IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received December 22, 2021, accepted January 5, 2022, date of publication January 7, 2022, date of current version January 20, 2022.

*Digital Object Identifier 10.1109/ACCESS.2022.3141494*

# A Fast and Scalable Algorithm for Prior Art Search

**JUHYUN LEE[1], SANGSUNG PARK[2], AND JUNSEOK LEE[3]**
[1]Department of Industrial Management Engineering, Korea University, Seoul 02841, Republic of Korea
[2]Department of Big Data Statistics, Cheongju University, Cheongju 28503, Republic of Korea
[3]Micube Solution Inc., Seoul 06719, Republic of Korea

Corresponding author: Junseok Lee (js.lee@micube.co.kr)

This work was supported by Brain Korea 21 FOUR.

**ABSTRACT** Companies and research institutes are acquiring monopolized technologies through patents to survive in a rapidly changing market. The applicant is guaranteed by law the exclusive right to the technology for a certain period. A prior art search is an essential task for this purpose. Therefore, the need for a successful prior art search is increasing. Previous studies have disadvantages because relevant fields are limited, requiring an in-depth understanding of specific technologies. This paper proposes a method to overcome these limitations using patent text and citation information. The proposed method searches for prior art using a citation network based on the shortest path derived from a breadth-first search. We examined based on experiments whether the proposed method could be applied to the industry. Based on the results, the proposed algorithm can efficiently search for prior art. Furthermore, it was possible to construct a counterstrategy based on the characteristics of the applicant. The code will be made available at github.com/lee-ju/k_step_PIC.

**INDEX TERMS** Prior art search, graph traversal, citation network, document similarity, patent analysis.

## I. INTRODUCTION

Patents give the inventor the right to exclude others from making, using, or selling technology [1]. Furthermore, patents are open to the public to promote industrial development. As diverse industries grow, patents create value for intelligent, autonomous, and interconnected technologies [2], [3]. Companies and researchers conduct research and development (R&D) of new technologies and register them as patents to keep pace with the rapidly changing technology market.

Recently, as technology has rapidly developed, the need to improve the patent registration process has increased [4]. A patent is registered through a complex procedure—from a specification prepared by an attorney to examination by an examiner. Examiners search for prior art to evaluate the novelty, right, and inventive steps of the technology. Prior art is a document related to the novelty and inventive steps claimed by the applicant. They cite prior art to improve the quality and patentability of the technology [5].

Applicants, attorneys, and examiners should enhance the completeness of patents using a prior art search (PAS) [6], [7]. When applicants conceal prior art, it prevents examiners from rejecting applications, but the patent tends to be invalidated by the courts [8]. Furthermore, examiners may represent true knowledge spillovers by citing the prior art omitted by applicants [9]. Accordingly, PAS is a crucial task in the patent registration process.

Nevertheless, examiners have spent significant time processing one patent, which causes many problems [10]–[12]. Efficient PAS simplifies the registration process and promotes industrial innovation by increasing patent quality [13]. The difficulties of PAS are as follows:

- PAS is a complex, cumbersome task even for a well-trained examiner, and different results may be obtained for the same patent [14], [15]. Therefore, the new PAS method must produce efficient and consistent results.
- As new technologies proliferate, examiners have an increasing amount of literature and knowledge to consider, delaying reviews [16]. The improved PAS should be a general-purpose method not limited to a specific technology.

An inadequate PAS leads to unnecessary patent litigation [17]. Therefore, a novel method that considers the factors behind legal disputes should be designed. A successful PAS affects applicants in several ways. First, applicants can increase technology competitiveness through PAS [18].

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

Applicants and attorneys examine prior art to determine the superiority of their technology.

Second, the examiner decides whether to register the patent accordingly and search for the existence of prior art. If no prior art is similar to a new patent, the technology satisfies novelty. Furthermore, the examiner requests an advanced patent by notifying the applicant of the prior art. The applicant then modifies the patent to satisfy the inventive step. PAS helps to increase the quality, registration possibility, and competitiveness of patents.

Second, PAS can enhance the market competitiveness of applicants [19]. In business management, a company's decision-making on ''maintenance or sale'' and ''manufacture or purchase'' of R&D resources is crucial [20]. Companies must develop sustainable technologies to increase their competitiveness. Furthermore, patents function as catalysts for innovation. Companies can sustain their business through patents. Consequently, PAS was the first of these processes.

Third, PAS can lower the risk of patent litigation [21]–[25]. Companies need to recognize the importance of the PAS because they can be protected from external threats from technological competition [26]. As contention intensifies, patent litigation is widely used as a strategic weapon [27]. Under United States patent law, if defendants are found to infringe on plaintiffs' rights, courts can issue orders to compel them to stop manufacturing and selling products [28], [29]. Because these consequences can be fatal to companies and cause huge losses, it is imperative to prepare for them [30]–[32].

PAS is critical for increasing the competitiveness of applicants in the technology and market and lowering the risk of patent litigation. Thus, the PAS algorithm has several objectives:

- We propose an algorithm to search for prior art that is similar but has no citation relationship.
- The proposed method minimizes the parameters that the examiner is involved in to derive consistent results quickly.
- Because the algorithm uses natural language processing (NLP), it can be universally applied to new technologies.
- The novel method may also explore prior art that the applicant, attorney, and examiner were unaware of.
- The algorithm's output is not limited to the patent level but can be extended to the applicant level, allowing researchers to use the analysis results for diverse purposes.

We propose an algorithm to overcome the shortcomings of the existing method and increase its utility. The remainder of this paper is organized as follows. Section 2 provides related work on PAS. In Section 3, we describe the proposed PAS method. Section 4 describes an experiment to verify the applicability of the proposed method. Finally, conclusions are presented in Section 5. The code of our algorithm is available on the GitHub page: github.com/lee-ju/k_step_PIC.

## II. RELATED WORK

Shalaby and Zadrozny [33] proposed a taxonomy of three patent retrieval methods. The first is a keyword-based method. Prior art can be searched using queries in a patent database. Studies based on this method attempt to increase the efficiency of the PAS by expanding or reducing the query.

The second method exploits the unique features of patents. When a patent is filed, an international patent classification (IPC) code is assigned by an examiner. Furthermore, the patent cites prior art and is cited by other patents. The metadata-based method uses an IPC code or F-term, which morphologically divides patents for an efficient search.

The last method uses the semantic information of patents through text mining or NLP. This method searches for prior art based on document similarity.

### A. KEYWORD-BASED APPROACHES

A query is a combination of keywords used when searching a patent database. Researchers select relevant keywords based on their knowledge and combine them to create a query. A query using too few keywords searches for an extensive range of patents. The results of the analysis based on this search may contain errors. Conversely, a query that uses too many keywords searches for a very narrow range of patents. Thus, the analysis of these results may be biased.

PAS requires an appropriate range of queries. Jones [34] argued that visualizing query aids with an efficient search. This method formats the query by structuring the data to be retrieved in a Venn diagram. Anick *et al.* [35] and Ressell-Rose *et al.* [36] proposed a query-based tool for accurate, repeatable, and transparent PAS. Furthermore, Tiwana and Horowitz [37] conducted a study using keywords used in queries to search for similar patents in a database.

Although the keyword-based approach is a traditional method from previous studies, it has two limitations. First, it relies on queries. This method may have different prior art collected depending on the examiner. Second, it is essential to understand the target technology for selecting the keywords constituting the query. Therefore, work efficiency may be low because the examiner has to spend significant time investigating the prior art of the new technology.

