IEEE Access
Multidisciplinary : Rapid Review : Open Access Journal

Received February 18, 2020, accepted March 24, 2020, date of publication March 30, 2020, date of current version April 20, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.2984024

Generic SAO Similarity Measure via Extended Sørensen-Dice Index

XIAOMAN LI[1], CUI WANG[1,2], XUEFU ZHANG[1], AND WEI SUN[1]
[1]Agricultural Information Institute, Chinese Academy of Agricultural Sciences, Beijing 100081, China
[2]School of Information Science and Engineering, Shandong Agriculture and Engineering University, Jinan 250100, China

Corresponding authors: Xiaoman Li (lixiaoman@caas.cn) and Xuefu Zhang (zhangxuefu@caas.cn)

This work was supported in part by the National Social Science Found of China under Grant 16BTQ067, in part by the Innovation Project of Chinese Academy of Agricultural Sciences under Grant CAAS-ASTIP-2016-AII, and in part by the Specialized Fundamental Research Operational Fees of Chinese Academy of Agricultural Sciences under Grant Y2017ZK04.

ABSTRACT As an essential component of many Natural Language Processing applications, semantic similarity measure has been studied for decades. Recent research results indicate that the Subject-Action-Object (SAO) structure in sentences is more desirable for describing the technological information, and SAO-based similarity measure outperforms classical text-based ones. The typical approach in the literature to finding the similarity between two SAO structures relies on a term matching technique, which produces the similarity score by the Sørensen-Dice index, i.e., the proportion of the total number of matching terms. However, in this paper, we observe that the entities in the SAO structures usually have a small number of terms, which makes the currently acknowledged methods have a high recurrence rate and poor accuracy. To settle this issue, we extend the Sørensen-Dice index, and present a new unified framework for the SAO similarity measure that can give a higher discrimination. The effectiveness of our measure is evaluated on the basis of patent data sets in the Nano-Fertilizer field. The results show that our measure can significantly improve the accuracy than the currently acknowledged ones. The proposed measure has an excellent flexibility and robustness, and can be easily used for patent similarity measure. In addition, the extended Sørensen-Dice index is of independent interest, and has potential applications for other similarity measures.

INDEX TERMS Similarity measurement, Sørensen-Dice index, semantic information, Subject-Action-Object, computational linguistics.

I. INTRODUCTION

Semantic similarity analysis is an indispensable module for applications in natural language processing (NLP) and related areas [1], such as text mining [2], information retrieval [3], machine learning [4], [5], and patent analysis [6], [7]. The measure of semantic similarity can be defined as a metric assessing the degree to which two texts are similar to each other in terms of meaning. According to the measuring object, we can group the semantic similarity measures into three categories, the similarity between words/terms, the similarity between sentences, and the similarity between documents/paragraphs. The typical approach to finding the similarity between two text segments is to use a simple matching method (e.g., Sørensen-Dice index [8], [9]), and produce a similarity score based on the number of units that occur in both input segments [10]. Although such a method has been improved by considering stop-words removal [11], part-of-speech tagging [12], syntactic (word order) information [13], [14], and as well as various weighting and normalization factors [15], measuring sentence similarity [14], [16]–[20] is still challenging due to the ambiguity and variability of linguistic expression.

In linguistic typology, Subject-Action-Object (SAO) is a triple syntactic structure extracted from sentences. The subject entity and object entity are terms or phrases, which are connected by the action entity that is usually verbs. SAO is also denoted by SPO (Subject-Predicate-Object) [21] or SVO (Subject-Verb-Object) [22] in the literature. Owing to the rapid development of the NLP techniques, SAO structure can be efficiently identified, and used to express the semantic information of sentence [23]. Recently, based on the analyses of the SAO structures, a lot of new text-mining approaches

The associate editor coordinating the review of this manuscript and approving it for publication was Muhammad Afzal.

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/

are proposed [24]–[26], and widely used in patent analysis and technological evolution analysis [27].

As the case of measuring the sentences similarity, the typical method of detecting the similarity between two SAO structures is to use the Sørensen-Dice index, and evaluate the proportion of the total number of matching terms that appear in both SAO structures. Here, two terms are said to be matching if their semantic similarity score exceeds some fixed threshold. That is, the underlying term-vs-term similarity scores are compressed into two levels. Such a method is effective, and has been used for patent infringement identification [28]–[31], technological trend identification [25], [32]–[36], strategic technology planning [6], [37], document mapping [38], and etc.

However, as shown by Wang *et al.* [7], the performance of this commonly used method is far from desirable due to the relatively high recurrence rate and poor discrimination. Such a situation is caused by the fact that the number of terms (words or phrases) in the SAO structures is small. In general, in order to improve the efficiency, the collected data needs to be preprocessed, e.g., stop-words removal and transformations from complex sentence to simple sentence. Sometimes, one sentence can be dismembered and recombined into several SAO structures. In our experiment, we find that the action entity is usually just one verb, and the subject entity and object entity rarely has more than five words. In order to better demonstrate the causality, we consider following extreme situation, where all the entities in the SAO structures have just one word. Thus, according to the aforementioned method, the similarity score between the corresponding entities (including subject entity, action entity, and object entity) in SAO structures is just 0 (unmatched) or 1 (matched). We note that the overall similarity score between two SAO structures is calculated by averaging the similarity scores between corresponding entities. Then, the final similarity score can only be one of four discrete values, i.e., 0, $1/3$, $2/3$, and 1. Apparently, for such a situation, the typical similarity measure in the literature must lead to a relatively high recurrence rate and poor discrimination.

A. OUR CONTRIBUTION

In this paper, we revisit the measure of similarity between two SAO structures.

- We observe that the currently acknowledged Sørensen-Dice index is not desirable for the case where the number of terms is small. To address this issue, we extend the Sørensen-Dice index by reducing the information loss of underlying term-vs-term similarity. In particular, the acknowledged Sørensen-Dice index can just support two-levels compression, while our extended one can support arbitrary levels compression. Based on the extended Sørensen-Dice index, we presented a unified framework for the SAO similarity measure in a modular way, which can give a higher discrimination.
- The experiments are conducted based on the patent data sets in the Nano-Fertilizer field. The results show that our extended Sørensen-Dice index can dramatically reduce the recurrence rate, and our proposed SAO similarity measure can significantly improve the accuracy and F-measure compared with the acknowledged one. The application of our SAO similarity measure to the patent similarity analysis is also demonstrated.

B. ORGANIZATION

Sec. II introduce the related works. Our extended Sørensen-Dice index is shown by Sec. III. The unified framework for SAO similarity measure is given in Sec. IV. The experiment and evaluation are presented by Sec. V. In Sec. VI, we conclude our work and discuss the potential application of our proposed method.

II. RELATED WORKS

A. WORD SEMANTIC SIMILARITY

The metrics of semantic similarity between words are mainly grouped into two categories [10]. One is corpus-based measures that determine the semantic similarity using the information exclusively gained from a large corpus, a collection of written or spoken material assembled for the purpose of studying linguistic structures, frequencies, etc. Among the corpus-based measures, word relationships are derived analyzing the co-occurrence distribution in a corpus, e.g., latent semantic analysis [39] and PMI-IR algorithm [40], turning words (or terms) as high-dimensional vectors by wikipedia-based technique, e.g., Explicit Semantic Analysis [41], and using the web and search engine, e.g., Google Distance [42].

The other is knowledge-based measures, which quantify the degree of semantic similarity using information drawn from semantic network. There are several well-known measures with relatively high computational efficiency, e.g., Leacock and Chodorow [43], Wu and Palmer [44], Resnik [45], Jiang and Conrath [46] and Lin [47]. In particular, Leacock-Chodorow and Wu-Palmer are based on path and depth in the taxonomy, while Resnik, Jiang-Conrath and Lin are based on information content. A short description of these measures can be found in Sec. IV-A.

B. SENTENCE SEMANTIC SIMILARITY

Measures for detecting semantic similarity between two sentences usually utilize linguistic knowledge such as semantic relations between words and their syntactic composition. Mandreoli *et al.* [13] propose a method based on a purely syntactic approach for searching similarities within sentences. The semantic measure, given by Mihalcea *et al.* [10], combines word semantic similarity scores with word specificity scores, but the syntax structure of sentences is ignored. Li *et al.* [14] present an algorithm that takes account of semantic information and word order information. The semantic similarity of two sentences is calculated using information from a structured lexical database and from corpus statistics. Based on dynamic time warping, Liu *et al.* [48] propose a similarity measure that takes into account the semantic information, word order and the contribution of different parts of speech in a sentence. Quan *et al.* [19] combine syntactic information, semantic features, and attention weight mechanism together, and propose an efficient framework for sentence similarity.

C. SAO SEMANTIC SIMILARITY

SAO is a syntactic structure that expresses the semantic relationship between things, i.e., how the entity subject (S) of a sentence relates to the entity object (O) of a sentence through an entity action (A) [7]. Subjects can represent ''solutions'', actions can represent either the ''effect'' or the ''influence'' of the solution, and objects can represent the ''invention problem'' [49].

SAO structures can be efficiently identified and extracted using the method given by [23]. In particular, Yang *et al.* [23] introduce term clumping, and design a co-word algorithm (considering the co-occurrence with keywords) to identify SAO core components. Based on syntax-tree, they construct a hierarchical SAO extraction model, and perform the SAO cleaning and consolidation function.

Using the SAO structures to exploit the technological content of patents has significant advantages over traditional patent features [7], [50]. Hence, there is an increasing interest in studying the SAO semantic similarity metric, which has been widely used for various patent analyses, e.g., patent infringement identification [28]–[31].

Currently, the SAO-vs-SAO similarity is measured by first evaluating the entity-vs-entity similarity with the Sørensen-Dice index, and then calculating the final similarity score using weighted average. Such a method is acknowledged and widely used in [6], [25], [28]–[38]. However, as we observe in Sec. III, the entities in the SAO structures has small number of terms, which will lead to the fact that the current acknowledged measure has a high recurrence rate and poor discrimination.

D. PATENT SIMILARITY ANALYSIS

The research on analyzing patent similarity has a long history. The similarity measures can be divided into three categories, co-classification analysis, citation analysis, and keyword-based analysis. The co-classification analysis [51] relies on the patent classification codes, e.g., IPC codes, and does not involve the content information of a patent. Citation analysis relies on a patent citation network [52]. Keyword-based analysis is the most widely used method for measuring patent similarity, please refer to [53], [54]. In particular, text matching is used to measuring the technological similarity between patents [54]. SAO-based analysis is an extension of the keyword-based analysis that involves the relationships between entities. Various methodologies including co-word analysis, SAO structures, bibliographic coupling, co-citation analysis, and self-citation links are compared by [38]. The results show that the two former ones tend to describe rather semantic similarities that differ from knowledge flows as expressed by the citation-based methodologies.

III. EXTENDED SØRENSEN-DICE INDEX

A. SØRENSEN-DICE INDEX

The Sørensen-Dice index that is independently proposed by Dice [8] and Sørensen [9], is a statistic used to gauge the similarity of two samples. Originally, this index was intended for discrete data. Given two sets, X and Y, the original Sørensen-Dice index is defined as

$$\mathrm{SD}_{\mathrm{Original}} = \frac{2|X \cap Y|}{|X| + |Y|} \tag{1}$$

where $|X|$ ($|Y|$, resp.) is the cardinality of the set X (Y, resp.), i.e., the number of elements in the set. That is, the Sørensen-Dice index is equal to twice the ratio of the number of elements appearing in both sets to the sum of the number of elements in each set. We remark that in the context the sets will be instantiated by entities in SAO structure, and the elements will be instantiated by terms or words accordingly.

B. ACKNOWLEDGED SØRENSEN-DICE INDEX FOR SAO STRUCTURES

When measuring the semantic similarity between two SAO structures, direct adoption of the original Sørensen-Dice index as the metric will ignore the semantic relations between words, and result into universally low scores, poor discrimination and accuracy. This is due to the inherent flexibility of natural language enabling to express similar meanings using quite different sentences in terms of structure and word content. Thus, the SAO semantic similarity is usually measured by the following acknowledged Sørensen-Dice index exploiting the information of the underlying semantic similarity among elements in sets [7].

