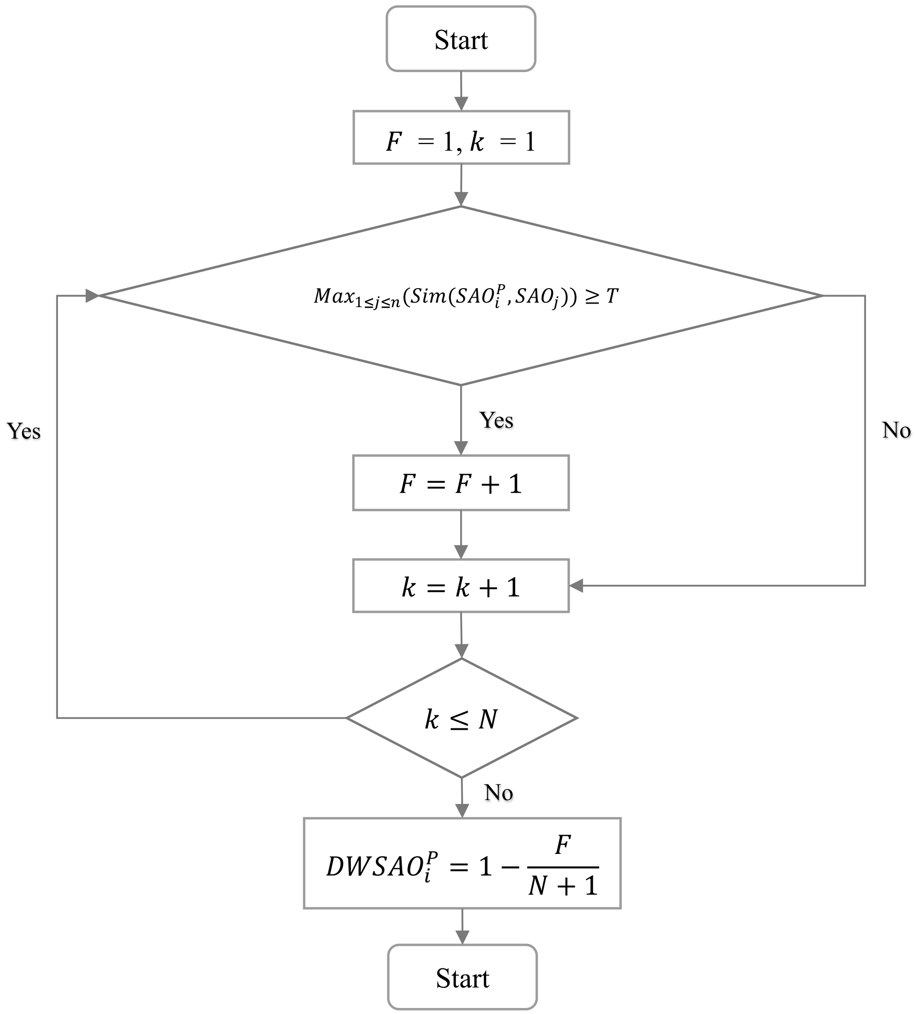

Fig. 6 Algorithm flowchart of calculating the DWSAO

3. If $k \le N$, (N is the number of the related patents) and proceed to Step (4); otherwise, proceed to Step (5);
4. If P_k contains SAO_j (j is the SAO structure number in Patent_k) and $\text{Sim}(\text{SAO}_i^P, \text{SAO}_j)$ is greater than or equal to the threshold Q and $0 < T < 1$), add 1 to F and add 1 to k and proceed to Step (3); otherwise, add 1 to k and proceed to Step (3);
5. Calculate the DWSAO of SAO_i^P by the formula (7);

In general, the larger the DWSAO_i^P, the stronger the SAO structure's ability to characterize the technology information in P. The smaller the DWSAO_i^P, the more common the SAO structure is in relation to the other patents, and the weaker its ability to represent the technical features of P.

Calculating patent similarity

The optimization method for measuring patent similarity is shown as follows:

$$\mathrm{Sim}\left(P, P_k\right) = \frac{2 * \sum_{i=1}^{m} \mathrm{DWSAO}_i^P * \mathrm{MatchSAO}_i^P}{\mathrm{NumSAO}(P) + \mathrm{NumSAO}\left(P_k\right)} \tag{8}$$

where $\mathrm{Sim}\left(P, P_k\right)$ indicates the similarity between the target patent P and P_k, ranging from 0 to 1, m represents the number of SAO structures in P, i denotes the number of SAO semantic structures in P, and DWSAO_i^P represents the DWSAO value for SAO_i^P in P. $\mathrm{MatchSAO}_i^P$ is a Boolean value when P_k contains a SAO structure that can be matched with SAO_i^P, and 0 otherwise. $\mathrm{NumSAO}(P)$ and $\mathrm{NumSAO}\left(P_k\right)$ represent the number of SAO structures contained in P and P_k, respectively.