### B. METADATA-BASED APPROACHES

Patents are classified into chemical, mechanical, electrical, and other categories according to the technology field. PAS can also be used as a code for patent classification. Harris *et al.* [38] proposed a method for investigating prior art using the hierarchical structure of IPC. PAS using a patent classification code such as IPC has the advantage that consistent results can be derived. However, this method has a disadvantage because it is challenging to investigate new technical fields. Another limitation is that when two or more technologies are combined, the meaning of the patent classification code may become ambiguous. In this case, it was difficult to determine the scope of the prior art.

Citations are among the relevant metadata for patents. Citations related to legal rights can be used for the PAS because they reflect the quality of a patent [39]. Furthermore, forward citations, although often classified as innovative or discontinuous concerning prior art, include the history of technical prior art [40]. Couteau [41] argued that prior art could be found by combining forward-searching and citation analysis. Furthermore, Von Wartburg *et al.* [42] proposed a citation network using family patents as weights to search for similar patents. They emphasized that this method could measure the cumulative inventive progress of technology.

Co-citation refers to different prior art being cited simultaneously by a later registered patent [43]. Previous studies have scored the relationship between co-citations to examine prior art [44], [45]. Other studies have followed the knowledge transfer of technology by co-citation [46], [47]. However, these methods are difficult to apply to new technical fields and have limitations in not reflecting the flow of citations—because they include the development trend of technology. Previous studies used only forward or backward patents. Therefore, improved PAS methods need to reflect the flow of citations.

### C. SEMANTIC-BASED APPROACHES

A patent is a document composed of titles, abstracts, and claims. The similarity between patents can be measured using a computer using NLP. Recently, as algorithms for NLP have advanced, many studies have investigated prior art using semantic information. No *et al.* [48] and Nakamura *et al.* [49] proposed a matrix-based method that merges co-citation information and co-word frequency. They pointed out the problem of using either metadata or semantic information when judging the similarity of a patent. Nevertheless, their methods are difficult to apply to co-words used in new technologies.

Sanguri *et al.* [50] proposed a latent semantic analysis method to address the problems of dictionary-based semantic approaches. They extracted the latent information from the text and calculated the similarity of the document along with the co-citation relationship. Furthermore, Cammarano *et al.* [51] suggested a correlation between the citation relationship and a document's latent semantics. Korobkin *et al.* [52] developed multiple pre-trained latent semantic models and obtained topic distribution of patents for PAS. Their method made it possible to apply deep syntactic relations and to calculate semantic similarities from large amounts of text data.

A methodology for scientifically verifying the use of a patent was recently proposed to measure the similarity and novelty of a patent. Arts *et al.* [53] introduced a novel similarity measure based on the United States Patent Classification System. They demonstrated through large-scale experiments in the entire patent population that text-matched patents were more likely to be cited alongside each other in the same document. Korobkin *et al.* [54] argued that a patent is relevant if the key feature set extracted from the patent contains all keywords based on chemical effect descriptions. They also introduced a coefficient of relevance to rank sets of related patents and developed a method for extracting chemical effect descriptions and technical features from patent databases.

PAS is also related to the specialization of examiners. Righi and Simcoe [55] focused on results that depended on examiner specialization when examining prior art. They found a positive correlation between examiner specialization and a more rigorous screening process. Furthermore, they measured the similarity of the technology and examiners based on the co-words of the patents and found that a patent can be delegated to an examiner of the prior art. It can then be registered quickly through professional review.

Arts *et al.* [56] proposed an NLP technique that can identify, based on the content of patent documents, the development and impact of new and subsequent technologies. Their method has been verified to replace the existing matrix by experimenting with patents with different examination results depending on the country. They argued that such a patent lacks novelty due to its high similarity with the prior art. Borodin *et al.* [57] introduced a process for identifying the similarity of a solution to the problem a patent is trying to solve. They assigned similarity coefficients for context synonyms using structural parsing and an embedding model based on neural network.

In previous studies, algorithms developed for PAS used keywords, metadata, and semantic information. As the use of NLP increases, the need for a semantic-based approach, rather than keywords, increases. Semantic-based approaches can mitigate the limitations of algorithms using keywords. Therefore, the improved PAS algorithm should use semantic information and metadata together. Thus, the improved method is more efficient than before. Accordingly, this paper proposes a PAS algorithm that can use both sources of information.

## III. PRIOR ART SEARCH

Let $\mathrm{DATE_i}$ be the registration date of $\mathrm{p_i}$, which is the i-th patent among n registered patents, and FC ($\mathrm{p_i}$) be the set of patents citing $\mathrm{p_i}$. When the cited network is $\mathrm{G_c} = (\mathrm{V_c}, \mathrm{E_c})$, $\mathrm{V_c}$ is a patent, and $\mathrm{E_c}$ is a citation relationship. For example, in this study, $\mathrm{p_j} \in \mathrm{FC}\ (\mathrm{p_i})$ for two patents $\mathrm{p_i}$, $\mathrm{p_j} \in \mathrm{V_c}$ with $\mathrm{i} \neq \mathrm{j}$, $\mathrm{p_j}$ cites $\mathrm{p_i}$, expressed as $\mathrm{e}(\mathrm{p_i}, \mathrm{p_j}) \in \mathrm{E_c}$. Furthermore, when the network representing the similarity of patents is $\mathrm{G_s} = (\mathrm{V_s}, \mathrm{E_s})$, $\mathrm{e_s}(\mathrm{p_i}, \mathrm{p_j}) \in \mathrm{E_s}$ indicates that $\mathrm{p_i}$ and $\mathrm{p_j}$ satisfy (1).

$$SIM\left(f\left(p_i\right), f\left(p_j\right)\right) \geq SIM_{min} \tag{1}$$

Let $f: p_i \rightarrow \mathbb{R}^d$ be a function that maps the natural language to an embedding vector that reflects the frequency or context. SIM is a function that calculates the degree of similarity of documents mapped to vectors. If the similarity between the two documents is greater than the threshold $\mathrm{SIM_{min}}$, they are linked to $\mathrm{G_s}$. For example, if SIM is a cosine similarity, it is calculated as in (2), and the range of values is $-1$ to 1. The closer the cosine similarity is to 1, the stronger