Given two sets $X = \{x_1, \ldots, x_m\}$ and $Y = \{y_1, \ldots, y_n\}$, and the similarity scores $\mathrm{Sim}(x_i, y_i)$ between x_i and y_j ($0 \leq \mathrm{Sim}(x_i, y_i) \leq 1$), where $i \in \{1, \ldots, m\}$ and $y \in \{1, \ldots, n\}$, the widely used Sørensen-Dice index for SAO structures is defined by[1]

$$\mathrm{SD}_{\mathrm{Acknowledged}} = \frac{2\sum_{k=1}^{\min(m,n)} \mathrm{F}(x_k, y_k)}{|X| + |Y|} \tag{2}$$

where the matching function $\mathrm{F}(x_k, y_k)$ indicates two terms x_k and y_k are matching or not, and $\sum_{k=1}^{\min(m,n)} \mathrm{F}(x_k, y_k)$ essentially counts the number of the matching terms between X and Y. In detail, $\mathrm{F}(x_k, y_k)$ is given by

$$\mathrm{F}(x_k, y_k) = \begin{cases} 1 & \text{if } R \leq \mathrm{Sim}(x_k, y_k) \leq 1 \\ 0 & \text{if } 0 \leq \mathrm{Sim}(x_k, y_k) < R \end{cases} \tag{3}$$

[1] In Sec. III, we assume the elements in sets are well ordered.

Remark: We note that above acknowledged index (2) is essentially the generalization of the original Sørensen-Dice index (1). In particular, $|X \cap Y|$ in (2) can also be interpreted as the number of the matching terms, i.e., $\sum_{k=1}^{\min(m,n)} \mathrm{F}(x_k, y_k)$, where $\mathrm{F}(x_k, y_k)$ is equal to 1 if $x_k = y_k$, and 0 otherwise.

C. OUR EXTENDED SØRENSEN-DICE INDEX FOR SAO STRUCTURES

We remark that the acknowledged Sørensen-Dice index in (2) is not desirable for the sets with small amount of elements. For example, assume the set X has just single element, i.e., $|X| = 1$. Then, according to (2), the similarity score between X and Y can only be either 0 or $2\big/(|X| + |Y|)$. Thus, such a semantic similarity measure has a quite lower discrimination.

The entities of SAO structure extracted from sentences, e.g., in the patent text, usually have small amount of terms. In particular, most of the "Action" entities have only single terms. This might be the key reason why the current widely used SAO similarity measure brings a relatively high recurrence rate, and poor accuracy.

We note that the acknowledged Sørensen-Dice index essentially gives a conversion from the term-vs-term (or element-vs-element) similarity to the entity-vs-entity (set-vs-set) similarity. However, the information loss during the conversion is very high, which is the key reason for the lower discrimination. In (3), the domain and codomain of the matching function are [0, 1] and {0, 1}, respectively. That is, the original term-vs-term similarity is compressed into two levels, 0 (unmatched) and 1 (matched). In the view of information theory, the entropy is also decreasing heavily. For example, assume the original term-vs-term similarity with precision 0.01 obeys the uniform distribution over the discrete set $\{0.01 * \lceil(100 * x)\rceil : x \in [0, 1]\}$, with Shannon entropy $\log 101 \approx 6.66$. Let the threshold value R be 0.5. Thus, the value of the matching function obeys the uniform distribution over {0, 1}, with Shannon entropy $\log(2) = 1$. That is, roughly speaking, a lot of information is compressed using the current matching function.

To solve this, we extend the Sørensen-Dice index by modifying the matching function to make support multiple-level compression and reduce the information loss. Given $R_0 = 0 < R_1 < R_2 < \ldots < R_t = 1$, the modified matching function can be defined by

$$\widetilde{\mathrm{F}}(x_k, y_k) = \begin{cases} w_1 & \text{if } R_0 \le \mathrm{Sim}(x_k, y_k) < R_1 \\ w_2 & \text{if } R_1 \le \mathrm{Sim}(x_k, y_k) < R_2 \\ \vdots & \\ w_t & \text{if } R_{t-1} \le \mathrm{Sim}(x_k, y_k) \le R_t \end{cases} \tag{4}$$

Then, accordingly, our extended Sørensen-Dice index will be

$$\mathrm{SD}_{\mathrm{Our}} = \frac{2\sum_{k=1}^{\min(m,n)} \widetilde{\mathrm{F}}(x_k, y_k)}{|X| + |Y|} \tag{5}$$

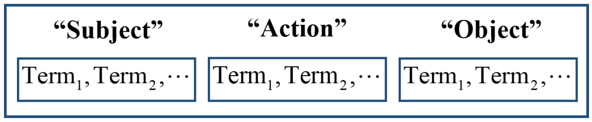

FIGURE 1. SAO structure with entities "Subject", "Action", and "Object".

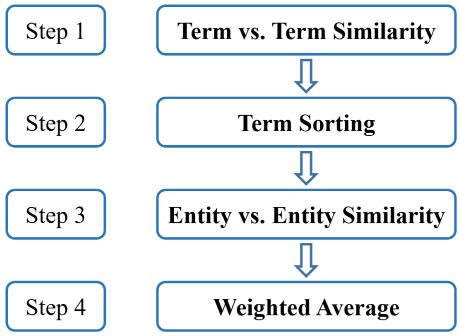

FIGURE 2. Overall procedure for measuring the similarity between two SAO structures.

1) FLEXIBILITY

Essentially, the modified matching function divides the interval [0, 1] into t subintervals and assigns fixed weights accessing the matching degree for these subintervals. We note that if we set $t = 2$, $w_1 = 0$ and $w_t = 1$, then our extended Sørensen-Dice index will be totally the same as the acknowledged one in Sec. III-B. For the aforementioned example with uniform distribution, the Shannon entropy will be $\log t$. If we choose $t \ge 3$, apparently, the information loss will be reduced.

2) ROBUSTNESS

One may argue that if we directly choose the term-vs-term similarity score $\mathrm{Sim}(x_k, y_k)$ as the matching function $\widetilde{F}(x_k, y_k)$, there will be no information loss for the matching function. However, we note that the underlying term-vs-term semantic similarity score is usually not precise enough, due to incomplete corpus. In fact, there exists even no domain thesaurus for some frontier field, which results in that many excellent word-vs-word semantic similarity measure will not work. As we have pointed, our extended matching function is essentially a compressing function. Thus, with this function, some noises (errors) existing in the underlying term-vs-term similarity score can be eliminated (corrected). We also remark that the Sørensen-Dice index is sometimes not the final similarity score, e.g., as an intermedium for our SAO similarity measure. Thus, eliminating noises in time can avoid error accumulation. Thus, our extended Sørensen-Dice index can also help improve the robustness of similarity measure systems.

IV. A UNIFIED FRAMEWORK FOR SAO SIMILARITY MEASURE

In this section, using the extended Sørensen-Dice index presented in Sec. III, we give a unified framework for SAO similarity measure. A SAO structure consists of three entities including "Subject", "Action", and "Object", see Fig. 1. Every entity is composed of several terms, which refer to words or phrases.

Given two SAO structures, we can quantify the degree of similarity by four steps, see Fig. 2. First, we calculate the term-vs-term similarity. Next, using the term-vs-term similarity scores, we reorder the terms in the entities. Then, with the extended Sørensen-Dice index, we can calculate the entity-vs-entity similarity scores. Finally, the SAO-vs-SAO similarity can be measured by a weighted average method.

A. TERM-VS-TERM SIMILARITY

The semantic similarity between terms/words has been well studied, and there are a relatively large number of metrics that have been proposed in the literature [1], [10], [55]. Below, we present five measures that have excellent performance and relatively high computational efficiency in NLP application. We remark that although we just select following five term-vs-term measures to test the effectiveness of our methods, the other term-vs-term measures can also work well with this framework.

We note that most term-vs-term similarity measures are defined for concepts,[2] but they can be easily turned into a word-to-word similarity metric by selecting for any given pair of words those two meanings that lead to the highest concept-vs-concept similarity [10]. In the following, we give a short description for each of these five metrics. These metrics use the WordNet [56] as a knowledge source. WordNet[3] is a large lexical database for English, where Nouns, verbs, adjectives and adverbs are grouped into sets of cognitive synonyms (synsets), each expressing a distinct concept. Synsets are interlinked by means of conceptual-semantic and lexical relations. Let c_1 and c_2 be two concepts.

Leacock and Chodorow [43]: This measure of Leacock-Chodorow Similarity is in basis of the shortest path that connects the concepts and the maximum depth of the taxonomy in which the concepts occur. The similarity is quantified by

$$\mathrm{Sim}_{lch}(c_1, c_2) = -\log \frac{length(c_1, c_2)}{2D} \tag{6}$$

where *length* is the length of the shortest path between two concepts using node-counting, and D is the maximum depth of the taxonomy.

Wu and Palmer [44]: The Wu-Palmer Similarity is based on the depth of the two concepts in the taxonomy and that of their Least Common Subsumer (LCS, most specific ancestor node). The similarity score is given by

$$\mathrm{Sim}_{wup}(c_1, c_2) = \frac{2 \times depth(\mathrm{LCS})}{depth(c_1) + depth(c_2)} \tag{7}$$

[2]Concept in this paper refers to a particular sense of a given word.

[3]More details about WordNet can be found at https://wordnet.princeton.edu/.

Resnik [45]: The Resnik similarity is based on the information content of the LCS. The similarity is identified by

$$\mathrm{Sim}_{res}(c_1, c_2) = -\log \Pr[\mathrm{LCS}] \tag{8}$$

where $\Pr[c]$ is the probability of encountering an instance of concept c in a large corpus.

Jiang and Conrath [46]: The Jiang-Conrath Similarity is based on the information content of the LCS and that of the two input Synsets. The similarity score is given by

$$\mathrm{Sim}_{jcn}(c_1, c_2) = \frac{1}{2\log \Pr[\mathrm{LCS}] - \log(\Pr[c_1] \cdot \Pr[c_2])} \tag{9}$$

Lin [47]: The Lin Similarity is based on the same elements as the Jiang-Conrath Similarity. The similarity score is given by

$$\mathrm{Sim}_{lin}(c_1, c_2) = \frac{2\log \Pr[\mathrm{LCS}]}{\log(\Pr[c_1] \cdot \Pr[c_2])} \tag{10}$$

Remark: We note that the ranges of the similarity scores for measures Leacock-Chodorow, Resnik and Jiang-Conrath are not [0, 1]. We use the following normalization method suggested by [57] to make the ranges between 0 and 1,

$$\mathrm{Sim}_{norm}(c_1, c_2) = \frac{\mathrm{Sim}(c_1, c_2)}{\mathrm{Sim}(c_1, c_1) * \mathrm{Sim}(c_2, c_2)}$$

B. TERM SORTING

Before evaluating the degree of similarity between entities, we need to adjust the order of terms in entities for the subsequent term matching. The goal of term sorting is to achieve a globally optimal term matching, i.e., the sum of the term-vs-term similarity scores between the corresponding terms with the same position in entities is maximum.

Let $E_1 = \{\mathrm{Term}_1^1, \ldots, \mathrm{Term}_m^1\}$ ($E_2 = \{\mathrm{Term}_1^2, \ldots, \mathrm{Term}_n^2\}$, resp.) be an entity with m (n, resp.) terms. Without loss of generality, we assume $m \geq n$. Then, mathematically, we need to search for a permutation of $\{1, \ldots, n\}$ such that the following objective function reaches the maximum value,

$$f(E_1, E_2) = \sum_{i=1}^{n} \mathrm{Sim}(\mathrm{Term}_i^1, \mathrm{Term}_i^2),$$

where $\mathrm{Sim}(\mathrm{Term}_i^1, \mathrm{Term}_i^2)$ is the similarity score between Term_i^1 and Term_i^2 obtained by Sec. IV-A.

We note that this problem can be considered as a combinatorial optimization problem of finding the maximum-weight matching in the weighted bipartite graphs, for which many algorithms have been proposed, e.g., the Hungarian method [58]. In this paper, considering the number of terms in entities is not large, we will adopt an efficient greedy algorithm to solve this problem, which gives a nearly optimal solution. The algorithm is illustrated by Algorithm 1.

Algorithm 1: Greedy Algorithm for Term Sorting

Input: Entities E_1 and E_2
Output: E_1 and E_2 with updated term order

$m := length(E_1)$, $n := length(E_2)$;
for $k \leftarrow 1$ ***to*** $\min(m, n)$ **do**
  $max_{temp} := -1$;
  /*Search for the k-th maximum matching */
  for $i \leftarrow k$ ***to*** m **do**
    for $j \leftarrow k$ ***to*** n **do**
      $sim_{temp} = \mathrm{Sim}(\mathrm{Term}_i^1, \mathrm{Term}_j^2)$;
      if $sim_{temp} > max_{temp}$ **then**
        $flag_i := i$;
        $flag_j := j$;
        $max_{temp} := sim_{temp}$;
  /*Reorder the terms in E_1 and E_2 */
  /*Swap Term_k^1 with $\mathrm{Term}_{flag_i}^1$ */
  $\mathrm{Temp} := \mathrm{Term}_k^1$;
  $\mathrm{Term}_k^1 := \mathrm{Term}_{flag_i}^1$;
  $\mathrm{Term}_{flag_i}^1 := \mathrm{Temp}$;
  /*Swap Term_k^2 with $\mathrm{Term}_{flag_j}^2$ */
  $\mathrm{Temp} := \mathrm{Term}_k^2$;
  $\mathrm{Term}_k^2 := \mathrm{Term}_{flag_j}^2$;
  $\mathrm{Term}_{flag_j}^2 := \mathrm{Temp}$;

C. ENTITY-VS-ENTITY SIMILARITY

After sorting the terms in entities, we can quantify the degree of similarity between entities by using our extended Sørensen-Dice index in Sec. III-C.

