A Novel Approach for Patent Similarity Measurement Based on Sequence Alignment

Xin An[1]
School of Economics & Management
Beijing Forestry University
Beijing, P.R. China
anxin@bjfu.edu.cn

Jinghong Li[2]
School of Economics & Management
Beijing Forestry University
Beijing, P.R. China
724298617@qq.com

Shuo Xu[3*]
Research Base of Beijing Modern Manufacturing Development, College of Economics and Management
Beijing University of Technology
Beijing, P.R. China
xushuo@bjut.edu.cn

Liang Chen[4]
Institute of Scientific and Technical Information of China
Beijing, P.R. China
25565853@qq.com

Sainan Pi[5]
School of Economics & Management
Beijing Forestry University
Beijing, P.R. China
silencepipi@bjfu.edu.cn

* Corresponding author

ABSTRACT

Patent similarity measurement, as one of fundamental building blocks for patent analysis, not only can derive technical intelligence efficiently, but also can detect the risk of infringement and evaluate whether the invention meets the criteria of novelty and innovation. However, traditional approaches make implicitly several assumptions, such as bag of words in each component, semantic direction irrelevance and so on. In order to relax these assumptions, this study proposes a novel approach on the basis of sequence alignment, which takes semantic direction of each sequence structure and the word order information of each component into consideration. Meanwhile, an algorithm for calculating the global importance of each sequence structure is put forward. Finally, to verify the effectiveness and performance of the improved semantic analysis, a case study is conducted on the *thin film head* subfield in the field of *hard disk drive*. Extensive experimental results show that our approach is significantly more accurate and is not sensitive to several core parameters.

KEYWORDS

Patent similarity measurement, Semantic analysis, Entities and semantic relations, Sequence alignment

1 Introduction

According to many surveys of authorities, patents cover more than 90% latest technical information of the world, of which 80% would not be published in other forms [5]. Thus, patents analysis is increasingly vital for mining technical intelligence. Patent similarity measurement, as one of fundamental building blocks for patent analysis, not only can derive technical intelligence efficiently, but also can detect the risk of infringement and evaluate whether the invention meets the criteria of novelty and innovation [13].

Nowadays, Subject-Action-Object (SAO) semantic analysis [2, 4, 9, 17] is the most widely used method to measure patent similarity, which stresses the key concepts and functional relations. By function, it means "the action changing a feature of any object" [18]. That is to say, SAO structure explicitly describes a relation between components in the patent documents. However, on closer examination, one can see that traditional SAO semantic analysis [2, 4, 9, 17] has several shortcomings. First, the semantic direction of each SAO structure and the word order in each component of a SAO structure are not taken into account. Second, intuitively, each SAO structure carries different amount of domain-specific information. To say it in another way, the importance of each SAO structure should be different [13]. But the SAO semantic analysis usually assigns equal weight to each SAO structure. Last but not least, the SAO semantic analysis only focuses on the functional relations, but ignores the valuable technology intelligence underlying in the non-functional relations which is based on the prepositions [1].

In order to overcome these issues, this article proposes an improved semantic analysis approach for assessing patent similarity on the basis of sequence alignment. Different from previous studies, the sequence structures are used in this paper. A sequence structure can be explained as an "Entity$^{(1)}$ – Relation – Entity$^{(2)}$" sequence. This type of structure embraces the functional and non-functional relations. For example, the phrases, "*…the seed film acting as a stop layer…*" and "*…planar layers on opposing sides of a pole piece…*", reflecting the *form* and *spatial* relation respectively, will generate two sequence structures as "*[seed film]* $^{(E)}$ – *form*$^{(R)}$ – *[stop layer]* $^{(E)}$" and "*[planar layers]*$^{(E)}$ – *spatial*$^{(R)}$ – *[pole piece]*$^{(E)}$". It is worth mentioning that the "sequence" emphasizes two aspects in this study: the semantic direction of these functional and non-functional structures and the word order of each entity. Meanwhile, an algorithm for calculating the global importance of each sequence structure is put forward.

2 Related Work

Before delving into more specifies, discussion of the literature pertinent to patent similarity measurement is in order.