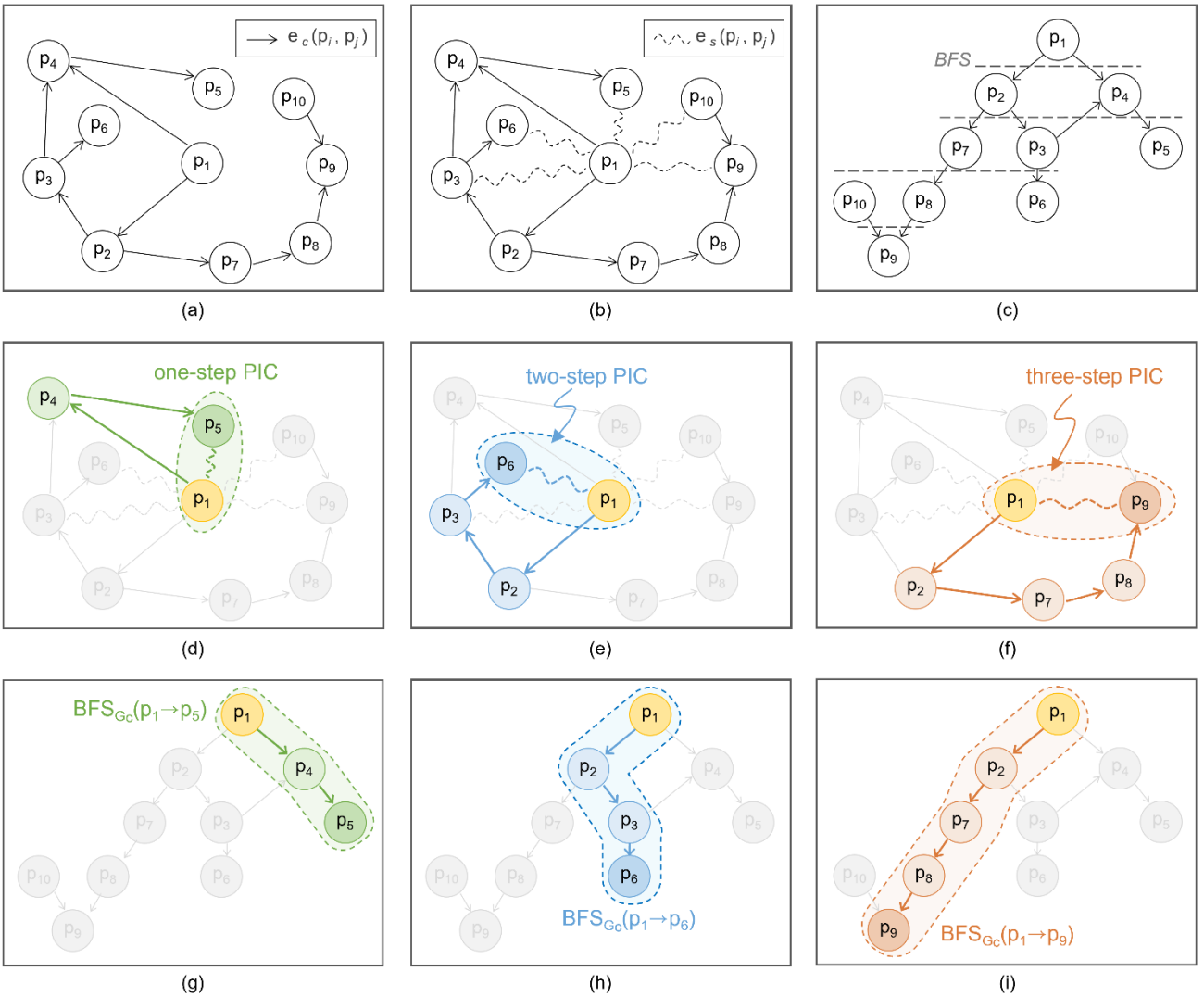


**FIGURE 1.** The process of exploring k-step PIC in the proposed method. (a) An example of a patent citation network ($G_c$) with 10 nodes. (b) The result of merging $G_c$ and similarity networks ($G_s$). In the figure, $p_1$ is similar to $p_2$ but not to $p_3$. (c) The process of finding the SP between two nodes in $G_c$. (d) $p_1$ is similar to $p_5$, but has no direct citation. (e) $p_1$ is similar to $p_6$, but without direct citation. (f) $p_1$ is similar to $p_9$, but without direct citation. The length of SP found as BFS for $p_1$ and $p_9$ is 4. (g) With the graph for (d), the shortest path can be found with BFS. The length of the SP is 2. Therefore, $p_1$ and $p_5$ are one-step PICs. (h) The SP for (e). (i) The SP for (f).

the positive correlation [58]–[60].

$$SIM\left(f\left(p_i\right), f\left(p_j\right)\right) = \frac{\sum^d f\left(p_i\right) \times f\left(p_j\right)}{\sqrt{\sum^d f\left(p_i\right)^2}\sqrt{\sum^d f\left(p_j\right)^2}} \quad (2)$$

Fig. 1(a) illustrates an example of $G_c$, where n is 10. In this case, $p_2 \in FC\ (p_1)$ is established because $p_2$ cites $p_1$. Fig. 1(b) illustrates a network that combines $G_c$ and $G_s$. $p_1$ is similar to $p_3$, $p_5$, $p_6$, $p_9$, and $p_{10}$. SIM ($f(p_1)$, $f(p_3)$) is greater than $SIM_{min}$, and SIM ($f(p_1)$, $f(p_7)$) is not. Thus, $p_1$ is more similar to $p_3$ than $p_7$. This study measures the similarity between a patent and its prior art based on this logic.

### A. PIC-EXPLORER

The inventor searches for prior art to register the developed technology as a patent. When they find prior art, they cite a document. These activities can increase patentability by inventing a technology that is more advanced than the prior art. Therefore, it is essential for inventors to quickly and accurately find similar patents. Furthermore, the attorney and examiner should be able to find prior art to increase the completeness of a patent.

An algorithm to solve this problem was proposed [61]. The researchers defined patents with indirect connections (PIC) as a pair of similar patents with no direct citation relationship. Furthermore, they proposed PIC-Explorer (PIC-E) to find PICs. PIC-E explores the relationship between $p_1$ and $p_5$, as depicted in Fig. 1(d). $p_1$ and $p_3$ have no direct citation relationship but are similar to each other. $p_1$ and $p_3$ have the same technological development flow as $p_1 \rightarrow p_2 \rightarrow p_3$. PIC-E detects this by searching for patents citing $p_1$ and FC ($p_1$). Next, it confirms

whether $p_3$ cites any patents among those belonging to FC ($p_1$).

PIC-E has several limitations:

- PIC-E can only find PICs, such as $p_1$ and $p_3$, that share only one patent (=$p_2$). This algorithm cannot search for PICs, such as $p_1$ and $p_6$.
- This method finds a patent that connects the two patents. Because of this mechanism, PIC-E has a very slow search speed.
- The PIC-E approach produced ambiguous results. In Fig. 1(b), $p_1$ and $p_5$ are PIC. If PIC-E is used, the two patents are PIC with flow $p_1 \rightarrow p_4 \rightarrow p_5$. However, the two patents can also be connected in the flow of $p_1 \rightarrow p_2 \rightarrow p_3 \rightarrow p_4 \rightarrow p_5$. Therefore, different conclusions may be drawn depending on which of the two paths are used for PIC-E.

### B. K-STEP PIC-EXPLORER

Several improvements have been proposed to solve the limitations of PIC-E mentioned in the previous section: (i) it should be possible to explore prior art that takes a longer path, (ii) the search time is fast, and it should have broader scalability, and (iii) the PIC derivation process should be clarified. In this paper, we propose an algorithm that addresses the problems with PIC-E. We define the k-step PIC as a relationship in which two similar patents without direct citations are connected through k patents. PIC-E, which was proposed in previous studies, can only find a one-step PIC. However, the proposed method can find a k-step PIC. We referred to this as the "k-step PIC-Explorer (k-step PIC-E)."