Specifically, the similarity score between two entities E_1 and E_2 is calculated by

$$\mathrm{Sim}(E_1, E_2) = \frac{2 \sum_{k=1}^{\min(m,n)} \widetilde{\mathrm{F}}(\mathrm{Term}_k^1, \mathrm{Term}_k^2)}{m+n} \tag{11}$$

where the matching function $\widetilde{\mathrm{F}}$ is given by (4).

D. WEIGHTED AVERAGE

Finally, starting from the similarity between entities, we can identify the degree of similarity between two SAO structures with weighted average.

Let SAO_i ($i \in \{1, 2\}$) be a SAO structure with entities S_i ("Subject"), A_i ("Action"), and O_i ("Object"). Note that subjects and objects are nouns, actions are verbs. Thus, A_1 can only match A_2, S_1 (O_1) can match S_2 or O_2. With weighted average, the similarity score between SAO_1 and SAO_2 can be evaluated by

$$\mathrm{Sim}(\mathrm{SAO}_1, \mathrm{SAO}_2) = \max(\mathrm{Comb}_1, \mathrm{Comb}_2) \tag{12}$$

where

$$\mathrm{Comb}_1 = \alpha_1 \mathrm{Sim}(S_1, S_1) + \alpha_2 \mathrm{Sim}(A_1, A_2) + \alpha_3 \mathrm{Sim}(O_1, O_2),$$
$$\mathrm{Comb}_2 = \alpha_1 \mathrm{Sim}(S_1, O_1) + \alpha_2 \mathrm{Sim}(A_1, A_2) + \alpha_3 \mathrm{Sim}(O_1, S_2),$$

TABLE 1. The patents used in the experiments.

No.	Patent number	No.	Patent number	No.	Patent number
1	CN108046929A	16	CN108358700A	31	CN108484329A
2	CN108101666A	17	CN108358703A	32	CN108503043A
3	CN108142059A	18	CN108358710A	33	CN108503429A
4	CN108147925A	19	CN108424225A	34	CN108516884A
5	CN108191520A	20	CN108424300A	35	CN108530160A
6	CN108249989A	21	CN108440162A	36	CN108530202A
7	CN108264422A	22	CN108440206A	37	CN108530210A
8	CN108285400A	23	CN108440210A	38	CN108530212A
9	CN108314486A	24	CN108456062A	39	CN108546161A
10	CN108314540A	25	CN108456063A	40	CN108558504A
11	CN108314541A	26	CN108456083A	41	CN108558512A
12	CN108314556A	27	CN108456115A	42	CN108558515A
13	CN108329088A	28	CN108484295A	43	CN108558524A
14	CN108329151A	29	CN108484298A	44	CN108586082A
15	CN108341706A	30	CN108484309A	45	CN108623399A

and α_1, α_2 and α_3 are non-negative weight coefficients such that $\alpha_1 + \alpha_2 + \alpha_3 = 1$.

Remark: We note that Verb (or Action) is usually a single word, Subject and Object are usually a noun-phrase. But, the most-right noun in a noun-phrase can generally represent the noun-phrase. Thus, for most cases, our method can be simplified by removing step 1 and step 2. But, in some cases (e.g., ecological fertilizer vs. composite fertilizer), the left adjective plays a more important role in evaluating the similarity between two noun-phrases. This paper focuses on a unified and generic framework for SAO similarity measure that can apply more complicated cases. Therefore, the step 1 and step 2 are necessary.

V. EVALUATION AND RESULTS

To evaluate the effectiveness of our semantic similarity measure, we perform several experiments using the computer with Intel(R) Core(TM) i5-4210U processor 4GHz and 8GHz RAM. The programming language used is Python2.7. The knowledge source WordNet used in the term-vs-term similarity measures is loaded by NLTK (Natural Language Toolkit).

A. DATA COLLECTION AND PREPROCESSING

As one of the most important and effective ways to protect technological achievements, patent documents contain a lot of new scientific and technological information. As we have showed in the introduction, the measure of semantic similarity between the SAO structures is widely used in patent analysis. Therefore, in our experiments, we choose the patent documents as data sets. In particular, we downloaded 45 patent documents in the Nano-Fertilizer field published in 2018 from the Derwent Innovation Index patent database, where Nano-Fertilizer is a new fertilizer constructed by nano material and pharmaceutical microencapsulation technology, and has a landmark application in agriculture [59]. The patent numbers are given by Table 1.

We remark that the SAO structures can be extracted from any description in textual format including title, abstract, claims, and description sections of a patent document. But, in this paper, considering the title and the abstract are precise and have been regarded as the most meaningful part in a patent document, we follow prior works, e.g., [7], and just extract SAO structures from the title and abstract. The SAO extractor is designed by following a standard procedure, as given by [7]. For the sentences in the abstract, we perform a syntactic analysis using the Stanford parser, and every entities in the SAO structure are elaborately determined. Thus, 1126 SAO structures are collected from the 45 patents in Table 1. Finally, we clean the SAO structures by removing meaningless stop words, extraneous parts of speech, etc.

B. THE SEMANTIC SIMILARITY MEASURES BETWEEN THE SAO STRUCTURES

1) TERM-VS-TERM SIMILARITY

In the experiments, the term-vs-term similarity measures presented in IV-A, including Leacock-Chodorow, Wu-Palmer, Resnik, Jiang-Conrath, and Lin, can be directly implemented using the NLTK WordNet. Note that these five measures are used to quantify the degree of simialrity from different aspects. The Leacock-Chodorow and Wu-Palmer similarity measures are based on the path and depth in the taxonomy, while the Resnik, Jiang-Conrath, and Lin similarity measures are based on an information content dictionary from the WordNet corpus. Except Wu-Palmer, the other four similarity measures require the concepts having the same part of speech (POS). The Resnik, Jiang-Conrath, and Lin similarity measures can not apply to the concepts with the *adjective* and *adverb* POS. The experiments in [10] show that the best performance can be achieved by combing these measures with a simple average. Therefore, in this paper, we take the average of similarity scores obtained using above five measures as the final scores indicating the similarity between terms.

2) ENTITY-VS-ENTITY SIMILARITY

To show the effectiveness of our extended Sørensen-Dice index, we perform similarity measures for 101 entities randomly selected from the 1126 SAO structures, please refer to Table 6 in the Appendix. In particular, we take the first entity as the target entity, and the remaining as the entities to be compared. That is, the similarity will be evaluated among 100 pairs of entities.

For simplicity, in our extended Sørensen-Dice index, we set $R_i - R_{i-1} = \frac{1}{t}$ $(i \in \{1, \ldots, t\})$, i.e., the internal $[0, 1]$ is divided into t subintervals with equal length indicating different levels. The weight w_i corresponding to the i-th subintervals is set to be $\frac{R_{i-1}+R_i}{2}$. Our method for entity-vs-entity similarity is implemented with level $t = 3, 4, 5, 10, 20, 30, 40, 50, 100$. For a comprehensive comparison, we also conduct the acknowledged similarity measure in (2) with threshold $R = 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9$. The recurrence rate is calculated by

$$P_{recurr} = \frac{N - N_{diff}}{N} \tag{13}$$

TABLE 2. The recurrence rate comparison for the entity-vs-entity similarity.

Our method			Acknowledged method		
Level t	N_{diff}	P_{recurr}	Threshold R	N_{diff}	P_{recurr}
3	21	0.79	0.1	8	0.92
4	24	0.76	0.2	9	0.91
5	32	0.68	0.3	11	0.89
10	38	0.62	0.4	11	0.89
20	43	0.57	0.5	10	0.90
30	46	0.54	0.6	8	0.92
40	50	0.50	0.7	7	0.93
50	51	0.49	0.8	8	0.92
100	60	0.40	0.9	8	0.92

where N is the total experiment number, and N_{diff} is the number of different similarity scores.

The recurrence rate comparisons between our method and the acknowledged method for the entity-vs-entity similarity are given by Table 2. We can find that the recurrence rate is significantly reduced with our method. This is consistent with our theoretical analysis in Sec. III, which shows that our extended Sørensen-Dice index can reduce the loss of the underlying term-vs-term similarity information, and further reduce the recurrence rate. We remark that the concrete value of recurrence rate is also highly influenced by the total experiment number N. If the similarity score has a precision of two decimal figures, then the recurrence rate is at least $\frac{N-100}{N}$ when $N > 100$. Thus, the recurrence rate can not be very low, e.g., approximately approaching 0. From Table 2, we can see that even though we set the level to be 100, corresponding the precision 0.01, the recurrence rate is still 0.4. We also remark that the lower recurrence rates do not always increase the accuracy, please see Sec. V-B3 for details.

To further reveal the relationship between our method with different levels and the acknowledged method with different thresholds, we calculate the Pearson correlation factor among all the obtained similarity scores. As shown by Table 3, the Pearson correlation factors among different levels from $t = 3$ to $t = 100$ for our method is at least 0.964. In particular, for the levels from $t = 5$ to $t = 100$, the Pearson correlation factors can reach at least 0.991. For the levels $t \geq 20$, the Pearson correlation factors can reach maximum 1. That is, our extended Sørensen-Dice index with more levels makes no sense since they essentially give the same similarity metric. Fig. 3 shows the entity-vs-entity similarity scores using our method with levels $t = 5, 20$ and 100.

From Table 3, we can see the similarity scores using the acknowledged method with different thresholds have some certain positive correlation, although lower than our method with different levels. The lowest Pearson correlation factor is 0.443 between thresholds $R = 0.1$ and $R = 0.6$. While the maximum Pearson correlation factor is 1 between thresholds $R = 0.8$ and $R = 0.9$. That is, the acknowledged

TABLE 3. **Pearson correlation factor among the entity-vs-entity similarity scores using our method with different levels and the acknowledged method with different thresholds.**

	R=0.1	R=0.2	R=0.3	R=0.4	R=0.5	R=0.6	R=0.7	R=0.8	R=0.9	t=3	t= 4	t=5	t=10	t=20	t=30	t= 40	t=50	t=100
R=0.1	1.000	0.938	0.780	0.637	0.535	0.443	0.438	0.462	0.462	0.682	0.737	0.739	0.768	0.780	0.782	0.785	0.784	0.786
R=0.2	0.938	1.000	0.804	0.654	0.532	0.480	0.473	0.492	0.492	0.690	0.753	0.769	0.784	0.793	0.795	0.799	0.798	0.799
R=0.3	0.780	0.804	1.000	0.782	0.658	0.592	0.587	0.599	0.599	0.837	0.819	0.809	0.847	0.848	0.851	0.851	0.850	0.852
R=0.4	0.637	0.654	0.782	1.000	0.842	0.786	0.781	0.788	0.788	0.912	0.890	0.931	0.916	0.917	0.917	0.917	0.917	0.917
R=0.5	0.535	0.532	0.658	0.842	1.000	0.898	0.884	0.853	0.853	0.864	0.907	0.880	0.894	0.883	0.884	0.881	0.881	0.881
R=0.6	0.443	0.480	0.592	0.786	0.898	1.000	0.982	0.937	0.937	0.875	0.877	0.890	0.883	0.874	0.871	0.869	0.870	0.868
R=0.7	0.438	0.473	0.587	0.781	0.884	0.982	1.000	0.948	0.948	0.882	0.875	0.886	0.881	0.873	0.868	0.867	0.868	0.866
R=0.8	0.462	0.492	0.599	0.788	0.853	0.937	0.948	1.000	1.000	0.872	0.889	0.896	0.886	0.883	0.880	0.879	0.879	0.878
R=0.9	0.462	0.492	0.599	0.788	0.853	0.937	0.948	1.000	1.000	0.872	0.889	0.896	0.886	0.883	0.880	0.879	0.879	0.878
t=3	0.682	0.690	0.837	0.912	0.864	0.875	0.882	0.872	0.872	1.000	0.958	0.966	0.968	0.967	0.966	0.965	0.965	0.964
t=4	0.737	0.753	0.819	0.890	0.907	0.877	0.875	0.889	0.889	0.958	1.000	0.979	0.986	0.985	0.984	0.984	0.983	0.983
t=5	0.739	0.769	0.809	0.931	0.880	0.890	0.886	0.896	0.896	0.966	0.979	1.000	0.993	0.992	0.991	0.991	0.991	0.991
t=10	0.768	0.784	0.847	0.916	0.894	0.883	0.881	0.886	0.886	0.968	0.986	0.993	1.000	0.999	0.998	0.998	0.998	0.998
t=20	0.780	0.793	0.848	0.917	0.883	0.874	0.873	0.883	0.883	0.967	0.985	0.992	0.999	1.000	1.000	1.000	1.000	1.000
t=30	0.782	0.795	0.851	0.917	0.884	0.871	0.868	0.880	0.880	0.966	0.984	0.991	0.998	1.000	1.000	1.000	1.000	1.000
t=40	0.785	0.799	0.851	0.917	0.881	0.869	0.867	0.879	0.879	0.965	0.984	0.991	0.998	1.000	1.000	1.000	1.000	1.000
t=50	0.784	0.798	0.850	0.917	0.881	0.870	0.868	0.879	0.879	0.965	0.983	0.991	0.998	1.000	1.000	1.000	1.000	1.000
t=100	0.786	0.799	0.852	0.917	0.881	0.868	0.866	0.878	0.878	0.964	0.983	0.991	0.998	1.000	1.000	1.000	1.000	1.000