2.1 Patent Similarity Measurement based on SAO structures

Some researchers utilized SAO structures based on semantic similarity to evaluate the risk of patent infringement [2, 9], identified the evolving technological trend for R&D planning [17], build a technology tree for technology planning [4] and so on. But in these approaches, each SAO structure is assigned the same weight. As an improvement, Wang et al. [13] has constructed a DWSAO indicator through assigning different weights to SAO structures for measuring patent similarity. However, it neglects the influence of the number of SAO structures of patents, which may result in the phenomenon that patents with high similarity values are actually not similar. Besides, it is not a symmetrical indicator.

In addition, previous methods implicitly omit the word order information of each component in a SAO structure. As we all know, the meaning of a phrase may be varied when the words are permutated. For example, the phrases "car gasoline" and "gasoline car" both consist of the same words but in different orders. The former is a kind of fuels while the latter is one kind of cars, so they should not be seen as the same thing.

Finally, just as An et al. [1] mentioned, the SAO analysis only focuses on functional relations between the components, but ignores the valuable technology intelligence in the form of non-functional relations. They proposed an approach based on preposition semantic network where prepositions aid to revealing the relations between keywords related to technologies and applied it to mine intelligence information in the patents. Thus, prepositional semantic analysis can be viewed to be complementary to SAO semantic analysis. This study integrates functional and non-functional relations, which are collectively referred to sequence structures.

2.2 WordNet

In order to calculate lexical semantic similarity, WordNet is usually chosen as the source of word relations. WordNet is a lexical database which groups English concepts into sets of synonyms called "synsets" and constructs the hierarchical structure to connect "synsets" by means of hypernym/hyponym relations. Just because of this property, WordNet is commonly used to calculate the semantic similarity of concepts. In this paper, the information-content (IC) based approach is used, which measures semantic similarity between concepts based on the notion of IC that is calculated in accordance to the probability of encountering a concept [6, 7, 12]. The IC-based approach can be formally defined as follows [6]:

$$sim(c_1, c_2) = \frac{2 \times IC(LCS)}{IC(c_1) + IC(c_2)} \quad (1)$$

Here, $sim(c_1, c_2)$ is the similarity between two concepts c_1 and c_2. *LCS* is the *Least Common Subsumer* (hypernym) of two concepts, and *IC* represents the information content value of the concepts.

Note that a word may express different meaning (concept) in different context, viz. polysemy. This paper uses the concepts corresponding to the highest similarity between two words. At length, given that the synset of $word_1$ and $word_2$ in WordNet is Syn_1 and Syn_2 respectively, the similarity of two words can be defined as follows.

$$sim(word_1, word_2) = \max_{c_i \in Syn_1} \max_{c_j \in Syn_2} sim(c_i, c_j) \quad (2)$$

3 Methodology

As shown in Figure 1, our research framework consists of four phases. The first is to extract sequence structures (functional and non-functional semantic relations) from patent documents through natural language processing (NLP) techniques and tools. At the second phase, the similarity between sequence structures is measured, which takes semantic direction of each sequence structure and the word order information of each component into consideration. The third phase is to calculate the global importance of each sequence structure based on the *TV_LinkA* algorithm [16]. Finally, the similarity between patents is assessed with the well-known optimal transportation problem solver [10, 14]. These phases are described in more details in the following subsections.

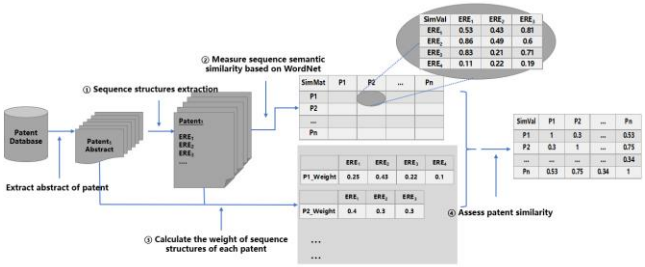

Figure 1: The overall procedure for measuring patent similarity.

3.1 Sequence structures extraction

Recently, Chen et al. [3] have proposed a promising patent information extraction framework, where two deep-learning models are respectively used for entity identification and semantic relation extraction. This framework can be used here to extract the sequence structures mentioned in the patent documents. For more elaborate and detailed descriptions, we refer the readers to Chen et al. [3].

3.2 Similarity between sequence structures

After extracting sequence structures, each patent can be represented by a collection of different number of sequence structures. In this way, patent similarity calculation problem can be transformed to compute the similarity between the collections of sequence structures. Before this, this subsection illustrates how to calculate the semantic similarity between two sequence structures, as shown in Figure 2. Since each sequence structure

consists of three components: $E^{(1)}$ (Entity$^{(1)}$), R (relation) and $E^{(2)}$ (Entity$^{(2)}$), the key is how to align the components from different structures and even the words in each component.