The following is a more detailed description of the proposed algorithm. First, our method determines whether the two patents are similar. If $p_i$ and $p_j$ are similar, $e_s(p_i, p_j) \in E_s$ hold. Furthermore, $e_c(p_i, p_j) \notin E_c$ is satisfied if the two patents are not cited. The input of the k-step PIC-E is a patent pair that is similar but has no citation relationship. The k-step PIC calculates the difference in registration dates to determine which of the two patents was registered earlier. The time difference (TD) calculated in (3) is the interval between the registration dates of the two patents.

$$TD = DATE_i - DATE_j \tag{3}$$

For example, when $DATE_i$ is January 1, 2021, and $DATE_j$ is July 30, 2021, TD is $-210$, so it is negative. Therefore, the algorithm determines that $p_i$ is a patent registered earlier than $p_j$ through the TD. We define a patent with an earlier registration date as $P_E$ and a later date as $P_L$. In the proposed method, if TD is 0, then $p_i$ is $P_E$, to consider the case of the same registration date.

Next, the k-step PIC-E limits the maximum value of TD to $TD_{max}$ (day) because when a researcher explores prior art, it is necessary to adjust the search scope according to the characteristics of the technology. When the user sets the time range for researching the prior art to the last five years, $TD_{max}$ is 1,825. If they search for the k-step PIC over all ranges, $TD_{max}$ is set to 7,300. A total of 7,300 (20 years) is the maximum time the patent rights are guaranteed.

When $P_E$ and $P_L$ are determined, the k-step PIC-E finds the shortest path (SP) from $P_E$ to $P_L$ in $G_c$ that connects two nodes in the graph. This study finds SP in $G_c$ using breadth-first search (BFS), a representative graph traversal method [62]–[64]. The characteristics of the BFS are as follows:

- BFS is a search algorithm that uses a queue and first visits all adjacent nodes from the start node.
- BFS searches all nodes of the current level before moving from the starting node to the next depth.
- BFS guarantees SP from the starting node to the target node.

The searched SP was $BFS_{Gc}$ ($P_E \rightarrow P_L$). For example, in Fig. 1(d), $p_1$ and $p_5$ are one-step PICs. In Fig. 1(a), the paths from $p_1$ to $p_5$ are $p_1 \rightarrow p_2 \rightarrow p_3 \rightarrow p_4 \rightarrow p_5$ and $p_1 \rightarrow p_4 \rightarrow p_5$. Furthermore, the path lengths are 4 and 2, respectively. However, as depicted in Fig. 1(g), the BFS returns the SP of the two patents as $p_1 \rightarrow p_4 \rightarrow p_5$. The k-step PIC-E measures the length of the SP of $P_E$ and $P_L$. If the two patents are k-step PICs, the length of SP is $k + 1$ (Fig. 1(g)–(i)). Therefore, $p_1$ and $p_5$ are one-step PICs, not three-step PICs.

**Algorithm 1** k-Step PIC-Explorer Algorithm

**Input:** $G_c = (V_c, E_c)$, $G_s = (V_s, E_s)$
**Output:** list of k-step PIC
Set: k, $TD_{max}$
$e_s(p_i, p_j) \in E_s$ and $e_c(p_i, p_j) \notin E_c$, $i < j$
*LOOP Process*
**for** i, j = 1 to n **do**
  TD = $DATE_i$ - $DATE_j$
  **if** $|TD| \leq TD_{max}$ **then**
    **if** TD $\leq 0$ **then**
      $P_E \leftarrow p_i$, $P_L \leftarrow p_j$
    **else**
      $P_E \leftarrow p_j$, $P_L \leftarrow p_i$
  **end if**
  SP = $BFS_{Gc}(P_E \rightarrow P_L)$
  **if** Length of SP = k + 1 **then**
    $P_E$ and $P_L$ are k-step PIC
  **end if**
**end if**
**end for**

**TABLE 1.** Results of searching for k-step PIC from Fig. 1(d)–(f).

| $P_E$ | $P_L$ | Number of steps | SP with BFS |
|---|---|---|---|
| $p_1$ | $p_3$ | 1 | $p_1 \rightarrow p_2 \rightarrow p_3$ |
| $p_1$ | $p_5$ | 1 | $p_1 \rightarrow p_4 \rightarrow p_5$ |
| $p_1$ | $p_6$ | 2 | $p_1 \rightarrow p_2 \rightarrow p_3 \rightarrow p_6$ |
| $p_1$ | $p_9$ | 3 | $p_1 \rightarrow p_2 \rightarrow p_7 \rightarrow p_8 \rightarrow p_9$ |

Table 1 illustrates the results of searching for the k-step PIC in Fig. 1(d)–(f). For example, in Fig. 1(e), $p_1$ and $p_9$ are three-step PICs. The implication is that there is no citation even though the two patents are similar to each other. Consequently, the k-step PIC-E determines the SP of $p_1$ and $p_9$ in $G_c$. As depicted in Fig. 1(i), the SP is $p_1 \rightarrow p_2 \rightarrow p_7 \rightarrow p_8 \rightarrow p_9$, and the length is 4 (=k + 1). Therefore, the two patents have a three-step PIC relationship.

## IV. EXPERIMENTS

We collected the actual patents to verify the proposed algorithm. The first experiment measured the sensitivity and scalability of the parameters of the k-step PIC-E. The second proceeds for the industrial applicability of the proposed algorithm. Finally, we propose a method for extending the k-step PIC-E to the applicant level. All experiments were conducted using Python 3.7.3 on a system with an Intel Core i5 processor and 16 GB of memory.

### A. DATASET AND EXPERIMENTAL SETUP

The experiment used two datasets: $D_1$ and $D_2$. $D_1$ includes data from 253 examiner citations (EC) registered in the United States. The examiner finds prior art similar to the patent to be filed. Moreover, the applicant cites prior art. The patent cited here is EC. EC is prior art that is not recognized by the applicant and the attorney. Accordingly, the applicant's technology is highly likely to infringe on the scope of the EC's rights. Therefore, the applicant increases patentability by citing the technology searched by the examiner.

We measure the sensitivity and scalability of the parameters of the k-step PIC-E by assuming that the patent cited by the examiner is the k-step PIC. Thus, a patent and its EC are k-step PICs. The sensitivity and scalability of k-step PIC-E are measured while finding a patent and its EC because EC is prior art that the applicant and attorney are unaware of.

$D_2$ includes 11,408 Information and Communication Technology (ICT) patents registered in the United States. These data were used to confirm the practical applicability of the k-step PIC-E. Furthermore, the applicant-level experiment was conducted using k-step PICs derived from the ICT field.

The last subsection describes the case of extending the proposed algorithm to the applicant level. The PAS is a patent-level approach. Accordingly, previous studies can only determine the relationship between individual technologies. Therefore, this paper presents a method for transforming the PAS to the applicant level to illustrate the scalability of the proposed algorithm.

### B. PARAMETER SENSITIVITY AND SCALABILITY

The experiment uses $D_1$ to examine the variability of $SIM_{min}$ and $TD_{max}$, the parameters of the k-step PIC-E. We found PIC by increasing k from 1 to 5; one-step PIC had the most with 238, two-step PIC had 10, and the rest had 5 each.

Fig. 2 illustrates the sensitivity of the $SIM_{min}$ and $TD_{max}$. The sensitivity of the parameter was expressed by accumulating the number of k-step PICs searched according to the threshold. Approximately 60% of the 253 k-step PICs had a similarity of ≥0.5 and a difference of 1,000 days or more on registration dates. As the value of $TD_{max}$ increases, the number of searched k-step PICs decreases.