By introducing DWSAO into a semantic similarity measurement method for patents, common SAO structures can be filtered out, which helps to better identify patents with similar target technical features.

Case study

Data collection and preprocessing

Robotics is a research hotspot. Among the many other uses, robots can improve production efficiency, enhance military strength and national defense, improve quality of life, and stimulate economic development. To illustrate the method presented in this paper, we downloaded 220 patents related to robot docking station technology published between 1 Jan 1997 and 20 July 2017 from the Derwent Innovation patent database. A selection of these patents appears in Table 3.

Generally, due diligence occurs prior to lodging a patent application and during patent examination. However, given we are not in the process of preparing to lodge an application, we chose a very recently-published patent as the target—FR3046259A. Detailed information about this patent is shown in Table 4; however, its core technical innovation is two sets of infrared light emitting diodes (LEDs) that are placed in the robot docking station area. One set of LEDs emits a ray that guides the robot's approach to the docking station in the

Table 3 20 of the 220 patents in the data sample (see "Appendix" for all patents)

No.	Patent number	No.	Patent number
1	FR3046259A1	11	US20170072568A1
2	US20170159199A1	12	US20170057760A1
3	US9672184B1	13	US20170050311A1
4	WO2017091066A1	14	CN106444736A
5	US20170105592A1	15	US20170037648A1
6	CN206115269U	16	US20170020064A1
7	US20170102709A1	17	US9527217B1
8	CN106551659A	18	US20160363933A1
9	US20170086325A1	19	WO2016196622A1
10	US20170075962A1	20	US20160349756A1

Table 4 Target patent information

Patent number	FR3046259A1
Inventor	CAUSSY Ramesh; DELARBOULAS Pierre Jean-Luc Sylvain; HASSON Cyril; ROLLAND DE RENGERVE Antoine Marie Anne
Title-DWPI	Docking station for mobile robot, has set of infrared LEDs arranged to emit attracting rays in robot approach region, and another set of infrared LEDs arranged to emit repelling rays outside robot approach region
Abstract—DWPI	The docking station (10) has a set of infrared LEDs (21–23) arranged around a robot parking zone, so as to emit attracting rays ($R1$–$R3$) in a robot approach region. Another set of infrared LEDs (24, 25) are arranged on each side of the robot parking zone, so as to emit repelling rays ($R4$, $R5$) outside the robot approach region, where the repelling rays have a shorter range than the attracting rays. The former set of LEDs is arranged such that the attracting rays are emitted in directions ($X1$–$X3$) intersecting at a fixed point (P) of the robot parking area. Docking station for a mobile robot. The sets of infrared LEDs emit attracting rays and the repelling rays, respectively, thus ensuring a mobile robot to approach the docking station according to appropriate directions defined by the attracting rays while avoiding approach to the docking station in improper directions defined by the repelling rays. The drawing shows a schematic top view of a docking station showing attracting rays and repelling rays. P Fixed point $R1$–$R3$ Attracting rays $R4$, $R5$ Repelling rays $X1$–$X3$ Directions 10 Docking station 21–25 Infrared LEDs
IPC Subclass	G05D
Publication date	2017/6/30

correct direction and within a limited area. The other set of LEDs emits a ray that repels robots away from inappropriate or incorrect approaches. A summary of technical information in the target patent combined with TRIZ theory reveals the invention problem is "how to dock a mobile robot to a docking station according to the correct route". The invention solution is "attaching infrared LEDs to the docking station".

The numbering for each of the 220 patents, P1 to P220, is too cumbersome to show, but each patent is sorted from the most recent to the oldest filing date. P1 is the target patent. With the patent set assembled, we extracted 2833 SAO structures, 2744 of which remained after cleaning. 15 SAO structures were extracted from the DWPI-abstract of P1. To help with data processing and clarity of reference, the SAO structures for each patent were numbered—for example, from 1 to 15 for the 15 SAO structures extracted from the target patent.

Determining the optimal thresholds for (*R*) and (*Q*)

In order to distinguish the similarity between related patents and target patents, we hope that the proportion of patents with the same similarity and those with the similarity of 0 is as small as possible. Prior to calculating the initial level of patent similarity, thresholds for matching words (R) and SAO structures (Q) needed to be established. Given that different thresholds may produce different results, we designed 12 different pairs of thresholds to identify the optimal settings. The results for each pair are shown in Table 5. As shown in the table, with the same R, a larger Q results in a larger proportion of recurrence similarity and a larger proportion of patents with a similarity of 0. Compared to the other 8 threshold combinations, combinations 1, 5, and 9 meet the above requirements.