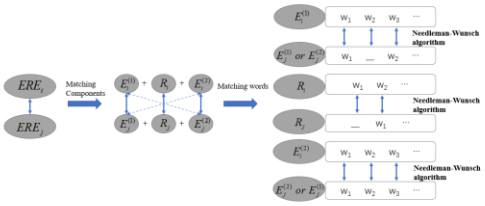

Figure 2: The overall procedure for calculating the similarity between sequence structures.

As for the alignment of components, we argue that the semantic direction between $E^{(1)}$ and $E^{(2)}$, which can be judged by R (relation), is also very important. According to the relation types, we can define corresponding semantic directions to match the components. In this study, the similarity between two sequence structures is defined as the average of the similarity of the matched components as follows.

$$sim(ERE_i, ERE_j) = \begin{cases} \frac{sim(E_i^{(1)},E_j^{(1)})+sim(R_i,R_j)+sim(E_i^{(2)},E_j^{(2)})}{3}, E_i^{(1)} \text{ matches } E_j^{(1)} \text{ and } E_i^{(2)} \text{ matches } E_j^{(2)} \\ \frac{sim(E_i^{(1)},E_j^{(2)})+sim(R_i,R_j)+sim(E_i^{(2)},E_j^{(1)})}{3}, E_i^{(1)} \text{ matches } E_j^{(2)} \text{ and } E_i^{(2)} \text{ matches } E_j^{(1)} \end{cases} \quad (3)$$

Here, $sim(ERE_i, ERE_j)$, ranging from 0 to 1, represents the similarity between ERE_i and ERE_j. The larger this index is, the greater the similarity between the sequence structures is. $sim(E_i^{(1)},E_j^{(1)})$, $sim(R_i,R_j)$, $sim(E_i^{(2)},E_j^{(2)})$, $sim(E_i^{(1)},E_j^{(2)})$ and $sim(E_i^{(2)},E_j^{(1)})$ denote the similarity between the matched components of ERE_i and ERE_j.

Of course, there exist undirected and bidirectional relations. As for these two case, we cannot assert whether $E^{(1)}$ of one sequence structure matches with $E^{(1)}$ or $E^{(2)}$ of another. In this situation, Eq. (4) is used to calculate the similarity between two sequence structures.

$$sim(ERE_i, ERE_j) = max \begin{cases} \frac{sim(E_i^{(1)},E_j^{(1)})+sim(R_i,R_j)+sim(E_i^{(2)},E_j^{(2)})}{3} \\ \frac{sim(E_i^{(1)},E_j^{(2)})+sim(R_i,R_j)+sim(E_i^{(2)},E_j^{(1)})}{3} \end{cases} \quad (4)$$

Now there remains how to align the words in the matched component. Here, the Needleman-Wunsch algorithm [8, 15] is utilized here to construct the correspondences of words in the focused words. As a comparison, the method of Wang et al. [13] is considered, which adopts the alignment form of Cartesian product. That is, each word from one component is aligned to each word from another component. If the similarity between aligned words is greater than a threshold, these two words are deemed to be matched.

Example 1. Consider two entities "car gasoline" and "gasoline car". Following Wang et al. [13], one can generate the correspondences of words as shown in Figure 3 (a) and the similarity between these two entities is 1.0000. It is counter-intuitive for the two entities to have a high degree of similarity, since the former is a kind of fuels while the latter is one type of cars. In our opinion, main reason for counter-intuitive similarity is that Wang et al. [13] omits the word order information. Figure 3 (b)-(c) illustrates the alignment of words in the entities "car gasoline" and "gasoline car" based on our approach, in which the symbol "_" denotes a gap. When a word corresponds with "_", the resulting similarity is regarded as zero. Thus, the similarity between two entities is the average of the similarity of the aligned words, that is, 0.3333. Compared to Wang et al. [13], this result seems be more realistic and credible.

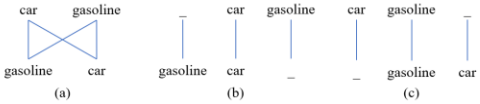

Figure 3: The correspondence of words in the entities "car gaosoline" and "gasoline car".

To more understandably show the procedure of calculating the similarity between two sequences, Figure 4 illustrates an example.