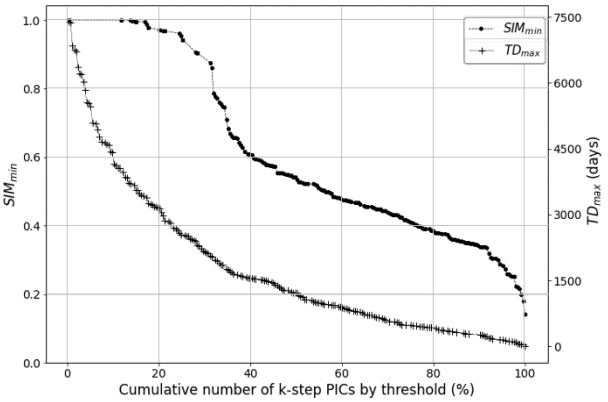


**FIGURE 2. Cumulative number of k-step PICs searched according to the threshold.**

**TABLE 2. Comparison of search time (s) between the proposed method and PIC-Explorer [61]. The method proposed in the previous study can only explore k-step PICs, where k is 1.**

| Algorithm | Number of steps | | | |
|---|---|---|---|---|
| | 1 | 2 | 3 | 4+ [a] |
| PIC-Explorer [61] | 29.08 | - | - | - |
| k-step PIC-Explorer | 0.35 | 0.01 | 0.01 | 0.02 |

[a] "4+" is when the number of steps is greater than or equal to 4.

Table 2 presents the results of comparing the search speeds of the algorithms. PIC-E required 29.08 s to find 238 one-step PICs from 253 patents. In contrast, k-step PIC-E requires 0.35 s, which is approximately 83 times faster than before. Furthermore, k-step PIC-E searches for two-step or higher PICs are very fast.

We statistically test whether the difference in search times of the two algorithms is greater than zero. Δtime calculated by (4) is the difference between the time required for PIC-E and the k-step PIC-E to find one-step PICs.

$$\Delta time = time_{PIC-E} - time_{k-step\ PIC-E} \tag{4}$$

For statistical testing, the paired t-test and Wilcoxon test were used to determine the difference in time required to find the same EC. The paired t-test tests the difference between a sample before and after a specific treatment. The Wilcoxon test is a non-parametric approach to the paired t-test. If k-step PIC-E finds the same EC faster than PIC-E, Δtime is expected to be greater than zero. Therefore, the hypotheses for the statistical tests are as follows:

- $H_0$: Mean of Δtime ≤ 0
- $H_1$: Mean of Δtime > 0

When the null hypothesis cannot be rejected at the significance level, the mean of Δtime is expected to be less

than or equal to zero—implying that k-step PIC-E searches the same EC more slowly than PIC-E. Conversely, if the null hypothesis is rejected, the k-step PIC-E searches faster than PIC-E.

**TABLE 3. Results of statistical test for Δtime.**

| Variable | Avg [a] | SD [b] | paired t-test | | Wilcoxon test | |
|---|---|---|---|---|---|---|
| | | | Statistic | p-value | Statistic | p-value |
| Δtime | 0.12 | 0.18 | 10.63 | < 0.001 | 0.00 | < 0.001 |

[a] Avg and [b] SD are the average and standard deviation of Δtime.

Table 3 presents the results of the statistical tests for the mean of Δtime. Because the average is 0.12, PIC-E requires more time to search than the k-step PIC-E. The p-values for both tests were less than 0.001. Thus, when using a significance level of 0.05, the null hypothesis was rejected. Therefore, the search speed of the k-step PIC-E is faster than before, with statistical significance.

### C. EXPERIMENTAL STUDY

ICT is a technology that partially or entirely replaces existing industrial structures and business models. Furthermore, it is a representative technology in which companies, research institutes, and even the nation must continuously conduct research to enhance future competitiveness [65]. Thus, we collected 11,408 ICT patents to experiment with the k-step PIC-E.

Modern search systems do not rely solely on matching keywords; therefore, we should use semantic information [66], [67]. The similarity between patents was calculated using the title of the invention, abstract, and representative claims. We preserve the context in the document using sentence bidirectional encoder representations from transformers (SBERT) [68] as function $f$. Thus, our implementation uses a pre-trained model (paraphrase-MiniLM-L6-v2).[1] The text of the patent is embedded in $\mathbb{R}^{384}$ using a pre-trained SBERT. SIM uses cosine similarity to calculate the distance of a patent.

For exploring the k-step PIC of ICT, $SIM_{min}$ and $TD_{max}$ were set to 0.5 and 7,300, respectively. Moreover, the range of k was set from 1 to 5. Table 4 presents the number of PICs searched for each step and the distribution of the TDs. When k is 1 and 2, the median of the TD is less than five years. When k is 3 or more, it is less than six years. Consequently, the patents registered within six years of ICT are very similar to each other, although there are no citations.

Table 5 presents a list of the searched prior art. Patents filed by QUALCOMM and GHCOMM have a one-step PIC. The patents of QUALCOMM and GHCOMM were registered on August 23, 2016, and January 26, 2021, respectively.

Both patents are related to orthogonal frequency-division multiplexing (OFDM). This method is used to transmit high-speed data to wireless channels. Mathematical transformations are performed to obtain and transmit a frequency response to a processor device for generating wireless communications. Some of the representative claims of QUALCOMM's patent are as follows:

- The circuit is configured to perform a discrete Fourier transform (DFT) on a discrete signal of complex data symbols to obtain a first frequency response vector.
- The circuit is configured to generate a low peak-to-average OFDM signal representing the vector of digital signal samples for transmission.

GHCOMM's patent describes how a radio transceiver uses codes to produce symbols. This technique uses a matrix-based mathematical transformation formula to create a spreading code. Some representative claims of this technology are as follows:

- Spread the plurality data symbols, where the set of complex-valued orthogonal spreading codes comprises rows or columns of a DFT matrix.
- The spread symbols are modeled onto the set of OFDM subcarriers to generate a discrete-time OFDM transmission signal, where the spreading reduces the peak-to-average power.

As a result of comparing the claims, the two patents were similar in obtaining and converting data using mathematical transformation and transmitting it again. The two companies can use these results to build five counterstrategies.

First, because QUALCOMM's patent is prior art, GHCOMM's technology is likely to infringe on the scope of its rights. Accordingly, QUALCOMM can claim royalties from them in litigation. Second, QUALCOMM can assert the logic of invalidating GHCOMM's patent if their technological competitiveness is high. Third, the GHCOMM prepares patent litigation using QUALCOMM. They may argue for invalidation or non-infringement of QUALCOMM's patents if they believe they are more likely to win the dispute. Fourth, GHCOMM can reduce damage by licensing with QUALCOMM when it is expected to lose. Finally, GHCOMM can design avoidance to avoid infringing on the scope of the QUALCOMM's rights. Consequently, GHCOMM will be able to reduce the risk of patent litigation. The 1,459 k-step PICs we found in our experiments, including

**TABLE 4. Basic statistics of TD according to k.**

| Basic statistic | | Number of steps | | | |
|---|---|---|---|---|---|
| | | 1 | 2 | 3 | 4+ |
| Frequency [a] | | 806 | 409 | 187 | 57 |
| TD | Maximum | 6,300 | 6,188 | 6,727 | 5,719 |
| | $Q_3$ [b] | 2,499 | 2,338 | 2,938 | 2,468 |
| | Median | 1,610 | 1,533 | 1,967 | 2,009 |
| | $Q_1$ [c] | 735 | 859 | 879 | 494 |
| | Minimum | 0 | 0 | 0 | 7 |

[a] "Frequency" is the number of k-step PICs searched according to k. [b, c] $Q_3$ and $Q_1$ represent the third and first quartiles.