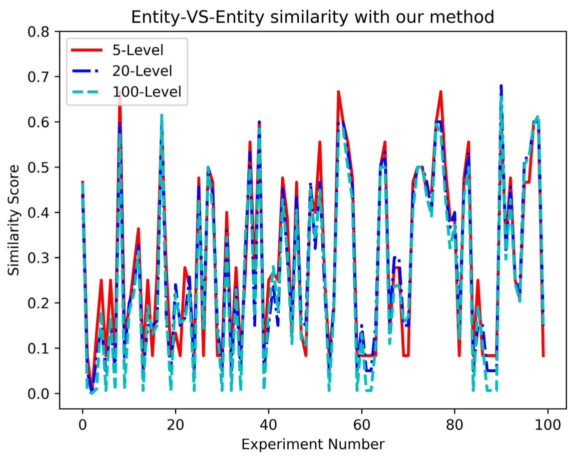

FIGURE 3. **Entity-vs-entity similarity with our method.**

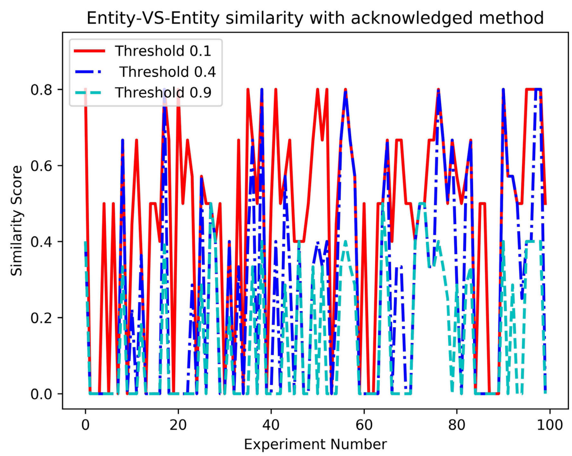

FIGURE 4. **Entity-vs-entity similarity with acknowledged method.**

method with thresholds $R = 0.8$ and $R = 0.9$ can essentially give the same similarity scores. Overall, the acknowledged method is more sensitive to the parameter change than our method. Thus, our method has a better robustness. Fig. 4 shows the entity-vs-entity similarity scores using the acknowledged method with thresholds $R = 0.1, 0.4$ and 0.9.

Table 3 also shows the Person correlation factors between our method with differ levels and the acknowledged method with different thresholds, please see the bottom left or top right of the table. We can see that the minimum is 0.682 and the maximum is 0.917. Thus, generally, our method is positively correlated with the acknowledged method. This is because that our extended Sørensen-Dice index is essentially the generalization of the acknowledged one, which can be seen as the our method with two levels, i.e., $t = 2$.

3) SAO-VS-SAO SIMILARITY

For the 1126 SAO structures extracted in Sec. V-A, we choose the first SAO structure as a target SAO structure, and take the other 1125 ones as the SAO structures to be compared. Thus, 1125 pairs of SAO structures are prepared. First, these pairs are manually labelled by three human annotators who are familiar with expertise in the field of Nano-Fertilizer, and together determine if the two SAO structures in a pair are semantically equivalent ("1") or not ("0"). We take these manual classification as the actual class of these pairs of SAO structures. Then, using the method presented in Sec. IV, we calculate the similarity scores among the 1125 pairs of SAO structures, and then identify them by "1" ("0", resp.) when the similarity score exceeds (does not exceed, resp.) a threshold of 0.5. In addition, we also label these pairs with the acknowledged method, which is the same as our method except using the acknowledged Sørensen-Dice index

TABLE 4. Performance comparisons between similarity measures for the SAO structures.

Our method				
Level	Accuracy	Precision	Recall	F-measure
t=3	0.878	0.860	0.926	0.892
t=4	0.876	0.875	0.898	0.887
t=5	**0.900**	0.890	**0.931**	**0.910**
t=10	0.887	0.889	0.905	0.897
t=20	0.889	**0.894**	0.902	0.898
t=30	0.890	0.893	0.905	0.899
t=40	0.885	0.891	0.898	0.895
t=50	0.887	0.891	0.902	0.896
t=100	0.887	0.891	0.902	0.896
Acknowledged method				
Threshold	Accurate	Precision	Recall	F-measure
R=0.1	0.550	0.547	**0.997**	0.706
R=0.2	0.553	0.549	0.984	0.705
R=0.3	0.572	0.563	0.946	0.706
R=0.4	**0.810**	0.789	0.887	**0.835**
R=0.5	0.588	0.858	0.287	0.430
R=0.6	0.535	0.922	0.156	0.266
R=0.7	0.531	**0.927**	0.146	0.252
R=0.8	0.511	0.905	0.110	0.196
R=0.9	0.511	0.905	0.110	0.196

(see Sec. III-B) to quantify the degree of the entity-vs-entity similarity. Table 7 shows the concrete similarity scores derived by human annotators, our method with $t = 5$, and the acknowledged method with $R = 0.4$. The complete similarity scores by our method and the acknowledged method with other parameters are posted at `https://github.com/l-x-m/SAO-similarity-measure`, where the data sets and python script are also provided.

We evaluate the results in terms of accuracy, representing the percentage of correctly identified true or false classifications. We also measure precision, recall and F-measure, calculated with respect to the true values in the classifications. The F-measure is the weighted average of precision and recall, and can be calculated by

$$\text{F-measure} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{14}$$

As shown by Table 4, the maximum accuracy and F-measure using our method can reach 90% and 91%, respectively, with level $t = 5$. While the highest accuracy and F-measure of the currently acknowledged method the can only attain 81% and 83.5%, respectively, with threshold $R = 0.4$. That is, using our extended Sørensen-Dice index for SAO similarity measure can significantly improve the accuracy and F-measure than the currently acknowledged one. We also remark that our method also has an excellent robustness, and the accuracy and F-measure vary little with the change of the level t. But the accuracy and F-measure of the currently acknowledged method is sensitive to the threshold R, as shown by Table 4.

We note that the lowest recurrence rate for our method is achieved with highest level $t = 100$ (see Table 2), but the highest accuracy is obtained with level $t = 5$. That is, lower loss of the underlying term-vs-term similarity information does not always bring into higher accuracy. This is due to the fact that the underlying term-vs-term similarity is not always accurate enough, i.e., there are some noises, especially when the data comes from some specific field. We also note that lower level t can reduce the influences of noises in the underlying term-vs-term similarity. Therefore, there is a balance between reductions of the information loss and noise influences. In our experiments, optimal balance can be achieved by setting $t = 5$.

TABLE 5. Similarity scores between the first patent and the remaining patents with Nos. from 2 to 45.

No.	$\mathrm{Sim}(1,\cdot)$	No.	$\mathrm{Sim}(1,\cdot)$	No.	$\mathrm{Sim}(1,\cdot)$	No.	$\mathrm{Sim}(1,\cdot)$
2	0.610	13	0.705	24	0.451	35	0.544
3	0.416	14	0.595	25	0.624	36	0.692
4	0.726	15	0.651	26	0.447	37	0.630
5	0.633	16	0.570	27	0.591	38	0.615
6	0.546	17	0.407	28	0.350	39	0.447
7	0.569	18	0.659	29	0.600	40	0.648
8	0.507	19	0.488	30	0.526	41	0.515
9	0.617	20	0.675	31	0.714	42	0.549
10	0.400	21	0.481	32	0.655	43	0.531
11	0.589	22	0.608	33	0.439	44	0.378
12	0.576	23	0.529	34	0.660	45	0.539

C. APPLICATION TO MEASURING PATENT SIMILARITY

Patent has been proved to be one of the most important and effective ways to protect technological inventions. The rapid increase of the patent number has called for the development of sophisticated patent analysis tools, of which many are based on patent similarity identification techniques. In particular, patent similarity analysis has been used for infringement identification [28]–[31], technological trend identification [25], [32]–[36], strategic technology planning [6], [37], document mapping [38], and etc.

With our SAO-vs-SAO similarity measure, we can easily evaluate the similarity between two patents. We view the SAO structure as a term, and the patent as a new entity composed of several SAO structures. Then, using the same method in the entity-vs-entity similarity measure, we first sort the SAO structures in the patents, and then utilize our extended Sørensen-Dice index to calculate the similarity score of two patents. Setting the level to be 5, i.e., $t = 5$, we calculate the similarity scores between the first patent and the remaining patents. The results are given by Table 5.

VI. CONCLUSION AND DISCUSSION

In this paper, we observe that the currently acknowledged SAO similarity measure has a relatively high recurrence rate and poor discrimination, which is caused by the fact that the entities in the SAO structure always have a small amount of terms. To settle such issues, we extend the Sørensen-Dice index by reducing the information loss of

TABLE 6. The entities selected from the SAO structures.

No.	Entities	No.	Entities
1	Water soluble fertilizer	2	composite fertilizer
3	recycles	4	Trichoderma reesei
5	is not easy to	6	potassium fulvate
7	accurate	8	acetylacetone tin(IV) dichloride
9	degrades	10	nano slow release fertilizer
11	can effectively absorb	12	deactivated silicon containing molecular sieve pretreatment agent
13	zinc ion porphyrin nanocomposites	14	Nano material modified waste oil coated controlled release fertilizer
15	is useful for	16	forms
17	first masterbatch	18	be sprayed onto
19	complex fertilizer	20	impact resistant chitosan coating
21	sprayed	22	bauxite tailings
23	has no	24	composite microbial inoculum
25	High efficiency modifying agent	26	can alleviate
27	Fertilizer anti-caking agent	28	colloid
29	foliar fertilizer	30	nanocarbon synergistic fertilizer
31	remains	32	enhances
33	Impact resistant peach chitosan coated slow release fertilizer	34	reduce
35	Anti caking agent	36	accelerate
37	compost material	38	Bio organic composite fertilizer
39	fermenting	40	Rice fertilizer
41	satisfies	42	membrane
43	The preparation method	44	be
45	Selenium enriched bio-organic fertilizer	46	raw material mixture
47	concentrated biogas slurry	48	Synergistic fertilizer
49	does not	50	detecting
51	fruit tree organicic fertilizer	52	soil conditioner
53	Nano modified calcium fertilizer	54	coating material
55	removing	56	inactive silicon containing molecular sieve preprocessing agent
57	Sustained release fertilizer	58	Potato fertilizer
59	decomposed organic fertilizer	60	Selenium enriched organic fertilizer
61	modifying	62	packaging
63	apply	64	ensures
65	additive	66	Bio organicic fertilizer
67	organic ecological fertilizer	68	mixed biogas slurry A
69	mixed biogas slurry B	70	mixed biogas slurry C
71	add	72	cooling
73	Odorless organicic fertilizer	74	Fertilizer
75	Multifunctional foliar fertilizer	76	Selenium enriched pitaya organicic fertilizer
77	Environmentally friendly fertilizer	78	organic fertilizer
79	trace element fertilizer	80	Selenium germanium enriched element fertilizer
81	Soil remediation agent	82	Selenium enriched agriculture fertilizer
83	stirring	84	nitrogen phosphate potassium fertilizer
85	solid organic fertilizer	86	regulate
87	Composition	88	monitoring
89	separating	90	drying
91	preventing	92	compound fertilizer
93	calcium chloride nano raw material	94	Selenium enriched pitaya organic fertilizer
95	synergbetic agent	96	natural high selenium nutritional powder
97	fertilizer core	98	silicon fertilizer
99	Nanocarbon organic fertilizer	100	Nano complex synergbetic fertilizer
101	second masterbatch		

underlying term-vs-term similarity. Based on that, we present a unified framework for the SAO similarity measure, which can give a higher discrimination. The effectiveness of our measure is evaluated on the basis of data sets from the Derwent Innovation Index patent database. The experiment results show that our measure can significantly improve the accuracy and F-meaure than the currently acknowledged ones.