Table 5 The proportion of recurrence similarity and 0 similarity with different combinations of threshold settings

No.	R	Q	Proportion of recurrence similarity (%)	Proportion of similarity of 0 (%)
1	≥ 0.6	> 0.5	17.70	2.70
2	≥ 0.6	≥ 0.6	27.70	5.00
3	≥ 0.6	≥ 0.7	65.00	39.10
4	≥ 0.6	≥ 0.8	83.20	59.50
5	≥ 0.7	> 0.5	20.50	8.20
6	≥ 0.7	≥ 0.6	42.70	15.50
7	≥ 0.7	> 0.7	75.00	59.50
8	≥ 0.7	≥ 0.8	86.80	76.80
9	≥ 0.8	> 0.5	33.20	11.80
10	≥ 0.8	≥ 0.6	54.10	25.00
11	≥ 0.8	≥ 0.7	82.30	74.10
12	≥ 0.8	≥ 0.8	92.70	88.60

To cross-check the results, we also asked several technology professionals to manually read the patents and ensure there were minimal differences between our measurement results and their own comprehension. The results for all combinations were confirmed, but the results for combination 9 were found to be the most accurate.

The similarities between the target patent and each of the 220 related patents were measured according to two metrics: recurrence similarity and patents with a similarity of 0. Recurrence similarity means the patent similarities have the same value. Smaller recurrence similarity values represent finer distinctions in the similarity between patents. The smaller the proportion of patents with a similarity of 0 is, the more detailed the text content analysis will be to some extent. Again, the results are provided as counts and as a proportion of the structure analyzed, and smaller values representing finer levels of detail.

The similarity of SAO structures between the target patent and related patents

We then conducted the initial similarity analysis of the SAO structures using the threshold combination 9. A portion of the results is shown in Table 6.

Calculating the DWSAO for every SAO structure in the target patent

We calculated the DWSAO for every SAO structure of the target patent using the method presented in the previous section. The results are shown in Table 7. There are some obvious differences in the DWSAO weights between SAO semantic structures. Combining the main technical innovations contained in the target patent, we analyzed the SAO structures with different DWSAO.

SAO Structure 12, which is 'approach$^{(S)}$–close$^{(A)}$–docking station$^{(O)}$' had the largest DWSAO, indicating that it best embodies the technical characteristics of the target patent. Semantically, this structure concerns proper incoming approach trajectories to the docking station, and while it does not express specific technical methods, it does reflect the core content of the target patent.

Table 6 Sample results of the initial SAO structure similarity analysis

No. of SAO structures in the target patent	No. of related patents	No. of SAO structures in the related patent	Similarity
4	2	13	0.400
6	2	1	0.400
7	2	1	0.360
8	2	1	0.400
11	2	1	0.400
14	2	1	0.400
14	2	13	0.520
15	2	1	0.400
1	3	15	0.200
2	3	15	0.200
3	3	15	0.360
4	3	16	0.400
5	3	1	0.200
6	3	1	0.600
6	3	2	0.200
6	3	3	0.400
6	3	4	0.400
6	3	5	0.467
6	3	16	0.200
7	3	1	0.600

Table 7 The DWSAO of each SAO structure of the target patent

No. of SAO of target patent	DWSAO	No. of SAO of target patent	DWSAO
1	0.959	9	0.823
2	0.950	10	0.832
3	0.806	11	0.714
4	0.964	12	0.982
5	0.877	13	0.841
6	0.650	14	0.541
7	0.655	15	0.405
8	0.609		

The SAO Structure 4 'repel ray$^{(S)}$–define$^{(A)}$–improper direction$^{(O)}$', had the second largest DWSAO. Hence, this structure is also highly reflective of the technical characteristics of the patent. It identifies inappropriate directions for the repelling rays—a concept that, again, embodies the main technical characteristics of the target patent.

The SAO Structure 1 'attract ray$^{(S)}$–place$^{(A)}$–robot docking station$^{(O)}$' with the third largest DWSAO, concerns the rays that guide the robot into the docking station, while SAO Structure 2 discusses the repelling rays that guide the robot away from improper approaches with the fourth largest DWSAO.