Example 2. One sequence structure is "*[insulating material]$^{(E)}$-partof$^{(R)}$-[planar layers]$^{(E)}$*" that means "insulating material" is a whole and "planar layers" is part of it, and another is "*[seed film]$^{(E)}$-form$^{(R)}$-[stop layer]$^{(E)}$*" that means "stop layer" is a whole or a product and "seed film" is part of it or the material making of it. We can define the semantic direction of the former as "*insulating material* ← *planar layers*", and the latter as "*seed film* → *stop layer*". Hence, "insulating material" and "stop layer" are the homogeneous components which can be considered to match, so do "planar layers" and "seed film". After matching the components, we use the Needleman-Wunsch algorithm to align words and then calculate the similarity between the aligned components. The similarity between two sequence structures is the average of the similarity of the aligned components.

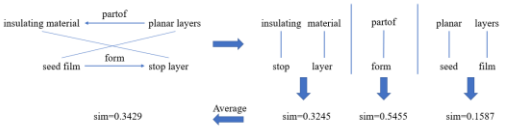

Figure 4: The procedure of calculating the similarity of example 2.

3.3 Weight estimation of sequence structures of each patent

Base on the concept that each sequence structure carries different amount of domain-specific information. This paper introduces a new method to calculate the global importance of each component of sequence structures based on *TV_LinkA* algorithm [16]. First, the network $\mathcal{G}(\mathcal{V},\mathcal{E})$ is constructed, where $\mathcal{V}$ is the set of nodes which consist of abstracts, sentences and components (entities and relations), and $\mathcal{E}$ is the set of edges. Each abstract links to the sentences which are original from it, and each sentence links to the components which are extracted from it. Second, the values of sentence and component nodes are preset to 1. Third, set the

appropriate number of iterations. For each iteration, the value of each component node is updated to the sum of the values of the sentence nodes connected to it and the updated values are standardized by the L2 norm. So does the value of each sentence node. Repeat the above steps to continuously update the value of the node until it is stable. At last, given that a terminology occurring a few times in domain-relevant sentences is more likely to be domain specific than another occurring many times in some general sentences, inverse document frequency (IDF) is multiplied the resulting value of each node.

After that, we can obtain the global importance of each component in all patent documents. Thus, the importance of each sequence structure is the average of the importance of the corresponding components. To let the weights lie from 0 to 1, for all the sequence structures in a same patent, the weights are normalized so that their summary is guaranteed to be equal to 1.

3.4 Patent similarity assessment

From the similarity matrix to the patent similarity, in order to make full use of all the information, patent similarity measurement problem can be transformed into the well-known optimal transportation problem [10, 14]. Just as Figure 5, the patent distance matrix, which can get from 1 minus patent similarity matrix, and the weight vectors are fed to an optimal transportation problem solver to obtain the shortest distance between two patents. The similarity of two interested patents is equal to 1 minus the shortest distance.

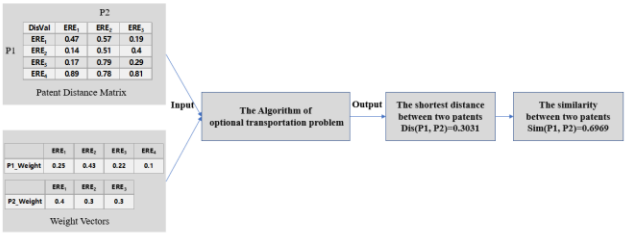

Figure 5: The procedure of calculating the similarity between two patents.

4 Case Study

4.1 Dataset

To evaluate the performance of our methodology, an annotated corpus[1] by [3] is used in this work. This dataset comes from *thin film head* subfield in the field of *hard disk drive*. It contains 1,010 patent documents. Note that, in this dataset, there are 84 pairs of patents coming from the same patent family. That is, each pair of patents both has the same abstract and the identical collection of sequence structures so that they should have higher similarity than others. These patents can be used to assess the effectiveness and performance of our method. If a method can better identify these 84 pairs of patents, its performance should be better.

Before comparing the sequence structures, we should judge the semantic direction in accordance to the type of semantic relations between the components $E^{(1)}$ and $E^{(2)}$ in a sequence structure so that they can correctly match to the $E^{(1)}$ and $E^{(2)}$ of another sequence structures. As shown in Table 1, we have defined 4 types of semantic directions. If the sequence structures are both single-direction, we can match the components $E^{(1)}$ and $E^{(2)}$ between two sequence structures and apply Eq. (3) to calculate the similarity, Eq. (4) otherwise.