The proposed SAO measure is generic and modular, and has an excellent flexibility and robustness. With this unified SAO measure, patent similarity metric can be easily established, which can be further used for various patent analyses, including patent infringement identification, technological trend identification, strategic technology planning, and etc. In addition, the extended Sørensen-Dice index is of independent interest, and has potential applications for other similarity measures, e.g., Jaccard index, Szymkiewicz-Simpson index, and etc.

APPENDIX

See tables 6 and 7.

TABLE 7. Similarity scores between the first SAO structure and the remaining structures with Nos. from 2 to 1126. The Hum. column shows the scores derived by human annotators. The Ours column shows the scores calculated by our method with $t = 5$. The Ack. column shows the scores evaluated by the acknowledged method with $R = 0 : 4$.

No.	Hum.	Ours t=5	Ack. R=0.4	No.	Hum.	Ours t=5	Ack. R=0.4	No.	Hum.	Ours t=5	Ack. R=0.4	No.	Hum.	Ours t=5	Ack. R=0.4	No.	Hum.	Ours t=5	Ack. R=0.4
2	1	0.738	0.708	227	0	0.288	0.233	452	0	0.381	0.650	677	1	0.592	0.533	902	1	0.600	0.675
3	1	0.773	0.708	228	0	0.382	0.408	453	0	0.522	0.592	678	1	0.627	0.708	903	1	0.544	0.640
4	1	0.662	0.767	229	0	0.335	0.233	454	0	0.522	0.592	679	1	0.571	0.673	904	1	0.530	0.517
5	1	0.545	0.533	230	0	0.510	0.533	455	0	0.494	0.557	680	0	0.493	0.580	905	1	0.530	0.500
6	1	0.629	0.533	231	0	0.335	0.233	456	1	0.530	0.592	681	0	0.353	0.300	906	1	0.553	0.617
7	1	0.650	0.580	232	0	0.335	0.233	457	1	0.565	0.675	682	0	0.376	0.300	907	1	0.516	0.533
8	1	0.662	0.767	233	0	0.382	0.408	458	1	0.565	0.697	683	0	0.423	0.475	908	1	0.530	0.557
9	1	0.808	0.883	234	0	0.382	0.408	459	1	0.530	0.500	684	0	0.487	0.440	909	0	0.355	0.257
10	1	0.405	0.233	235	0	0.335	0.233	460	1	0.577	0.733	685	0	0.465	0.533	910	0	0.367	0.433
11	0	0.475	0.408	236	1	0.615	0.813	461	1	0.565	0.500	686	0	0.457	0.515	911	0	0.367	0.433
12	0	0.580	0.533	237	1	0.592	0.708	462	1	0.565	0.500	687	0	0.382	0.280	912	1	0.600	0.675
13	1	0.895	0.883	238	1	0.615	0.813	463	1	0.553	0.617	688	0	0.347	0.233	913	1	0.544	0.640
14	1	0.738	0.708	239	1	0.592	0.708	464	1	0.483	0.500	689	0	0.393	0.467	914	1	0.530	0.517
15	0	0.465	0.533	240	1	0.580	0.673	465	1	0.572	0.640	690	0	0.347	0.350	915	1	0.530	0.500
16	1	0.773	0.708	241	1	0.592	0.708	466	1	0.577	0.733	691	0	0.557	0.533	916	1	0.553	0.617
17	1	0.662	0.767	242	1	0.545	0.533	467	0	0.330	0.400	692	0	0.347	0.233	917	1	0.516	0.533
18	1	0.545	0.533	243	1	0.592	0.708	468	0	0.273	0.200	693	0	0.557	0.533	918	1	0.530	0.557
19	1	0.629	0.533	244	1	0.545	0.533	469	0	0.355	0.257	694	0	0.510	0.533	919	1	0.601	0.533
20	1	0.650	0.580	245	1	0.592	0.708	470	0	0.367	0.433	695	1	0.557	0.533	920	1	0.685	0.720
21	1	0.662	0.767	246	1	0.580	0.673	471	0	0.285	0.200	696	1	0.510	0.533	921	1	0.650	0.533
22	1	0.808	0.883	247	0	0.395	0.708	472	1	0.565	0.500	697	1	0.557	0.533	922	1	0.615	0.533
23	1	0.505	0.475	248	1	0.615	0.673	473	1	0.530	0.592	698	1	0.557	0.533	923	1	0.615	0.533
24	1	0.552	0.553	249	1	0.615	0.813	474	1	0.530	0.500	699	1	0.627	0.720	924	1	0.615	0.533
25	1	0.622	0.615	250	1	0.475	0.533	475	1	0.483	0.500	700	1	0.557	0.615	925	1	0.685	0.708
26	1	0.587	0.615	251	1	0.545	0.533	476	1	0.577	0.733	701	1	0.662	0.720	926	1	0.685	0.708
27	0	0.140	0.165	252	1	0.587	0.615	477	1	0.553	0.617	702	1	0.592	0.708	927	1	0.615	0.533
28	0	0.230	0.165	253	1	0.475	0.533	478	1	0.565	0.500	703	1	0.557	0.615	928	1	0.650	0.708
29	0	0.277	0.311	254	1	0.580	0.533	479	1	0.565	0.675	704	1	0.571	0.557	929	1	0.615	0.533
30	0	0.360	0.553	255	1	0.545	0.633	480	1	0.565	0.697	705	1	0.599	0.673	930	0	0.500	0.580
31	0	0.308	0.563	256	1	0.545	0.673	481	1	0.577	0.733	706	1	0.662	0.883	931	0	0.447	0.373
32	0	0.308	0.563	257	1	0.475	0.533	482	1	0.650	0.860	707	1	0.592	0.533	932	0	0.405	0.280
33	0	0.300	0.368	258	1	0.587	0.767	483	1	0.580	0.720	708	0	0.477	0.813	933	0	0.405	0.233
34	0	0.342	0.315	259	1	0.592	0.708	484	1	0.580	0.580	709	0	0.417	0.513	934	0	0.419	0.257
35	0	0.342	0.175	260	1	0.615	0.673	485	1	0.594	0.697	710	0	0.405	0.455	935	0	0.475	0.408
36	0	0.342	0.315	261	1	0.580	0.650	486	1	0.622	0.697	711	0	0.417	0.513	936	0	0.568	0.533
37	0	0.342	0.175	262	1	0.531	0.650	487	1	0.622	0.813	712	0	0.405	0.455	937	0	0.568	0.533
38	1	0.505	0.475	263	1	0.568	0.633	488	1	0.650	0.708	713	0	0.405	0.455	938	0	0.405	0.233
39	1	0.552	0.553	264	1	0.545	0.708	489	1	0.615	0.720	714	0	0.370	0.315	939	0	0.500	0.580
40	1	0.622	0.615	265	0	0.372	0.633	490	1	0.615	0.720	715	0	0.370	0.350	940	1	0.601	0.533
41	1	0.587	0.615	266	0	0.395	0.673	491	1	0.580	0.580	716	0	0.370	0.280	941	1	0.685	0.720
42	0	0.358	0.292	267	0	0.300	0.233	492	1	0.622	0.697	717	0	0.370	0.280	942	1	0.650	0.533
43	0	0.358	0.333	268	0	0.395	0.533	493	0	0.465	0.580	718	0	0.370	0.280	943	1	0.615	0.533
44	0	0.382	0.408	269	0	0.348	0.533	494	0	0.370	0.455	719	0	0.440	0.455	944	1	0.615	0.533
45	0	0.335	0.233	270	0	0.335	0.233	495	0	0.440	0.420	720	0	0.440	0.560	945	1	0.615	0.533
46	0	0.335	0.233	271	1	0.475	0.533	496	0	0.370	0.280	721	0	0.405	0.455	946	1	0.685	0.708
47	0	0.370	0.373	272	1	0.545	0.633	497	0	0.370	0.280	722	0	0.440	0.455	947	1	0.685	0.708
48	0	0.288	0.233	273	1	0.545	0.533	498	0	0.335	0.280	723	0	0.417	0.513	948	1	0.615	0.533
49	0	0.335	0.233	274	1	0.580	0.533	499	0	0.440	0.420	724	0	0.335	0.315	949	1	0.650	0.708
50	0	0.382	0.408	275	1	0.475	0.533	500	0	0.444	0.580	725	0	0.417	0.513	950	1	0.615	0.533
51	0	0.358	0.292	276	1	0.545	0.673	501	0	0.450	0.760	726	0	0.370	0.280	951	1	0.580	0.720
52	0	0.288	0.233	277	0	0.395	0.533	502	0	0.422	0.597	727	0	0.412	0.420	952	1	0.545	0.592
53	0	0.358	0.292	278	1	0.615	0.813	503	0	0.370	0.280	728	0	0.335	0.280	953	1	0.510	0.615
54	1	0.568	0.633	279	1	0.475	0.533	504	0	0.440	0.560	729	0	0.370	0.280	954	1	0.440	0.440
55	1	0.592	0.708	280	1	0.615	0.673	505	0	0.393	0.380	730	0	0.370	0.280	955	1	0.510	0.580
56	1	0.545	0.533	281	1	0.531	0.650	506	0	0.440	0.513	731	0	0.405	0.455	956	1	0.510	0.440
57	1	0.580	0.673	282	1	0.580	0.575	507	0	0.417	0.467	732	0	0.477	0.813	957	1	0.463	0.440
58	1	0.580	0.533	283	1	0.552	0.650	508	0	0.370	0.280	733	0	0.405	0.455	958	1	0.552	0.557
59	1	0.559	0.533	284	1	0.510	0.417	509	1	0.622	0.813	734	0	0.417	0.513	959	1	0.545	0.592
60	1	0.557	0.673	285	1	0.557	0.633	510	1	0.580	0.720	735	0	0.405	0.455	960	1	0.580	0.592
61	1	0.545	0.533	286	1	0.552	0.650	511	1	0.615	0.720	736	0	0.405	0.455	961	0	0.395	0.440
62	1	0.587	0.650	287	1	0.533	0.517	512	1	0.622	0.813	737	0	0.440	0.455	962	0	0.314	0.257
63	1	0.587	0.767	288	1	0.557	0.633	513	1	0.580	0.580	738	1	0.557	0.533	963	0	0.545	0.592
64	1	0.545	0.533	289	1	0.553	0.617	514	1	0.594	0.697	739	1	0.627	0.720	964	0	0.300	0.140
65	1	0.592	0.708	290	1	0.557	0.633	515	1	0.650	0.708	740	1	0.557	0.580	965	0	0.347	0.350
66	1	0.592	0.708	291	1	0.510	0.417	516	1	0.622	0.697	741	1	0.557	0.533	966	0	0.347	0.350