Table 8 Top 20 ranked related patents most similar to the target patent

No. of patent	Similarity	Rank	No. of patent	Similarity	Rank
1	1.000	1	162	0.242	11
78	0.482	2	133	0.241	12
8	0.400	3	91	0.241	13
77	0.368	4	171	0.240	14
215	0.357	5	7	0.237	15
168	0.351	6	97	0.232	16
38	0.346	7	212	0.226	17
5	0.344	8	81	0.225	18
104	0.328	9	33	0.217	19
44	0.324	10	37	0.214	20

Table 9 Sample of the patent similarity measurements using the traditional method

No. of patent	Similarity	Rank	No. of patent	Similarity	Rank
1	1.000	1	7	0.357	11
78	0.500	2	44	0.357	11
8	0.476	3	162	0.357	11
77	0.438	4	215	0.357	11
5	0.429	5	38	0.348	15
91	0.387	6	50	0.348	15
81	0.385	7	171	0.345	17
104	0.381	8	130	0.333	18
97	0.370	9	164	0.333	18
133	0.370	9	33	0.323	20

The structure with the lowest DWSAO was the number 15 'schematic top view$^{(S)}$–describe$^{(A)}$–robot$^{(O)}$'. This structure appears often in robot-related designs and, therefore, is not a useful similarity.

The above analysis confirms that the DWSAO indicator does accurately reflect the importance of certain SAO structures in characterizing the technical aspects of patents. The greater the DWSAO, the stronger the ability of the SAO structure to characterize the patent's innovation.

Measuring the similarity between patents

To measure the similarity between patents, we associated the similarity of each SAO structure in the target patent with its DWSAO value according to the method presented in the previous section. Table 8 lists the similarity scores of the top 20 patents, and shows that almost no patents have the same similarity, which shows a good degree of differentiation.

Result analysis

To further verify the effectiveness of the DWSAO method, we compared our results with the traditional SAO structure method. Table 9 shows the results of this analysis.

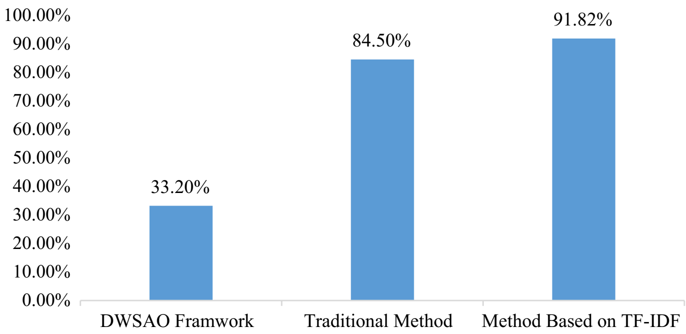

Fig. 7 The proportion of recurrence similarity with different methods

Table 10 Sample of the patent similarity measurements using a text similarity measuring method based on TF–IDF

No.	Similarity	Rank	No.	Similarity	Rank
1	1.000	1	154	0.076	11
156	0.209	2	171	0.076	11
78	0.154	3	187	0.073	13
163	0.148	4	85	0.073	13
102	0.114	5	186	0.072	15
147	0.083	6	130	0.071	16
83	0.081	7	129	0.070	17
137	0.081	7	79	0.066	18
114	0.077	9	107	0.065	19
160	0.077	9	96	0.063	20

Unlike Table 8, 17 identical patents were found and the order of similarity is very different. Further, the similarity scores show many recurrence values, which reflects poor differentiation.

We also compared our results with a text similarity measurement method based on TF–IDF. This approach only returned 3 identical patents, but in comparing these results to Tables 8 and 9, we note that the absolute values of similarity decreased significantly. This is somewhat unsurprising given that SAO semantic structures are better at reflecting specific key findings and structural relationships among technological components in inventions. This finding accords with Park et al. (2012, 2014), who showed that patent similarity measurement methods based on SAO semantic structures are superior to text-based methods. Hence, we did not conduct any further analysis on this issue (Table 10).

However, we did further examine the proportion of recurrence similarity and patents with a similarity of 0 between the traditional SAO semantic method and ours, as shown in Fig. 7. The proportion of recurrence similarity for the traditional method was 84.5%, and patents with a similarity of 0 accounted for 11.8% of the results. The corresponding results for the method presented in this paper was 33.2% for recurrence similarity and 11.8% for patents with a similarity of 0. While the results for patents with a similarity of 0 were the same for both methods, the recurrence proportion was much lower with DWSAO. From these results, we conclude that patent similarity measurement based on the DWSAO framework is significantly more accurate than traditional methods.