[1] https://www.sbert.net/docs/pretrained_models

**TABLE 5.** Results of searching for k-step PICs in ICT patents.

| k | No | RD [a] | Applicants | Title of Patent Document |
|---|---|---|---|---|
| 1 | 1 | 2016.08.23 | QUALCOMM | Signaling method in an OFDM multiple access system Pre-coding in OFDM |
| | | 2021.01.26 | GHCOMM | Pre-coding in OFDM |
| | 2 | 2011.08.09 | QUALCOMM | Method and apparatus for high rate packet data transmission |
| | | 2014.05.27 | InterDigital | Code division multiple access (CDMA) communication system |
| | 3 | 2008.08.19 | AVAGO | Network switching architecture with multiple table synchronization, and forwarding of both IP and IPX packets |
| | | 2010.11.23 | CISCO | Distributed forwarding in virtual network devices |
| | 4 | 2012.08.07 | KAIAM | Method and apparatus for requesting selected interlace mode in wireless communication systems |
| | | 2021.10.16 | QUALCOMM | Method and apparatus for processing simultaneous assignment in wireless communication systems |
| | 5 | 2009.09.08 | APPLE | Wireless communications system that supports multiple modes of operation |
| | | 2013.08.27 | QUALCOMM | Method and apparatus for monitoring other channel interference in wireless communication system |
| 2 | 6 | 2016.08.23 | QUALCOMM | Method and apparatus for high rate packet data and low delay data transmissions |
| | | 2016.08.23 | SAMSUNG | Apparatus and method for transmitting signal in a wireless communication system |
| | 7 | 2015.02.17 | ADAPTIX | Multi-carrier communications with adaptive cluster configuration and switching |
| | | 2019.03.26 | REARDEN | Systems and methods to enhance spatial diversity in distributed-input distributed-output wireless systems |
| | 8 | 2012.10.16 | QUALCOMM | Method and apparatus for processing open state in wireless communication system |
| | | 2017.02.07 | GOOGLE | Method and apparatus to detect the transmission bandwidth configuration of a channel in connection with reducing interference between channels in wireless communication systems |
| | 9 | 2012.08.21 | QUALCOMM | Method of transmitting and receiving a redirect message in a wireless communication system |
| | | 2012.08.21 | LG | Method of transmitting control signal in wireless communication system |
| 3 | 10 | 2012.10.16 | QUALCOMM | Method and apparatus for processing simultaneous assignment in wireless communication systems |
| | | 2015.01.13 | MOTOROLA | Method and apparatus using two radio access technologies for scheduling resources in wireless communication systems |
| | 11 | 2015.08.11 | QUALCOMM | Method and apparatus for determining a data rate in a high rate packet data wireless communications system |
| | | 2016.12.06 | COMCAST | Originator and recipient based transmissions in wireless communications |
| | 12 | 2011.11.29 | QUALCOMM | Method and apparatus for predicting favored supplemental channel transmission slots using transmission power measurements of a fundamental channel |
| | | 2017.02.07 | GOOGLE | Method and apparatus to detect the transmission bandwidth configuration of a channel in connection with reducing interference between channels in wireless communication systems |
| 4+ | 13 | 2011.11.22 | QUALCOMM | Method and apparatus using a multi-carrier forward link in a wireless communication system |
| | | 2015.01.13 | MOTOROLA | Method and apparatus using two radio access technologies for scheduling resources in wireless communication systems |
| | 14 | 2011.04.05 | INTEL | Mode selection for data transmission in wireless communication channels based on statistical parameters |
| | | 2016.05.03 | ODYSSEY | Systems/methods of adaptively varying a spectral content of communications |

ADAPTIX; *Adaptix, INC*. APPLE; *Apple, INC*. AVAGO; *Avago Technologies International Sales PTE. LIMITED*. CISCO; *Cisco Technology, INC*. COMCAST; *Comcast Cable Communications, LLC*. GHCOMM; *Genghiscomm Holdings, LLC*. GOOGLE; *Google Technology Holdings, LLC*. INTEL; *Intel CORP*. InterDigital; *InterDigital Technology CORP*. KAIAM; *Kaiam CORP*. LG; *LG Electronics, INC*. MOTOROLA; *Motorola Mobility, LLC*. ODYSSEY; *Odyssey Wireless, INC*. QUALCOMM; *Qualcomm, INC*. REARDEN; *Rearden, LLC*. SAMSUNG; *Samsung Electronics Co., Ltd*.

[a] "RD" is the registration date.