TABLE 7. *Continued.*

No.	Hum.	Ours t=5	Ack. R=0.4	No.	Hum.	Ours t=5	Ack. R=0.4	No.	Hum.	Ours t=5	Ack. R=0.4	No.	Hum.	Ours t=5	Ack. R=0.4	No.	Hum.	Ours t=5	Ack. R=0.4
67	1	0.615	0.673	292	1	0.568	0.633	517	0	0.465	0.580	742	1	0.603	0.767	967	0	0.300	0.140
68	1	0.545	0.533	293	1	0.510	0.557	518	1	0.615	0.720	743	1	0.580	0.650	968	0	0.300	0.140
69	0	0.377	0.350	294	0	0.248	0.175	519	0	0.587	0.767	744	1	0.557	0.580	969	0	0.300	0.140
70	0	0.545	0.673	295	0	0.283	0.175	520	0	0.533	0.708	745	1	0.662	0.883	970	0	0.395	0.440
71	0	0.377	0.350	296	1	0.580	0.575	521	0	0.540	0.673	746	1	0.543	0.557	971	1	0.580	0.720
72	0	0.580	0.533	297	1	0.552	0.650	522	0	0.545	0.767	747	1	0.627	0.708	972	1	0.545	0.592
73	1	0.559	0.533	298	1	0.510	0.417	523	0	0.498	0.533	748	1	0.749	0.883	973	1	0.510	0.615
74	1	0.557	0.673	299	1	0.557	0.633	524	0	0.545	0.767	749	1	0.627	0.708	974	1	0.440	0.440
75	1	0.545	0.533	300	1	0.552	0.650	525	0	0.540	0.673	750	0	0.417	0.373	975	1	0.510	0.580
76	1	0.587	0.650	301	1	0.533	0.517	526	0	0.540	0.673	751	0	0.417	0.408	976	1	0.510	0.440
77	1	0.587	0.767	302	1	0.524	0.475	527	0	0.498	0.533	752	0	0.417	0.420	977	1	0.463	0.440
78	1	0.545	0.533	303	1	0.571	0.673	528	0	0.383	0.708	753	0	0.389	0.373	978	1	0.552	0.557
79	1	0.592	0.708	304	1	0.717	0.790	529	0	0.302	0.533	754	0	0.347	0.233	979	1	0.545	0.592
80	1	0.592	0.708	305	1	0.682	0.615	530	0	0.463	0.533	755	0	0.347	0.233	980	1	0.580	0.592
81	1	0.615	0.673	306	1	0.647	0.615	531	0	0.358	0.513	756	0	0.347	0.233	981	1	0.533	0.440
82	1	0.594	0.615	307	1	0.804	0.790	532	0	0.358	0.408	757	0	0.453	0.650	982	1	0.580	0.650
83	1	0.524	0.475	308	1	0.559	0.615	533	0	0.323	0.408	758	1	0.557	0.533	983	0	0.270	0.233
84	1	0.524	0.475	309	1	0.454	0.475	534	0	0.288	0.233	759	1	0.627	0.720	984	0	0.314	0.175
85	1	0.594	0.615	310	1	0.524	0.650	535	0	0.288	0.233	760	1	0.557	0.533	985	0	0.216	0.000
86	1	0.594	0.615	311	1	0.524	0.615	536	0	0.288	0.233	761	1	0.603	0.767	986	0	0.270	0.233
87	1	0.524	0.615	312	1	0.524	0.615	537	0	0.288	0.233	762	1	0.580	0.650	987	0	0.370	0.350
88	1	0.559	0.475	313	1	0.559	0.615	538	0	0.383	0.708	763	1	0.557	0.580	988	0	0.365	0.257
89	1	0.524	0.650	314	1	0.580	0.673	539	0	0.302	0.533	764	1	0.662	0.883	989	0	0.393	0.292
90	1	0.545	0.615	315	1	0.314	0.175	540	0	0.587	0.767	765	1	0.650	0.755	990	0	0.323	0.140
91	1	0.594	0.615	316	1	0.279	0.175	541	0	0.533	0.708	766	1	0.615	0.615	991	0	0.393	0.292
92	0	0.314	0.175	317	1	0.559	0.615	542	0	0.540	0.673	767	1	0.580	0.615	992	0	0.370	0.350
93	0	0.314	0.175	318	1	0.804	0.790	543	0	0.545	0.767	768	1	0.580	0.580	993	0	0.358	0.292
94	0	0.356	0.292	319	1	0.647	0.615	544	0	0.498	0.533	769	1	0.580	0.580	994	0	0.370	0.350
95	0	0.267	0.175	320	1	0.682	0.615	545	1	0.650	0.720	770	1	0.580	0.580	995	0	0.370	0.350
96	0	0.349	0.315	321	1	0.571	0.673	546	1	0.678	0.755	771	1	0.615	0.755	996	0	0.337	0.257
97	0	0.489	0.475	322	1	0.524	0.475	547	0	0.468	0.455	772	1	0.510	0.580	997	1	0.627	0.813
98	1	0.594	0.615	323	1	0.717	0.790	548	0	0.468	0.455	773	1	0.627	0.813	998	1	0.566	0.592
99	1	0.524	0.475	324	1	0.524	0.475	549	0	0.314	0.292	774	1	0.650	0.755	999	1	0.650	0.755
100	1	0.524	0.475	325	1	0.571	0.673	550	0	0.314	0.280	775	1	0.580	0.580	1000	1	0.622	0.720
101	1	0.594	0.615	326	1	0.524	0.475	551	0	0.265	0.140	776	1	0.566	0.580	1001	1	0.627	0.813
102	1	0.594	0.615	327	1	0.571	0.673	552	0	0.377	0.350	777	1	0.566	0.580	1002	1	0.580	0.580
103	1	0.524	0.615	328	1	0.571	0.673	553	0	0.465	0.490	778	1	0.580	0.580	1003	1	0.580	0.580
104	1	0.559	0.475	329	1	0.571	0.673	554	0	0.319	0.373	779	1	0.650	0.860	1004	1	0.533	0.580
105	1	0.524	0.650	330	1	0.477	0.475	555	0	0.319	0.373	780	1	0.580	0.580	1005	1	0.533	0.580
106	1	0.545	0.615	331	1	0.524	0.475	556	0	0.342	0.315	781	0	0.405	0.455	1006	1	0.580	0.580
107	1	0.594	0.615	332	1	0.571	0.673	557	0	0.405	0.373	782	0	0.412	0.420	1007	0	0.335	0.280
108	1	0.440	0.388	333	0	0.349	0.315	558	0	0.251	0.175	783	0	0.417	0.397	1008	0	0.370	0.280
109	1	0.482	0.440	334	0	0.335	0.315	559	0	0.223	0.175	784	0	0.393	0.397	1009	0	0.405	0.455
110	1	0.493	0.475	335	0	0.384	0.455	560	0	0.573	0.580	785	0	0.323	0.280	1010	0	0.335	0.280
111	1	0.423	0.300	336	0	0.349	0.315	561	0	0.440	0.420	786	0	0.417	0.513	1011	1	0.627	0.813
112	0	0.240	0.365	337	0	0.405	0.408	562	0	0.433	0.455	787	1	0.650	0.755	1012	1	0.566	0.592
113	0	0.277	0.311	338	1	0.524	0.475	563	0	0.503	0.630	788	1	0.615	0.615	1013	1	0.650	0.755
114	0	0.272	0.218	339	1	0.524	0.475	564	0	0.468	0.455	789	1	0.580	0.615	1014	1	0.622	0.720
115	0	0.248	0.263	340	1	0.571	0.673	565	0	0.468	0.455	790	1	0.580	0.580	1015	1	0.627	0.813
116	0	0.300	0.368	341	0	0.314	0.175	566	0	0.398	0.350	791	1	0.580	0.580	1016	1	0.580	0.580
117	0	0.342	0.175	342	1	0.571	0.673	567	0	0.328	0.280	792	1	0.580	0.580	1017	0	0.347	0.233
118	0	0.402	0.650	343	1	0.524	0.475	568	1	0.524	0.475	793	1	0.615	0.755	1018	0	0.347	0.233
119	0	0.402	0.475	344	1	0.594	0.615	569	1	0.571	0.673	794	1	0.510	0.580	1019	0	0.361	0.257
120	0	0.575	0.575	345	1	0.568	0.697	570	1	0.615	0.755	795	1	0.627	0.813	1020	0	0.389	0.373
121	0	0.342	0.175	346	1	0.568	0.720	571	1	0.559	0.615	796	1	0.650	0.755	1021	0	0.417	0.408
122	0	0.587	0.475	347	1	0.493	0.580	572	1	0.594	0.615	797	1	0.580	0.580	1022	0	0.126	0.000
123	0	0.440	0.388	348	1	0.463	0.580	573	1	0.538	0.592	798	1	0.580	0.580	1023	0	0.216	0.000
124	0	0.482	0.440	349	1	0.603	0.860	574	1	0.717	0.790	799	1	0.580	0.580	1024	0	0.342	0.175
125	0	0.493	0.475	350	1	0.568	0.720	575	1	0.524	0.475	800	1	0.510	0.580	1025	0	0.342	0.175
126	0	0.423	0.300	351	1	0.580	0.860	576	1	0.571	0.673	801	1	0.603	0.697	1026	0	0.216	0.000
127	0	0.560	0.615	352	1	0.568	0.720	577	1	0.559	0.615	802	1	0.650	0.755	1027	0	0.300	0.292
128	0	0.538	0.580	353	1	0.638	0.767	578	1	0.615	0.580	803	1	0.510	0.580	1028	0	0.265	0.117
129	0	0.622	0.755	354	1	0.638	0.860	579	1	0.627	0.813	804	1	0.580	0.580	1029	0	0.265	0.140
130	0	0.622	0.755	355	1	0.568	0.720	580	1	0.566	0.580	805	1	0.580	0.580	1030	0	0.307	0.257
131	0	0.710	0.615	356	1	0.568	0.720	581	1	0.622	0.720	806	1	0.622	0.720	1031	0	0.286	0.280
132	0	0.599	0.673	357	1	0.533	0.720	582	0	0.158	0.200	807	1	0.615	0.755	1032	0	0.220	0.140
133	0	0.599	0.673	358	1	0.603	0.720	583	0	0.158	0.200	808	0	0.417	0.513	1033	0	0.255	0.140
134	0	0.745	0.790	359	1	0.533	0.720	584	0	0.212	0.200	809	0	0.545	0.580	1034	0	0.477	0.720

TABLE 7. Continued.