Since the purpose of this paper is to find the patents with a relatively high similarity to the target patent, we selected the top 7 most similar patents from each of the two measurement methods, i.e., the traditional method and the DWSAO framework. After eliminating the duplicates, we were left with 10 different patents. For the technical problem solved and the technical means involved in solving the problem in the patent, we invited three experts from the School of Automation at the Beijing Institute of Technology to manually read and rank the ten most similar patents for each measurement method. Each patent was compared to the target patent and ranked from high to low according to its similarity, as shown in Table 11. The higher the ranking, the higher the similarity. The average of all the rankings was used as the final result. "Manual reading" denotes the ranking results from the experts' readings. The "Traditional method" rankings were calculated using traditional SAO structure analysis. "DWSAO framework" was calculated using the DWSAO framework.

As shown in Table 11, the absolute difference between manual readings and the DWSAO framework was 28, and the average difference was approximately 2.8. Whereas, the absolute difference between the manual readings and the traditional method was 36 with an average difference of approximately 3.2. As Table 11 also shows, the overall rankings for the four most similar patents as determined by the DWSAO method were relatively near to those of the manual reading. This comparison further confirms our finding that the results obtained by the DWSAO framework are more accurate than traditional SAO structure analysis techniques.

Considering that the similarity based on TF–IDF method has many recurrence values, we only selected the top 5 most similar patents from each of the two measurement methods, i.e., the TF–IDF method and the DWSAO framework. After eliminating the duplicates, we were left with 9 different patents. We conducted this same comparative analysis to further confirms our finding that the results produced by the DWSAO framework are more accurate than those based on TF–IDF, as shown in Table 12. Compared to the above Table 11, although the patents with high similarity are not the same, we found that the similarity between all the same patents and the target patent have the same ranking order. And, again, there is a significant difference in the ranking for Patent 78. We find that this is largely because Patent 78 has a higher count of SAO structures that are similar to the target

Table 11 A comparison of the similarity ranks derived from manual readings, traditional method, and the DWSAO framework

No. patent	Manual reading	Traditional method	DWSAO framework
77	1	3	3
168	2	10	5
215	3	8	4
8	4	2	2
91	5	5	9
5	6	4	7
81	7	6	10
38	8	9	6
104	9	7	8
78	10	1	1
Rank change value sum	–	32	28
Average ranking change	–	3.2	2.8

Table 12 A comparison of the similarity ranks derived from manual readings, TF–IDF method and the DWSAO framework

No. patent	Manual reading	TF–IDF method	DWSAO framework
77	1	8	3
168	2	6	5
215	3	9	4
163	4	3	7
156	5c	1	8
8	6	7	2
102	7	4	6
78	8	2	1
147	9	5	9
Rank change value sum	–	36	24
Average ranking change	–	4.0	2.7

patent. Both this and the target patent address the same problem of how to connect the robot and the docking station, but each has developed a completely different technical solution. The target patent uses infrared LED technology, while Patent 78 patent adds a control unit. It is also worth noting that the TF–IDF method returned a high similarity recurrence rate, which shows it could not effectively distinguish between different patents and, hence, locating similar patents quickly would be difficult.

Conclusion

With the deepening of economic globalization, technological innovation has become a crucial means for many enterprises to remain competitive in the market. Patents are one of the most important ways to protect technological innovations and reap the maximum benefits from investments into technological development. However, before applying for a patent, applicants must search through a great deal of patent data to determine whether any similar inventions exist. In addition, before a patent is granted, patent examiners must perform a similar procedure to evaluate the novelty and innovation of an invention. Given the rapid growth in patent applications, these due diligence tasks are becoming more and more difficult. However, if applicants and examiners were to use natural language processing techniques coupled with an accurate patent similarity measurement method, these efforts would be much easier and much more effective. Hence, in view of the unique characteristics of patents, we designed a patent similarity measurement method based on SAO semantic analysis that combines text mining with a novel weighted text similarity measure called DWSAO. The DWSAO can be used to measure the mutual similarity between a 'target' patent and a set of relevant patents, or the similarities among a corpus of patents.

The inspiration for the DWSAO indicator comes from the idea of TF–IDF. DWSAO measures the importance of SAO structures to characterize patent technology by weighting similar semantic concepts that are not common in a domain more highly than those that are shared by many patents. The larger the DWSAO value, the more representative the structure is of innovation. The results from an empirical case study on robot docking stations demonstrate that weighting SAO structures according to their usefulness in indicating novelty can play a unique and effective role in identifying “relevant” similarity. In other words, DWSAO can improve the accuracy of identifying truly similar patents. The method proposed in this paper is suitable for patents written in English and is compatible with the term sets included in WordNet.