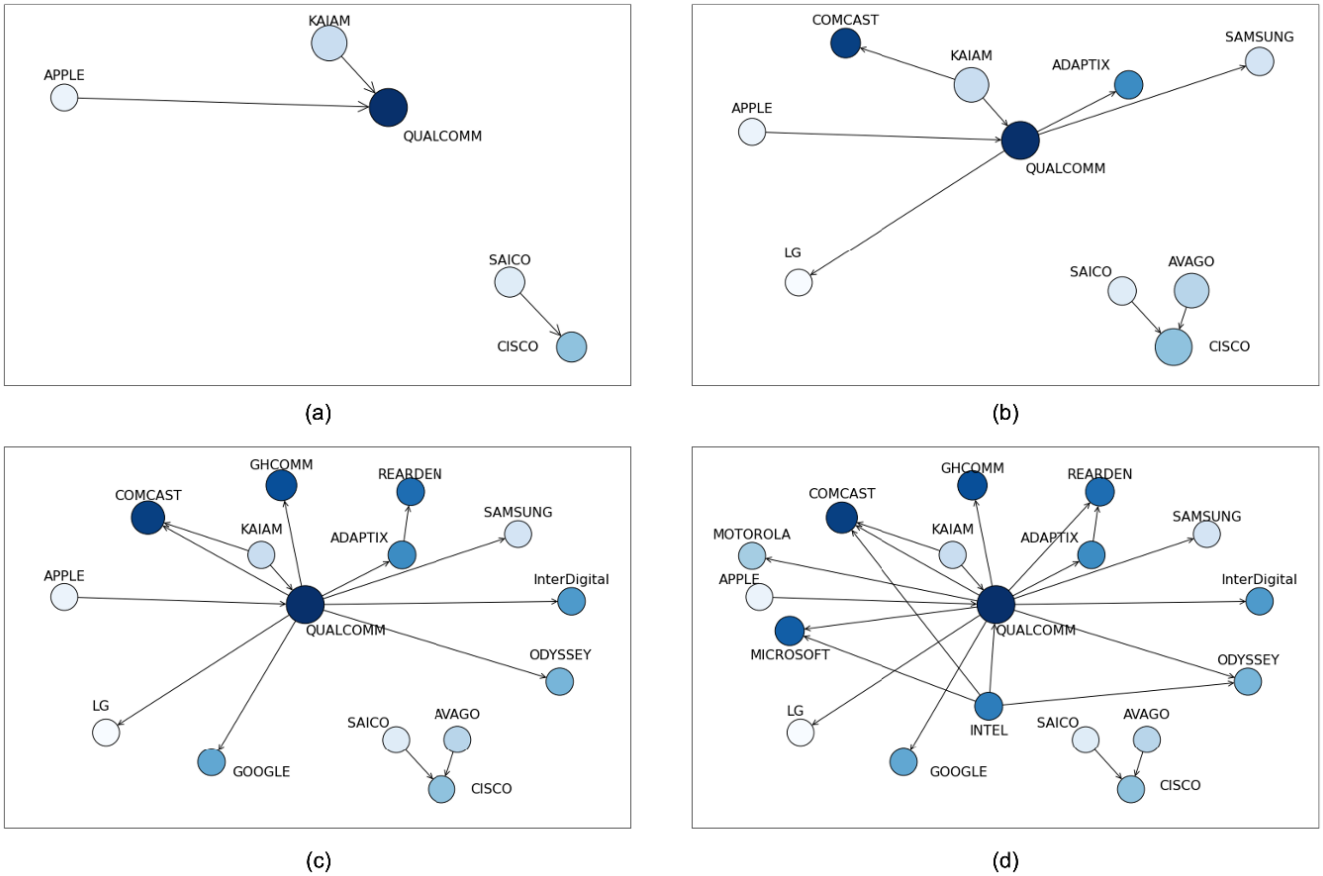


**FIGURE 3.** Visualization of k-step PIC converted to applicant level. (a) Results when k is 1. (b) and (c) are the results when k is less than or equal to 2 and 3, respectively. (d) Results of all k-step PICs.

those listed in Table 5, are in the github.com/lee-ju/k_step_PIC/paper/appendix.

### D. APPLICATIONS

The proposed algorithm is a patent-level method for identifying similar technologies without citations. This subsection extends our method to the applicant level. In Table 1, suppose that the applicants of $p_1$, $p_3$, $p_5$, and $p_6$ are A, B, C, and D, respectively, and that the applicant of $p_9$ is the same as $p_3$. Then, the technologies can be organized by the applicant. $p_1$ and $p_3$ are one-step PICs, indicating a relationship between the patents registered by A and B in the prior art. Moreover, $p_9$ is a patent of B. Therefore, applicants A and B have two cases of k-step PICs. Accordingly, the result of the k-step PIC-E can be extended to the applicant level.

Table 6 presents an extract from 1,459 k-step PICs searched for in ICT patents only when the applicants differ. One one-step PIC was searched for APPLE and QUALCOMM. This result indicates that among the patents of the two applications, there is one case that is similar but does not have a citation. When converted to the applicant level, QUALCOMM has a high frequency of appearance. This suggests that other applicants' patents are similar to those of QUALCOMM. Because most QUALCOMM's patents are $P_E$, it is highly likely that they possess ICT-related source technology.

Fig. 3 illustrates the result of expressing the k-step PIC converted to the applicant level as a network. For example, Fig. 3(a) illustrates only applicants with only a one-step PIC, as presented in Table 6. In Fig. 3(d), the number of edges is highest in the order of QUALCOMM, INTEL, and COMCAST. In QUALCOMM, 11 out of 14 edges extend outward, suggesting that their patents are likely to be prior art to other patents. In contrast to QUALCOMM and INTEL, COMCAST has all three edges facing inward, suggesting that the patent for COMCAST was registered without citing similar prior art.

The advantages of converting the proposed algorithm to the applicant level and expressing it as a network are as follows:

- Because graphs expressed as networks can be intuitively interpreted, researchers (even if they are not experts) can easily understand them.
- A company or research institute can devise a counter-strategy by expanding to the applicant level.
- Researchers can derive technological insights through subnetworks. Fig. 3(d) can be divided into subnetworks that include AVAGO, CISCO, and SAICO, and those

**TABLE 6. Results of converting k-step PIC to applicant level.**

| Fig. 3 | $P_E$ | $P_L$ | Number of steps | | | |
|---|---|---|---|---|---|---|
| | | | 1 | 2 | 3 | 4+ |
| (a) | APPLE | QUALCOMM | 1 | – | – | – |
| | SAICO | CISCO | 2 | – | – | – |
| | KAIAM | QUALCOMM | 4 | – | – | – |
| (b) | AVAGO | CISCO | 5 | 2 | – | – |
| | QUALCOMM | SAMSUNG | 1 | 1 | – | – |
| | QUALCOMM | LG | – | 1 | – | – |
| | QUALCOMM | ADAPTIX | – | 2 | – | – |
| | KAIAM | COMCAST | – | 3 | – | – |
| (c) | ADAPTIX | REARDEN | 26 | 14 | 2 | – |
| | QUALCOMM | COMCAST | 211 | 47 | 62 | – |
| | QUALCOMM | GHCOMM | 177 | 6 | 7 | – |
| | QUALCOMM | GOOGLE | 18 | 6 | 2 | – |
| | QUALCOMM | InterDigital | 3 | 8 | 17 | – |
| | QUALCOMM | ODYSSEY | 21 | – | 2 | – |
| (d) | INTEL | MICROSOFT | – | – | – | 1 |
| | INTEL | ODYSSEY | – | – | 1 | 1 |
| | INTEL | COMCAST | – | – | 3 | 4 |
| | QUALCOMM | MOTOROLA | – | 1 | 5 | 2 |
| | INTEL | QUALCOMM | 12 | 21 | 22 | 12 |
| | QUALCOMM | REARDEN | 59 | 21 | 8 | 1 |
| | QUALCOMM | MICROSOFT | 10 | 136 | 9 | 20 |
| Total | | | 550 | 269 | 140 | 41 |

MICROSOFT; *Microsoft Technology Licensing, LLC.* SAICO; *Saico Information Technology CO., LTD.*

that do not. Then, the researcher can analyze in-depth the differences by subnetwork.

The proposed method can be applied at the patent and applicant levels. The former can be used by researchers when searching for prior art. Accordingly, researchers can improve the quality of their patents and reduce the risk of unnecessary litigation. The latter can be used by applicants to track technological developments. Then, applicants can derive the relationship between companies through technology and plan various management strategies accordingly.

## V. CONCLUSION

Recently, the technology market has rapidly changed. Companies and research institutes are developing various technologies to keep pace with them. Patents give inventors exclusive rights instead of making the technologies publicly available. Because of these advantages, many inventors want to patent their technologies.

Patents are registered through a series of processes. PAS is one of the most critical tasks required for applicants, attorneys, and examiners. Prior art is historical data related to patentability, which they investigated and used to improve their technologies. However, as developed technologies expand, it is difficult for applicants, attorneys, and examiners to search for prior art because PAS must determine the level of understanding of the technology and the similarity of patents and infringement of the scope of rights. Nevertheless, a successful PAS increases an applicant's competitiveness and lowers the risk of patent litigation.

Previous studies have suggested methods using keywords, semantics, and metadata for efficient PAS. Keyword-based approaches expand or contract queries to search patent databases. However, this method is highly dependent on the query keywords.

Furthermore, metadata and semantic-based approaches must be used together to compensate for their respective limitations. Nevertheless, many previous studies have the disadvantage of using only one of the two types of information. Therefore, this study attempted to mitigate these limitations using a semantic-based approach with metadata.