Table 1: The semantic directions of each relation type.

	Relation Type	Semantic Direction
1	spatial relation	Undirected
2	part-of	$E^{(1)} \leftarrow E^{(2)}$
3	causative relation	$E^{(1)} \leftarrow E^{(2)}$
4	operation	$E^{(1)} \leftarrow E^{(2)}$
5	made-of	$E^{(1)} \leftarrow E^{(2)}$
6	instance-of	$E^{(1)} \rightarrow E^{(2)}$
7	attribution	$E^{(1)} \leftarrow E^{(2)}$
8	generate	$E^{(1)} \leftarrow E^{(2)}$
9	purpose	$E^{(1)} \leftarrow E^{(2)}$
10	in-manner-of	$E^{(1)} \leftarrow E^{(2)}$
11	alias	Bidirectional
12	formation	$E^{(1)} \rightarrow E^{(2)}$
13	comparison	Undirected
14	measurement	$E^{(1)} \leftarrow E^{(2)}$
15	others	Undirected

4.2 Experiment Setup

In this paper, we use WordNet as the source of word relations to calculate semantic similarity of words, but unfortunately, some words in the dataset are not included in WordNet. To solve this problem, we apply the "gestalt pattern matching" algorithm [11] as a supplement, which computes the similarity of two strings as the number of matching characters divided by the total number of characters in the two strings.

In our methodology, there are two parameters needed to be preset by user. The first one is the number of iterations when calculating the weight of each sequence structure, and the second one is the gap penalty in the Needleman-Wunsch algorithm.

As for the number of iterations, one can determine whether it is stable by observing the trend of the weights after several iterations. Through the experiment, we find that the weights of components gradually stabilize after 4 iterations. Thus, the number of iterations is fixed to 10 in this article.

As for the gap penalty, to assess its impact on patent similarity, we choose multiple values for comparison, such as -0.05, -0.1, -0.15, -0.2 and -0.3. But we find that no matter which value to choose, the word alignment, patent similarity matrix and patent similarity will not be affected. Hence, the gap penalty is set to -0.05 in this paper.

[1] https://github.com/awesome-patent-mining/TFH_Annotated_Dataset

4.3 Experimental results and discussions

To verify the effectiveness and performance of our approach, the result will be used to compare with the result of Wang et al. [13].

Figure 6 shows the results of these two approaches. Each patent is compared with other patents, Top 1 (@1), Top 2 (@2), Top 3 (@3), Top 4 (@4) and Top 5 (@5) highest similarity is chosen to form 5 collections and then to judge how many of 84 pairs of patents are covered. If we select Top 1 highest similarity of each patent, our method can obtain 54 pairs of patents that come from a patent family when the weights are determined by the weighting algorithm (section 3.3), while 58 pairs can be outputted by our approach with the same weights. But the DWSAO analysis can even recognize none of them. If Top 2 collection is considered, our weighted and non-weighted versions contain 70 pairs and 78 pairs respectively, while the DWSAO analysis only identifies 2 pairs. When we enlarge to Top 5 highest similarity of each patent, the weighted one can identify 81 pairs and the non-weighted one can fully recognize 84 pairs of patents while only 3 pairs are identified by the DWSAO analysis.

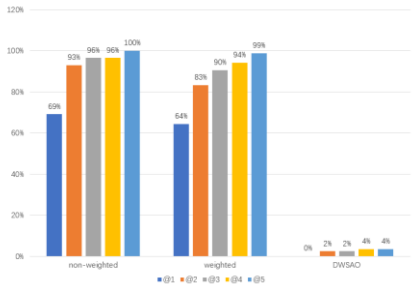

Figure 6: The performance of our approach and DWSAO method.

It is no doubt that our patent similarity measurement is significantly more accurate than the DWSAO approach. At the meanwhile, a significant advantage of the improved semantic analysis is that the results are not sensitive to several core parameters. But this method with different weights does not perform as well as the method with the equal importance. In our opinion, the main reason is that the weighting algorithm actually considers the importance of each sequence structure in the global context, not the importance in the local context (i.e., each patent). In the near future, a locally weighting method will be further investigated.