Like most studies, this research has some shortcomings. While the SAO extraction algorithm does extract the meaningful SAO structures from the patent, the initial structures do contain some noisy data. In future studies, we will further improve the extraction algorithm to reduce noise. Additionally, the similarity between words is measured based on the WordNet forest. However, some professional terms and abbreviations need to be supplemented. Therefore, future improvements to the method may construct a domain thesaurus to improve the efficiency of information processing. Third, the only source material used for extracting the SAO structures were DWPI enhanced abstracts. In future, we will look to extending the framework to accommodate other text information including the full-text and claims of the patents. This is a particularly important extension for some tasks. For example, when analyzing patent infringements, a viable analysis method would need to combine the opinions of intellectual property legal experts for a comprehensive judgment. Lastly, 12 different combinations of thresholds were tested, and the results were analyzed manually to obtain the word similarity and SAO structure similarity. In future, we intend to explore how to determine these thresholds using machine learning methods so as to improve accuracy.

Acknowledgements This work is partly supported by the General Program of the National Natural Science Foundation of China (Grant Nos. 71774012, 71673024, 71373019) and the strategic research project of the Development Planning Bureau of the Chinese Academy of Sciences (Grant No. GHJ-ZLZX-2019-42). The findings and observations present in this paper are those of the authors and do not necessarily reflect the views of the supporters or the sponsors. The authors would like to thank the anonymous reviewers for their constructive input into this paper.

Appendix

No.	Patent number	No.	Patent number	No.	Patent number	No.	Patent number
1	FR3046259A1	38	WO2016078517A1	75	GB2513912A	112	KR1179592B1
2	US20170159199A1	39	WO2016080615A1	76	KR2014120437A	113	WO2012086950A2
3	US9672184B1	40	US20160133491A1	77	KR1437778B1	114	KR1151449B1
4	WO2017091066A1	41	KR2016050285A	78	FR3002804A1	115	KR1146907B1
5	US20170105592A1	42	CN105487507A	79	US20140222197A1	116	WO2012064009A1
6	CN206115269U	43	WO2016045593A1	80	GB2509989A	117	US20120060320A1
7	US20170102709A1	44	WO2016038919A1	81	GB2509990A	118	KR2012001510A
8	CN106551659A	45	US20160062362A1	82	GB2509991A	119	WO2011136974A2
9	US20170086325A1	46	US20160057925A1	83	GB2510062A	120	DE102011010205A1
10	US20170075962A1	47	CN105361817A	84	PL402468A1	121	DE102010013297A1
11	US20170072568A1	48	US20160039541A1	85	WO2014105225A1	122	US20110236026A1
12	US20170057760A1	49	KR2016008856A	86	KR1411200B1	123	US20110238214A1
13	US20170050311A1	50	US20160011592A1	87	US20140135991A1	124	US20110214030A1
14	CN106444736A	51	WO2016000622A1	88	US20140122958A1	125	SE201100582A1
15	US20170037648A1	52	US20150356885A1	89	US20140122654A1	126	KR2011056660A
16	US20170020064A1	53	KR1569281B1	90	WO2014058106A1	127	KR2011041721A
17	US9527217B1	54	US20150314437A1	91	US20140100693A1	128	US20110092847A1

No.	Patent number	No.	Patent number	No.	Patent number	No.	Patent number
18	US20160363933A1	55	US20150314453A1	92	KR2014036653A	129	US20100324736A1
19	WO2016196622A1	56	WO2015166339A1	93	GB2505550A	130	US20100324731A1
20	US20160349756A1	57	WO2015150529A1	94	CN103576678A	131	US20100292884A1
21	CN205734931U	58	SE201500381A1	95	KR2014002841A	132	US20100228421A1
22	WO2016179782A1	59	KR2015105089A	96	EP2662742A1	133	KR2010092807A
23	US20160327959A1	60	WO2015123732A1	97	FR2987689A1	134	KR2010087820A
24	US20160325854A1	61	US9114440B1	98	DE102013101700A1	135	US20100145236A1
25	KR2016129515A	62	US20150228419A1	99	WO2013100938A1	136	US20100106298A1
26	CN205612410U	63	DE102014201203A1	100	US8373391B1	137	KR2010013362A
27	KR1660703B1	64	US20150164599A1	101	US20130035793A1	138	KR2010012351A
28	WO2016148327A1	65	EP2879010A1	102	US20120323365A1	139	KR2010007776A
29	US20160268823A1	66	KR2015053450A	103	US20120303190A1	140	SE200802217A
30	US20160237587A1	67	WO2015067225A1	104	CN102789232A	141	US20090245930A1
31	US20160240405A1	68	EP2870852A1	105	US20120277908A1	142	US20090240370A1
32	US20160236344A1	69	WO2015052588A2	106	KR2012117421A	143	PT104217A
33	US9411337B1	70	CN104416568A	107	US20120265391A1	144	WO2009092166A1
34	JP2016134081A	71	US20150063959A1	108	KR2012113188A	145	KR2009061461A
35	KR2016067351A	72	US20140379129A1	109	CN102692922A	146	KR2009053263A
36	SE201451644A1	73	WO2014201578A2	110	CN102687620A	147	KR2009051319A
37	US20160143500A1	74	KR1467887B1	111	US20120229433A1	148	US20090125174A1