No.	Hum.	Ours t=5	Ack. R=0.4	No.	Hum.	Ours t=5	Ack. R=0.4	No.	Hum.	Ours t=5	Ack. R=0.4	No.	Hum.	Ours t=5	Ack. R=0.4	No.	Hum.	Ours t=5	Ack. R=0.4
135	0	0.622	0.755	360	1	0.603	0.720	585	0	0.356	0.280	810	0	0.533	0.580	1035	0	0.389	0.373
136	0	0.745	0.790	361	1	0.493	0.580	586	0	0.267	0.175	811	0	0.533	0.580	1036	0	0.347	0.233
137	0	0.552	0.475	362	0	0.500	0.755	587	0	0.356	0.280	812	0	0.370	0.280	1037	0	0.347	0.233
138	0	0.622	0.755	363	0	0.533	0.580	588	0	0.267	0.175	813	0	0.417	0.513	1038	0	0.257	0.233
139	0	0.482	0.475	364	0	0.383	0.580	589	0	0.267	0.175	814	1	0.580	0.580	1039	0	0.347	0.233
140	0	0.552	0.475	365	0	0.430	0.650	590	0	0.328	0.292	815	1	0.510	0.580	1040	0	0.417	0.408
141	0	0.412	0.440	366	1	0.463	0.580	591	0	0.314	0.315	816	1	0.622	0.720	1041	0	0.417	0.408
142	0	0.465	0.440	367	1	0.580	0.860	592	0	0.300	0.292	817	1	0.622	0.720	1042	0	0.499	0.557
143	0	0.496	0.475	368	1	0.533	0.755	593	0	0.314	0.175	818	1	0.510	0.580	1043	0	0.520	0.580
144	0	0.465	0.440	369	1	0.638	0.767	594	0	0.361	0.373	819	1	0.650	0.755	1044	0	0.415	0.440
145	0	0.475	0.533	370	1	0.568	0.720	595	0	0.314	0.175	820	1	0.615	0.755	1045	0	0.417	0.408
146	0	0.500	0.615	371	1	0.638	0.860	596	0	0.384	0.420	821	1	0.580	0.580	1046	0	0.205	0.140
147	0	0.465	0.440	372	1	0.568	0.720	597	0	0.412	0.420	822	0	0.263	0.400	1047	0	0.377	0.567
148	0	0.412	0.440	373	1	0.493	0.580	598	0	0.417	0.513	823	1	0.460	0.517	1048	0	0.358	0.640
149	0	0.552	0.720	374	1	0.568	0.697	599	0	0.440	0.455	824	1	0.502	0.540	1049	0	0.377	0.673
150	0	0.290	0.331	375	1	0.568	0.720	600	0	0.393	0.420	825	1	0.530	0.575	1050	0	0.438	0.440
151	0	0.272	0.156	376	1	0.568	0.720	601	0	0.370	0.280	826	1	0.507	0.633	1051	0	0.499	0.557
152	0	0.333	0.373	377	1	0.603	0.860	602	0	0.370	0.280	827	1	0.502	0.540	1052	0	0.520	0.580
153	0	0.496	0.456	378	1	0.544	0.533	603	0	0.383	0.580	828	1	0.460	0.440	1053	0	0.415	0.440
154	0	0.412	0.440	379	1	0.600	0.675	604	0	0.477	0.813	829	1	0.502	0.540	1054	0	0.527	0.608
155	0	0.465	0.440	380	1	0.530	0.500	605	0	0.417	0.513	830	1	0.460	0.400	1055	1	0.580	0.673
156	0	0.496	0.475	381	1	0.565	0.675	606	0	0.356	0.280	831	1	0.460	0.400	1056	1	0.592	0.733
157	1	0.580	0.592	382	1	0.530	0.500	607	0	0.267	0.175	832	0	0.263	0.400	1057	1	0.530	0.533
158	1	0.580	0.720	383	1	0.635	0.850	608	0	0.356	0.280	833	0	0.250	0.140	1058	1	0.600	0.813
159	1	0.545	0.580	384	1	0.530	0.500	609	0	0.267	0.175	834	0	0.277	0.240	1059	1	0.627	0.767
160	1	0.668	0.592	385	1	0.530	0.592	610	1	0.524	0.475	835	0	0.320	0.275	1060	1	0.703	0.708
161	1	0.510	0.440	386	1	0.565	0.675	611	1	0.571	0.673	836	0	0.250	0.100	1061	1	0.773	0.883
162	1	0.552	0.557	387	1	0.600	0.697	612	1	0.615	0.755	837	1	0.502	0.540	1062	1	0.860	0.883
163	0	0.395	0.440	388	1	0.565	0.557	613	1	0.559	0.615	838	0	0.263	0.400	1063	1	0.475	0.533
164	0	0.370	0.292	389	1	0.577	0.733	614	1	0.538	0.592	839	1	0.530	0.575	1064	1	0.545	0.708
165	0	0.300	0.140	390	1	0.530	0.500	615	1	0.717	0.790	840	1	0.460	0.475	1065	0	0.430	0.650
166	0	0.347	0.350	391	0	0.320	0.257	616	1	0.580	0.580	841	1	0.460	0.517	1066	0	0.288	0.233
167	0	0.277	0.140	392	0	0.355	0.375	617	1	0.627	0.813	842	1	0.502	0.540	1067	0	0.300	0.373
168	0	0.265	0.140	393	0	0.285	0.200	618	1	0.580	0.615	843	1	0.507	0.633	1068	0	0.382	0.408
169	0	0.335	0.292	394	0	0.285	0.200	619	1	0.773	0.930	844	0	0.440	0.513	1069	0	0.335	0.408
170	1	0.440	0.440	395	0	0.334	0.200	620	1	0.580	0.580	845	0	0.440	0.420	1070	0	0.405	0.513
171	1	0.552	0.557	396	0	0.355	0.257	621	1	0.603	0.697	846	0	0.310	0.455	1071	0	0.345	0.433
172	1	0.545	0.580	397	1	0.568	0.592	622	1	0.615	0.755	847	0	0.415	0.583	1072	0	0.288	0.233
173	0	0.300	0.140	398	1	0.580	0.650	623	1	0.650	0.755	848	0	0.370	0.280	1073	0	0.382	0.408
174	0	0.347	0.350	399	1	0.533	0.615	624	1	0.510	0.580	849	0	0.412	0.397	1074	0	0.335	0.233
175	0	0.347	0.350	400	1	0.463	0.440	625	1	0.650	0.755	850	0	0.370	0.280	1075	1	0.580	0.673
176	0	0.580	0.720	401	1	0.568	0.592	626	1	0.615	0.580	851	0	0.450	0.620	1076	1	0.592	0.733
177	0	0.668	0.592	402	1	0.533	0.440	627	0	0.430	0.580	852	0	0.412	0.397	1077	1	0.530	0.533
178	0	0.395	0.440	403	1	0.533	0.440	628	0	0.405	0.455	853	0	0.440	0.420	1078	1	0.600	0.813
179	1	0.580	0.592	404	1	0.533	0.540	629	0	0.405	0.455	854	0	0.370	0.280	1079	1	0.627	0.767
180	1	0.514	0.475	405	1	0.603	0.720	630	0	0.290	0.315	855	0	0.440	0.513	1080	1	0.703	0.708
181	1	0.570	0.615	406	1	0.603	0.592	631	0	0.255	0.140	856	0	0.380	0.480	1081	1	0.773	0.883
182	1	0.574	0.708	407	1	0.568	0.580	632	0	0.405	0.280	857	0	0.345	0.480	1082	1	0.860	0.883
183	1	0.528	0.475	408	1	0.533	0.440	633	0	0.465	0.580	858	0	0.440	0.513	1083	1	0.475	0.533
184	1	0.528	0.475	409	0	0.522	0.592	634	0	0.440	0.455	859	0	0.440	0.420	1084	1	0.545	0.708
185	1	0.563	0.650	410	0	0.452	0.475	635	0	0.440	0.455	860	0	0.310	0.455	1085	0	0.266	0.275
186	1	0.720	0.825	411	0	0.494	0.557	636	0	0.417	0.513	861	0	0.415	0.583	1086	0	0.239	0.100
187	1	0.528	0.475	412	0	0.228	0.117	637	0	0.627	0.813	862	0	0.370	0.280	1087	0	0.254	0.240
188	1	0.528	0.563	413	0	0.242	0.117	638	0	0.580	0.580	863	1	0.615	0.580	1088	0	0.223	0.100
189	1	0.808	0.825	414	0	0.522	0.592	639	0	0.430	0.580	864	1	0.627	0.720	1089	0	0.298	0.228
190	1	0.598	0.650	415	0	0.522	0.592	640	1	0.773	0.930	865	1	0.510	0.580	1090	0	0.282	0.240
191	1	0.563	0.540	416	0	0.494	0.557	641	1	0.580	0.580	866	0	0.323	0.280	1091	0	0.239	0.100
192	1	0.650	0.650	417	1	0.524	0.475	642	1	0.615	0.755	867	0	0.370	0.280	1092	0	0.243	0.370
193	1	0.720	0.825	418	1	0.524	0.475	643	1	0.603	0.697	868	0	0.248	0.000	1093	0	0.358	0.528
194	1	0.640	0.615	419	1	0.524	0.475	644	0	0.528	0.455	869	0	0.213	0.000	1094	0	0.402	0.400
195	1	0.575	0.708	420	1	0.559	0.615	645	0	0.370	0.315	870	0	0.160	0.000	1095	0	0.449	0.400
196	0	0.353	0.350	421	1	0.477	0.475	646	0	0.650	0.755	871	0	0.417	0.420	1096	0	0.223	0.100
197	0	0.413	0.650	422	1	0.454	0.475	647	0	0.430	0.315	872	0	0.384	0.280	1097	0	0.414	0.400
198	1	0.528	0.475	423	1	0.538	0.475	648	0	0.283	0.175	873	0	0.487	0.300	1098	0	0.239	0.100
199	1	0.720	0.825	424	1	0.571	0.673	649	0	0.342	0.315	874	0	0.458	0.300	1099	0	0.266	0.275
200	1	0.808	0.825	425	1	0.559	0.475	650	0	0.358	0.420	875	0	0.405	0.300	1100	1	0.568	0.633
201	0	0.401	0.592	426	1	0.559	0.615	651	0	0.230	0.175	876	0	0.493	0.475	1101	1	0.592	0.708
202	1	0.650	0.650	427	1	0.566	0.580	652	0	0.440	0.455	877	0	0.504	0.408	1102	1	0.545	0.533

TABLE 7. ***Continued.***

No.	Hum.	Ours t=5	Ack. R=0.4	No.	Hum.	Ours t=5	Ack. R=0.4	No.	Hum.	Ours t=5	Ack. R=0.4	No.	Hum.	Ours t=5	Ack. R=0.4	No.	Hum.	Ours t=5	Ack. R=0.4
203	1	0.551	0.592	428	1	0.524	0.475	653	0	0.370	0.280	878	0	0.382	0.408	1103	1	0.592	0.708
204	1	0.574	0.708	429	0	0.374	0.475	654	0	0.323	0.280	879	0	0.417	0.513	1104	1	0.545	0.533
205	1	0.580	0.673	430	0	0.361	0.373	655	0	0.370	0.315	880	0	0.370	0.280	1105	1	0.592	0.708
206	1	0.615	0.813	431	1	0.566	0.592	656	0	0.370	0.315	881	0	0.580	0.580	1106	1	0.587	0.650
207	1	0.592	0.708	432	0	0.314	0.175	657	0	0.528	0.455	882	0	0.580	0.580	1107	1	0.580	0.673
208	1	0.592	0.708	433	1	0.524	0.475	658	0	0.370	0.315	883	0	0.545	0.580	1108	1	0.545	0.533
209	1	0.580	0.673	434	1	0.524	0.475	659	0	0.356	0.280	884	0	0.370	0.397	1109	1	0.592	0.708
210	1	0.592	0.708	435	1	0.524	0.475	660	0	0.384	0.420	885	1	0.615	0.580	1110	0	0.348	0.533
211	1	0.615	0.813	436	1	0.559	0.615	661	0	0.417	0.513	886	1	0.627	0.720	1111	0	0.349	0.350
212	1	0.615	0.673	437	1	0.477	0.475	662	0	0.405	0.455	887	1	0.510	0.580	1112	0	0.335	0.408
213	1	0.592	0.708	438	1	0.454	0.475	663	0	0.528	0.455	888	1	0.565	0.675	1113	0	0.335	0.233
214	1	0.545	0.533	439	1	0.538	0.475	664	0	0.370	0.315	889	1	0.516	0.533	1114	0	0.370	0.373
215	1	0.592	0.708	440	1	0.571	0.673	665	1	0.557	0.533	890	1	0.600	0.675	1115	0	0.335	0.373
216	1	0.531	0.650	441	1	0.559	0.475	666	1	0.510	0.533	891	1	0.572	0.640	1116	0	0.348	0.533
217	1	0.545	0.533	442	0	0.374	0.475	667	1	0.557	0.533	892	0	0.380	0.500	1117	1	0.568	0.633
218	1	0.603	0.767	443	1	0.559	0.615	668	1	0.557	0.533	893	0	0.335	0.233	1118	1	0.592	0.708
219	1	0.568	0.733	444	1	0.566	0.580	669	1	0.627	0.720	894	0	0.288	0.233	1119	1	0.545	0.533
220	1	0.545	0.533	445	1	0.524	0.475	670	1	0.557	0.615	895	0	0.335	0.233	1120	1	0.592	0.708
221	1	0.545	0.533	446	1	0.522	0.592	671	1	0.662	0.720	896	0	0.335	0.233	1121	1	0.545	0.533
222	1	0.498	0.533	447	1	0.522	0.592	672	1	0.592	0.708	897	0	0.300	0.233	1122	1	0.592	0.708
223	1	0.545	0.708	448	1	0.494	0.557	673	1	0.557	0.615	898	0	0.565	0.675	1123	1	0.587	0.650
224	1	0.545	0.533	449	0	0.381	0.650	674	1	0.571	0.557	899	0	0.516	0.533	1124	1	0.580	0.673
225	0	0.395	0.708	450	0	0.242	0.117	675	1	0.599	0.673	900	0	0.600	0.675	1125	1	0.545	0.533
226	0	0.288	0.233	451	0	0.242	0.117	676	1	0.662	0.883	901	0	0.572	0.640	1126	1	0.592	0.708

REFERENCES

[1] G. Majumder, P. Pakray, A. Gelbukh, and D. Pinto, "Semantic textual similarity methods, tools, and applications: A survey," *Computación y Sistemas*, vol. 20, no. 4, pp. 647–665.

[2] M. W. Berry and M. Castellanos, "Survey of text mining," *Comput. Rev.*, vol. 45, no. 9, p. 548, 2004.

[3] A. Singhal, "Modern information retrieval: A brief overview," *IEEE Comput. Soc. Tech. Committee Data Eng.*, vol. 24, no. 4, pp. 35–43, Jan. 2001.

[4] Y.-S. Lin, J.-Y. Jiang, and S.-J. Lee, "A similarity measure for text classification and clustering," *IEEE Trans. Knowl. Data Eng.*, vol. 26, no. 7, pp. 1575–1590, Jul. 2014.

[5] A. L. Fred and A. K. Jain, "Data clustering using evidence accumulation," in *Proc. 16th Int. Conf. Pattern Recognit. (ICPR)*, vol. 4, 2002, pp. 276–280.

[6] H. Park, K. Kim, S. Choi, and J. Yoon, "A patent intelligence system for strategic technology planning," *Expert Syst. Appl.*, vol. 40, no. 7, pp. 2373–2390, Jun. 2013.

[7] X. Wang, H. Ren, Y. Chen, Y. Liu, Y. Qiao, and Y. Huang, "Measuring patent similarity with SAO semantic analysis," *Scientometrics*, vol. 121, no. 1, pp. 1–23, Oct. 2019.

[8] L. R. Dice, "Measures of the amount of ecologic association between species," *Ecology*, vol. 26, no. 3, pp. 297–302, Jul. 1945.

[9] T. Sørensen, "A method of establishing groups of equal amplitude in plant sociology based on similarity of species content and its application to analyses of the vegetation on Danish commons," *Biologiske Skrifter*, vol. 5, pp. 1–34, Jun. 1948.

[10] R. Mihalcea, C. Corley, and C. Strapparava, "Corpus-based and knowledge-based measures of text semantic similarity," in *Proc. 21st Nat. Conf. Artif. Intell. 18th Innov. Appl. Artif. Intell. Conf.*, 2006, pp. 775–780.