5 Conclusion

This study proposes an improved semantic analysis for assessing patent similarity on the basis of entities and semantic relations (functional and non-functional relations), which takes semantic direction of each sequence structure and the word order information of each component into consideration. Meanwhile, we introduce a new method to calculate the global importance of each sequence structure. To verify the effectiveness and performance of the improved semantic analysis, a case study on patent similarity measurement related to *thin film head* subfield in the field of *hard disk drive* was used. Extensive experimental results demonstrate that our patent similarity measurement is significantly more accurate. Meanwhile, a significant advantage of the improved semantic analysis is that the results are not sensitive to several core parameters. But this method with different weights does not perform as well as the method with the equal importance. In our opinion, the main reason is that this weighting process actually considers the importance of each sequence structure in the global context, not the importance in the local context (i.e., each patent). In the near future, a locally weighting method will be further investigated.

ACKNOWLEDGMENTS

This research received the financial support from the Social Science Foundation of Beijing Municipality under grant number 17GLB074, and Natural Science Foundation of Guangdong Province (Grant Number 2018A030313695).

REFERENCES

[1] An, J., Kim, K., Mortara, L., & Lee, S. (2018). Deriving technology intelligence from patents: Preposition-based semantic analysis. Journal of Informetrics, 12(1), 217-236. doi:10.1016/j.joi.2018.01.001

[2] 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.

[3] Chen, L., Xu, S., Zhu, L., Zhang, J., Lei,X., & Yang, G. (2020). A deep learning based method for extracting semantic information from patent documents. Scientometrics.

[4] Choi, S., Park, H., Kang, D., Lee, J. Y., & Kim, K. (2012). An SAO-based text mining approach to building a technology tree for technology planning. Expert Systems with Applications, 39(13), 11443-11455. doi:10.1016/j.eswa.2012.04.014

[5] Zha, X., & Chen, M. (2010). Study on early warning of competitive technical intelligence based on the patent map. Journal of Computers, 5(2). doi:10.4304/jcp.5.2.274-281

[6] Lin, D. (1998). An information-theoretic definition of similarity. In Proceedings of the 15th International Conference on Machine Learning, pp. 296-304.

[7] Jiang, J. J., & Conrath, D. W. (1997). Semantic Similarity Based on Corpus Statistics and Lexical Taxonomy. arXiv: Computation and Language.

[8] Needleman, S. B., & Wunsch, C. D. (1970). A general method applicable to the search for similarities in the amino acid sequence of two proteins. Journal of Molecular Biology, 48(3), 443-453.

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

[10] Rachev, S.T. (1998). In L. Ruschendorf (Ed.), Mass transportation problems: Volume I: Theory (probability and its applications). New York, NY: Springer.

[11] Ratcliff, J. W., & Metzener, D. E. (1988). Pattern-matching-the gestalt approach. Dr Dobbs Journal, 13(7), 46.

[12] Resnik, P. (1995). Using information content to evaluate semantic similarity in a taxonomy. International Joint Conference on Artificial Intelligence, 448-453.

[13] Wang, X., Ren, H., Chen, Y., Liu, Y., Qiao, Y., & Huang, Y. (2019). Measuring patent similarity with SAO semantic analysis. Scientometrics, 121(1), 1-23. doi:10.1007/s11192-019-03191-z

[14] Xu, S., Zhai, D., Wang, F., An, X., Pang, H., & Sun, Y. (2019). A novel method for topic linkages between scientific publications and patents. Journal of the Association for Information Science and Technology, 70(9), 1026-1042. doi:10.1002/asi.24175

[15] Xu, S., Zhu, L., Qiao, X., & Xue, C. (2009). A novel approach for measuring chinese terms semantic similarity based on pairwise sequence alignment. In Proceedings of the 5th International Conference on Semantics, Knowledge and Grid (pp. 92-98). IEEE.

[16] Yang, Y., Lu, Q., & Zhao, T. (2010). A delimiter-based general approach for Chinese term extraction. Journal of the American Society for Information Science and Technology, 61(1), 111-125. doi:10.1002/asi.21221

[17] Yoon, J., & Kim, K. (2011). Identifying rapidly evolving technological trends for R&D planning using SAO-based semantic patent networks. Scientometrics, 88(1), 213-228. doi:10.1007/s11192-011-0383-0

[18] Savransky, S.D. (2000) Engineering of creativity: Introduction to TRIZ methodology of inventive problem solving. Boca Raton, FL: CRC Press.