No.	Patent number	No.	Patent number	No.	Patent number	No.	Patent number
149	US20090117011A1	167	KR2007103248A	185	EP1518784A2	203	US6228168B1
150	US20090049640A1	168	JP2007272301A	186	US20050010330A1	204	JP2001033357A
151	US20080275590A1	169	US20070226949A1	187	US20040210346A1	205	US6178361B1
152	WO2008106088A2	170	KR2007095558A	188	US20040204804A1	206	CA2300686A1
153	EP1961358A2	171	KR2007094288A	189	EP1435555A2	207	WO2000033355A2
154	KR2008073626A	172	US20070205215A1	190	US20040055746A1	208	EP997176A2
155	KR2008073628A	173	WO2007089269A2	191	US20040048550A1	209	WO1999065803A1
156	KR2008050278A	174	EP1806086A2	192	US6606784B1	210	US5993132A
157	US20080071417A1	175	EP1806085A2	193	US20020187024A1	211	WO1999059400A1
158	KR814784B1	176	US20070142972A1	194	EP1264935A2	212	WO1999038237A1
159	US20080062558A1	177	KR702147B1	195	US6443543B1	213	WO1999017263A1
160	US20080056933A1	178	US20060277423A1	196	WO2002055271A1	214	DE19738163A1
161	US20080038152A1	179	US20060232236A1	197	US6402846B1	215	WO1998033103A1
162	WO2008001275A2	180	US20060090320A1	198	WO2002044703A2	216	RD374022A
163	KR782863B1	181	US20060013646A1	199	US20020051700A1	217	US5324948A
164	KR2007111628A	182	GB2415252A	200	DE10033680A1	218	CA2054150A1
165	KR2007105477A	183	US20050235076A1	201	WO2002005313A2	219	US4792995A
166	US20070245511A1	184	WO2005074362A2	202	US6325808B1	220	RD246001A

References

Adams, S. R. (2006). *Information sources in patents* (pp. 234–235). Munich: K. G. Saur.

Ahlers, C. B., Fiszman, M., Demner-Fushman, D., Lang, F.-M., & Rindflesch, T. C. (2007). Extracting semantic predications from medline citations for pharmacogenomics. *Pacific Symposium on Biocomputing, 12,* 209–220.

Angeli, G., Premkumar, M. J. J., & Manning, C. D. (2015). Leveraging linguistic structure for open domain information extraction. In *Proceedings of the 53rd annual meeting of the association for computational linguistics and the 7th international joint conference on natural language processing* (Vol. 1: Long Papers, pp. 344–354).

Bär, D., Biemann, C., Gurevych, I., & Zesch, T. (2012). Ukp: Computing semantic textual similarity by combining multiple content similarity measures. In *Proceedings of the first joint conference on lexical and computational semantics-volume 1: Proceedings of the main conference and the shared task, and volume 2: Proceedings of the sixth international workshop on semantic evaluation* (pp. 435–440). Association for Computational Linguistics.

Bergmann, I., Butzke, D., Walter, L., Fuerste, J. P., Moehrle, M. G., & Erdmann, V. A. (2008). Evaluating the risk of patent infringement by means of semantic patent analysis: The case of DNA chips. *R&D Management, 38*(5), 550–562.

Boyack, K. W., Klavans, R., & Börner, K. (2005). Mapping the backbone of science. *Scientometrics, 64*(3), 351–374.

Braam, R. R., Moed, H. F., & Van Raan, A. F. (1988). Mapping of science: Critical elaboration and new approaches, a case study in agricultural biochemistry. *Journal of Informetrics, 87*(88), 15–28.

Finlayson, M. A. (2014). Java libraries for accessing the Princeton Wordnet: Comparison and evaluation. In *Proceedings of the 7th International Global WordNet Conference (GWC 2014), Tartu, Estonia* (pp. 78–85).

Keselman, A., Rosemblat, G., Kilicoglu, H., Fiszman, M., & Rindflesch, T. C. (2010). Adapting semantic natural language processing technology to address information overload in influenza epidemic management. *Journal of the American Society for Information Science and Technology, 61*(12), 2531–2543.