[11] M. A. Sultan, S. Bethard, and T. Sumner, "DLS@CU: Sentence similarity from word alignment and semantic vector composition," in *Proc. 9th Int. Workshop Semantic Eval. (SemEval)*, 2015, pp. 148–153.

[12] J. D. Choi and M. Palmer, "Fast and robust part-of-speech tagging using dynamic model selection," in *Proc. 50th Annu. Meeting Assoc. Comput. Linguistics, Short Papers*, vol. 2, 2012, pp. 363–367.

[13] F. Mandreoli, R. Martoglia, and P. Tiberio, "A syntactic approach for searching similarities within sentences," in *Proc. 11th Int. Conf. Inf. Knowl. Manage. (CIKM)*, 2002, pp. 635–637.

[14] Y. Li, D. McLean, Z. A. Bandar, J. D. O'Shea, and K. Crockett, "Sentence similarity based on semantic nets and corpus statistics," *IEEE Trans. Knowl. Data Eng.*, vol. 18, no. 8, pp. 1138–1150, Aug. 2006.

[15] G. Salton and C. Buckley, "Term-weighting approaches in automatic text retrieval," *Inf. Process. Manage.*, vol. 24, no. 5, pp. 513–523, Jan. 1988.

[16] R. M. Aliguliyev, "A new sentence similarity measure and sentence based extractive technique for automatic text summarization," *Expert Syst. Appl.*, vol. 36, no. 4, pp. 7764–7772, May 2009.

[17] R. Ferreira, G. D. C. Cavalcanti, F. Freitas, R. D. Lins, S. J. Simske, and M. Riss, "Combining sentence similarities measures to identify paraphrases," *Comput. Speech Lang.*, vol. 47, pp. 59–73, Jan. 2018.

[18] B. Hassan, S. E. Abdelrahman, R. Bahgat, and I. Farag, "UESTS: An unsupervised ensemble semantic textual similarity method," *IEEE Access*, vol. 7, pp. 85462–85482, 2019.

[19] Z. Quan, Z.-J. Wang, Y. Le, B. Yao, K. Li, and J. Yin, "An efficient framework for sentence similarity modeling," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 27, no. 4, pp. 853–865, Apr. 2019.

[20] S. Zhou, X. Xu, Y. Liu, R. Chang, and Y. Xiao, "Text similarity measurement of semantic cognition based on word vector distance decentralization with clustering analysis," *IEEE Access*, vol. 7, pp. 107247–107258, 2019.

[21] Y. Zhao, S. Gao, P. Gallinari, and J. Guo, "Knowledge base completion by learning pairwise-interaction differentiated embeddings," *Data Mining Knowl. Discovery*, vol. 29, no. 5, pp. 1486–1504, Sep. 2015.

[22] P. Velardi, P. Fabriani, and M. Missikoff, "Using text processing techniques to automatically enrich a domain ontology," in *Proc. Int. Conf. Formal Ontology Inf. Syst. (FOIS)*, 2001, pp. 270–284.

[23] C. Yang, D. Zhu, and X. Wang, "SAO semantic information identification for text mining," *Int. J. Comput. Intell. Syst.*, vol. 10, no. 1, p. 593, 2017.

[24] S. Choi, H. Kim, J. Yoon, K. Kim, and J. Y. Lee, "An SAO-based text-mining approach for technology roadmapping using patent information," *R&D Manage.*, vol. 43, no. 1, pp. 52–74, Jan. 2013.

[25] J. Yoon, H. Park, and K. Kim, "Identifying technological competition trends for R&D planning using dynamic patent maps: SAO-based content analysis," *Scientometrics*, vol. 94, no. 1, pp. 313–331, Jan. 2013.

[26] W. Ki and K. Kim, "Generating information relation matrix using semantic patent mining for technology planning: A case of nano-sensor," *IEEE Access*, vol. 5, pp. 26783–26797, 2017.

[27] A. Abbas, L. Zhang, and S. U. Khan, "A literature review on the state-of-the-art in patent analysis," *World Patent Inf.*, vol. 37, pp. 3–13, Jun. 2014.

[28] I. Bergmann, D. Butzke, L. Walter, J. P. Fuerste, M. G. Moehrle, and V. A. Erdmann, "Evaluating the risk of patent infringement by means of semantic patent analysis: The case of DNA chips," *R&D Manage.*, vol. 38, no. 5, pp. 550–562, 2008.

[29] H. Park, J. Yoon, and K. Kim, "Identifying patent infringement using SAO based semantic technological similarities," *Scientometrics*, vol. 90, no. 2, pp. 515–529, Feb. 2012.

[30] G. Cascini and M. Zini, "Measuring patent similarity by comparing inventions functional trees," in *Proc. 2nd Top. Session Comput.-Aided Innov. Comput.-Aided Innov. (CAI) IFIP 20th World Comput. Congr. (WG)*, vol. 277. Springer, 2008, pp. 31–42.

[31] I. Park and B. Yoon, "A semantic analysis approach for identifying patent infringement based on a product–patent map," *Technol. Anal. Strategic Manage.*, vol. 26, no. 8, pp. 855–874, Sep. 2014.

[32] J. Yoon and K. Kim, "Identifying rapidly evolving technological trends for R&D planning using SAO-based semantic patent networks," *Scientometrics*, vol. 88, no. 1, pp. 213–228, Jul. 2011.

[33] J. Yoon and K. Kim, "Detecting signals of new technological opportunities using semantic patent analysis and outlier detection," *Scientometrics*, vol. 90, no. 2, pp. 445–461, Feb. 2012.

[34] H. Park, J. J. Ree, and K. Kim, "An SAO-based approach to patent evaluation using TRIZ evolution trends," in *Proc. IEEE Int. Conf. Manage. Innov. Technol. (ICMIT)*, Jun. 2012, pp. 594–598.

[35] S. Choi, J. Yoon, K. Kim, J. Y. Lee, and C.-H. Kim, "SAO network analysis of patents for technology trends identification: A case study of polymer electrolyte membrane technology in proton exchange membrane fuel cells," *Scientometrics*, vol. 88, no. 3, pp. 863–883, Sep. 2011.

[36] C. Yang, D. Zhu, X. Wang, Y. Zhang, G. Zhang, and J. Lu, "Requirement-oriented core technological components' identification based on SAO analysis," *Scientometrics*, vol. 112, no. 3, pp. 1229–1248, Sep. 2017.

[37] S. Choi, H. Park, D. Kang, J. Y. Lee, and K. Kim, "An SAO-based text mining approach to building a technology tree for technology planning," *Expert Syst. Appl.*, vol. 39, no. 13, pp. 11443–11455, Oct. 2012.

[38] C. Sternitzke and I. Bergmann, "Similarity measures for document mapping: A comparative study on the level of an individual scientist," *Scientometrics*, vol. 78, no. 1, pp. 113–130, Jan. 2009.

[39] T. K. Landauer, P. W. Foltz, and D. Laham, "An introduction to latent semantic analysis," *Discourse Process.*, vol. 25, nos. 2–3, pp. 259–284, 1998.

[40] P. D. Turney, "Mining the Web for synonyms: PMI-IR versus LSA on TOEFL," in *Proc. Mach. Learn. (ECML)*. Springer, 2001, pp. 491–502.

[41] E. Gabrilovich and S. Markovitch, "Computing semantic relatedness using wikipedia-based explicit semantic analysis," *IJCAI*, vol. 7, pp. 1606–1611, Jun. 2007.

[42] R. L. Cilibrasi and P. M. B. Vitanyi, "The Google similarity distance," *IEEE Trans. Knowl. Data Eng.*, vol. 19, no. 3, pp. 370–383, Mar. 2007.

[43] C. Leacock and M. Chodorow, "Combining local context and wordnet similarity for word sense identification," *WordNet, An Electron. Lexical Database*, vol. 49, no. 2, pp. 265–283, 1998.

[44] Z. Wu and M. Palmer, "Verbs semantics and lexical selection," in *Proc. 32nd Annu. Meeting Assoc. Comput. Linguistics*, 1994, pp. 133–138.

[45] Y. Bin, L. Xiao-Ran, L. Ning, and Y. Yue-Song, "Using information content to evaluate semantic similarity on HowNet," in *Proc. 8th Int. Conf. Comput. Intell. Secur.*, vol. 1. San Mateo, CA, USA: Morgan Kaufmann, Nov. 2012, pp. 448–453.

[46] J. Jiang and D. Conrath, "Semantic similarity based on corpus statistics and lexical taxonomy," in *Proc. 10th Res. Comput. Linguistics Int. Conf.*, 1997, pp. 19–33.

[47] D. Lin, "An information-theoretic definition of similarity," in *Proc. 15th Int. Conf. Mach. Learn. (ICML)*. San Mateo, CA, USA: Morgan Kaufmann, 1998, pp. 296–304.

[48] X. Liu, Y. Zhou, and R. Zheng, "Sentence similarity based on dynamic time warping," in *Proc. Int. Conf. Semantic Comput. (ICSC)*, Sep. 2007, pp. 250–256.

[49] Y. Zhang, X. Zhou, A. L. Porter, and J. M. V. Gomila, "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*, vol. 101, no. 2, pp. 1375–1389, 2014.

[50] G. Angeli, M. J. Johnson Premkumar, and C. D. Manning, "Leveraging linguistic structure for open domain information extraction," in *Proc. 53rd Annu. Meeting Assoc. Comput. Linguistics 7th Int. Joint Conf. Natural Lang. Process.*, 2015, pp. 344–354.

[51] K. W. Boyack and R. Klavans, "Measuring science–technology interaction using rare inventor–author names," *J. Informetrics*, vol. 2, no. 3, pp. 173–182, Jul. 2008.

[52] A. Rodriguez, B. Kim, M. Turkoz, J.-M. Lee, B.-Y. Coh, and M. K. Jeong, "New multi-stage similarity measure for calculation of pairwise patent similarity in a patent citation network," *Scientometrics*, vol. 103, no. 2, pp. 565–581, May 2015.

[53] B. Yoon, "On the development of a technology intelligence tool for identifying technology opportunity," *Expert Syst. Appl.*, vol. 35, nos. 1–2, pp. 124–135, Jul. 2008.

[54] S. Arts, B. Cassiman, and J. C. Gomez, "Text matching to measure patent similarity," *Strategic Manage. J.*, vol. 39, no. 1, pp. 62–84, Jan. 2018.

[55] A. Budanitsky and G. Hirst, "Semantic distance in WordNet: An experimental, application-oriented evaluation of five measures," in *Proc. Workshop WordNet Other Lexical Resour.*, vol. 2, 2001, p. 2.

[56] G. A. Miller, "WordNet: A lexical database for english," *Commun. ACM*, vol. 38, no. 11, pp. 39–41, Nov. 1995, doi: 10.1145/219717.219748.

[57] G. K. Mazandu and N. J. Mulder, "Information content-based gene ontology semantic similarity approaches: Toward a unified framework theory," *BioMed Res. Int.*, vol. 2013, no. 292063, pp. 1–11, 2013.

[58] H. W. Kuhn, "The Hungarian method for the assignment problem," *Nav. Res. Logistics Quart.*, vol. 2, nos. 1–2, pp. 83–97, 1955.

[59] R. Prasad, A. Bhattacharyya, and Q. D. Nguyen, "Nanotechnology in sustainable agriculture: Recent developments, challenges, and perspectives," *Frontiers Microbiol.*, vol. 8, p. 1014, Jun. 2017.

XIAOMAN LI received the bachelor's degree in information management and information system from the University of Electronic Science and Technology of China, in 2015. She is currently pursuing the degree with the Agricultural Information Institute, Chinese Academy of Agricultural Sciences. Her research interests include patent analysis, technology evolution and prediction, text mining, and natural language processing.

CUI WANG received the master's degree in computer software and theory from Shandong University, in 2011. She is currently pursuing the degree with the Agricultural Information Institute, Chinese Academy of Agricultural Sciences. She is also a Lecturer with the School of Information Science and Engineering, Shandong Agriculture and Engineering University. Her research interests include rating systems, agricultural data analysis, and machine learning.

XUEFU ZHANG received the Ph.D. degree from the Chinese Academy of Sciences, in 2006. He is currently a Researcher with the Agricultural Information Institute, Chinese Academy of Agricultural Sciences. His main research interests include information visualization, knowledge organization and retrieval, strategic information research, and patent technology mining.

WEI SUN received the Ph.D. degree from the Chinese Academy of Sciences, in 2010. She is currently an Associate Researcher with the Agricultural Information Institute, Chinese Academy of Agricultural Sciences. Her main research interests include information visualization and patent technology mining.