Kim, Y., Tian, Y., Jeong, Y., Ryu, J., & Myaeng, S. (2009). Automatic discovery of technology trends from patent text. In *Proceedings of the 2009 ACM symposium on applied computing*, Hawaii, USA.

Lin, D. (1998). An information-theoretic definition of similarity. In *International conference on machine learning* (pp. 296–304).

Magerman, T., Looy, B. V., & Song, X. (2010). Exploring the feasibility and accuracy of latent semantic analysis based text mining techniques to detect similarity between patent documents and scientific publications. *Scientometrics, 82*(2), 289–306.

Manning, C. D., & Surdeanu, M., et al. (2014). The Stanford CoreNLP natural language processing toolkit. In *52nd ACL: System demonstrations*.

Miller, G. A. (1995). Wordnet: A lexical database for english. *Communications of the Association for Computing Machinery, 38*(11), 39–41.

Moehrle, M. G. (2005). How combinations of TRIZ tools are used in companies—Results of a cluster analysis. *R&D Management, 35*(3), 285–296.

Moehrle, M. G. (2010). Measures for textual patent similarities: A guided way to select appropriate approaches. *Scientometrics, 85*(1), 95–109.

Park, H., Kim, K., Choi, S., & Yoon, J. (2013a). A patent intelligence system for strategic technology planning. *Expert Systems with Applications, 40*(7), 2373–2390.

Park, H., Yoon, J., & Kim, K. (2012). Identifying patent infringement using SAO based semantic technological similarities. *Scientometrics, 90*(2), 515–529.

Park, H., Yoon, J., & Kim, K. (2013b). Identification and evaluation of corporations for merger and acquisition strategies using patent information and text mining. *Scientometrics, 97*(3), 883–909.

Park, I., & Yoon, B. (2014). A semantic analysis approach for identifying patent infringement based on a product–patent map. *Technology Analysis & Strategic Management, 26*(8), 855–874.

Saric, F., Glavas, G., Karan, M., Snajder, J., & Basic, B. D. (2012). TakeLab: Systems for measuring semantic text similarity. In *SEM 2012 and (SemEval 2012)* (pp. 441–448), Montreal, Canada.

Sternitzke, C., & Bergmann, I. (2009). Similarity measures for document mapping: A comparative study on the level of an individual scientist. *Scientometrics, 78*(1), 113–130.

Verbitsky, M. (2004). Semantic TRIZ.triz-journal.com. http://www.triz-journal.com/archives/2004/. Accessed January 18, 2013.

Wang, X., Ma, P., Huang, Y., Guo, J., Zhu, D., Porter, A. L., et al. (2017). Combining SAO semantic analysis and morphology analysis to identify technology opportunities. *Scientometrics, 111*(1), 3–24.

Wang, X., Qiu, P., Zhu, D., Mitkova, L., Lei, M., & Porter, A. L. (2015). Identification of technology development trends based on subject–action–object analysis: The case of dye-sensitized solar cells. *Technological Forecasting and Social Change, 98,* 24–46.

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

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

Yoon, J. (2012). Detecting signals of new technological opportunities using semantic patent analysis and outlier detection. *Scientometrics, 90*(2), 445–461.
Yoon, J., Park, H., & Kim, K. (2013). Identifying technological competition trends for R&D planning using dynamic patent maps: SAO-based content analysis. *Scientometrics, 94*(1), 313–331.
Yufeng, D. U., Duo, J. I., Lixue, J., & Guiping, Z. (2016). Patent similarity measure based on SAO structure. *Journal of Chinese Information Processing*, *30*(1), 30–35 **(in Chinese)**.
Zarrella, G., Henderson, J., Merkhofer, E. M., & Strickhart, L. (2015). Mitre: Seven systems for semantic similarity in tweets. In *Proceedings of the 9th international workshop on semantic evaluation (semeval 2015)* (pp. 12–17). Denver, CO: Association for Computational Linguistics. http://www.aclweb.org/anthology/S15-2002.
Zhang, Y., Shang, L., Huang, L., Porter, A. L., Zhang, G., Lu, J., et al. (2016). A hybrid similarity measure method for patent portfolio analysis. *Journal of Informetrics, 10*(4), 1108–1130.
Zhang, Y., Zhou, X., Porter, A. L., et al. (2014). How to combine term clumping and technology roadmapping for newly emerging science & technology competitive intelligence: "Problem & solution" pattern based semantic TRIZ tool and case study. *Scientometrics, 101*(2), 1375–1389.