We defined the concept of a k-step PIC for a successful PAS. Moreover, this paper proposes a k-step PIC-E algorithm to find k-step PICs. PIC is a relationship in which two patents are similar but not cited, and the k-step is the number of patents shared by two patents. Examiners can find PICs and discover prior art that they are not aware of. These results can be obtained by extending the scope of the search by increasing k.

The experiment on the parameter measured the number of k-step PICs found according to the $TD_{max}$, the maximum value of the difference between the registration dates of the two patents. $TD_{max}$ was able to explore much prior art when using a small value. We also conducted experiments on 11,408 ICT patents. The k-step PIC found in the ICT patent is a technology that primarily uses OFDM. The patents of QUALCOMM and GHCOMM are similar in terms of frequency conversion through mathematical methods. We confirmed the practical applicability of the proposed algorithm through experiments.

The limitations of our study are as follows:

- The proposed method uses forward citations of patents. Our method does not refer to backward citations. Because backward citation is cited when applying for a patent, it is possible that information contained in the prior art is included.
- This study includes the hyperparameter k but does not provide a guideline for determining the optimal value.
- The proposed algorithm does not consider an applicant. If two patents owned by the same applicant are k-step PICs, it is unnecessary to prepare a counterstrategy.

Future research should reflect the metadata of various patents (e.g., applicants and family patents) because patent citations are increasingly criticized for being used to measure knowledge spillover or relatedness [69]. Kuhn *et al.* [70] emphasized that a small number of patents generate a majority of patent citations, and the similarities between them are diminishing considerably. Because the effects of the problem are expected to grow gradually over time, a next-generation PAS must be designed to measure knowledge similarity and investigate prior art faster and more efficiently than before.

## REFERENCES


[1] P. S. Menell, "An analysis of the scope of copyright protection for application programs," *Stanford Law Rev.*, vol. 41, pp. 1045–1104, Nov. 1988.

[2] Y. Liao, F. Deschamps, E. D. F. R. Loures, and L. F. P. Ramos, "Past, present and future of industry 4.0—A systematic literature review and research agenda proposal," *Int. J. Prod. Res.*, vol. 55, no. 12, pp. 3609–3629, Mar. 2017, doi: 10.1080/00207543.2017.1308576.

[3] J. M. Müller, O. Buliga, and K.-I. Voigt, "Fortune favors the prepared: How SMEs approach business model innovations in industry 4.0," *Technol. Forecasting Social Change*, vol. 132, pp. 2–17, Jul. 2018, doi: 10.1016/j.techfore.2017.12.019.

[4] W. Seo, N. Kim, and S. Choi, "Big data framework for analyzing patents to support strategic R&D planning," in *Proc. IEEE 14th Int. Conf. Dependable, Autonomic Secure Comput., 14th Int. Conf Pervasive Intell. Comput., 2nd Int. Conf. Big Data Intell. Comput. Cyber Sci. Technol. Congress(DASC/PiCom/DataCom/CyberSciTech)*, Auckland, New Zealand, Aug. 2016, pp. 746–753, doi: 10.1109/DASC-PICom-DataCom-CyberSciTec.2016.131.

[5] F. M. Scherer and D. Harhoff, "Technology policy for a world of skew-distributed outcomes," *Res. Policy.*, vol. 29, nos. 4–5, pp. 559–566, Apr. 2000, doi: 10.1016/S0048-7333(99)00089-X.

[6] B. N. Sampat, "When do applicants search for prior art?" *J. Law Econ.*, vol. 53, no. 2, pp. 399–416, May 2010, doi: 10.1086/651959.

[7] C. Langinier and P. Marcoul, "The search of prior art and the revelation of information by patent applicants," *Rev. Ind. Org.*, vol. 49, no. 3, pp. 399–427, Apr. 2016, doi: 10.1007/s11151-016-9514-3.

[8] J. R. Allison and M. A. Lemley, "Empirical evidence on the validity of litigated patents," *AIPLA Q. J.*, vol. 26, no. 3, pp. 185–276, 1998.

[9] R. Lampe, "Strategic citation," *Rev. Econ. Statist.*, vol. 94, no. 1, pp. 320–333, Feb. 2012, doi: 10.1162/REST_a_00159.

[10] J. R. Thomas, "Collusion and collective action in the patent system: A proposal for patent bounties," in *Entrepreneurial Inputs and Outcomes: New Studies of Entrepreneurship in the United States*, vol. 13. West Yorkshire, U.K.: Emerald Group Publishing Limited, Nov. 2001. [Online]. Available: https://www.emerald.com/insight/content/doi/10.1016/S1048–4736(01)13006–7/full/html

[11] A. B. Jaffe and J. Lerner, *Innovation and its Discontents: How Our Broken Patent System is Endangering Innovation and Progress, and What to do About it*. Princeton, NJ, USA: Princeton Univ. Press. 2004, doi: 10.1016/S1048-4736(01)13006-7.

[12] T. S. S. DeSanti, W. E. Cohen, G. F. Levine, H. J. Greene, M. Bey, M. S. Wroblewski, R. Moore, M. Barnett, N. Gorham, C. Kohrs, D. Scheffman, M. Frankean, R. Levy, A. F. Abbott, S. Michel, P. Pidano, and K. Lubell, "To promote innovation: The proper balance of competition and patent law and policy: Executive summary," *Berkeley Technol. Law J.*, vol. 19, no. 3, pp. 861–883, 2004. [Online]. Available: https://www.jstor.org/stable/24116718/

[13] M. Kang, S. Lee, and W. Lee, "Prior art search using multi-modal embedding of patent documents," in *Proc. IEEE Int. Conf. Big Data Smart Comput. (BigComp)*, Busan, South Korea, Feb. 2020, pp. 548–550, doi: 10.1109/BigComp48618.2020.000-6.

[14] J. Michel and B. Bettels, "," *Scientometrics*, vol. 51, no. 1, pp. 185–201, 2001, doi: 10.1023/a:1010577030871.

[15] J. A. Jeffery, "Preserving the presumption of patent validity: An alternative to outsourcing the U.S. Patent examiner's prior art search," *Cath. UL Rev.*, vol. 52, no. 3, pp. 761–802, 2002.

[16] E. Marttin and A.-C. Derrien, "How to apply examiner search strategies in Espacenet. A case study," *World Pat. Inf.*, vol. 54, pp. S33–S43, Sep. 2018, doi: 10.1016/j.wpi.2017.06.001.

[17] S. W. Graf, "Improving patent quality through identification of relevant prior art: Approaches to increase information low to the patent office," *Lewis Clark Law Rev.*, vol. 11, no. 2, pp. 495–520, 2007.

[18] D. Somaya, "Patent strategy and management: An integrative review and research agenda," *J. Manage.*, vol. 38, no. 4, pp. 1084–1114, May 2012, doi: 10.1177/0149206312444447.

[19] S. Nambisan and M. Sawhney, "A buyer's guide to the innovation bazaar," *Harvard Bus. Rev.*, vol. 85, no. 6, pp. 109–118, Jun. 2007.

[20] M. Bogers, H. Chesbrough, S. Heaton, and D. J. Teece, "Strategic management of open innovation: A dynamic capabilities perspective," *California Manage. Rev.*, vol. 62, no. 1, pp. 77–94, Nov. 2019, doi: 10.1177/0008125619885150.

[21] J. O. Lanjouw and M. Schankerman, "Characteristics of patent litigation: A window on competition," *RAND J. Econ.*, vol. 32, no. 1, pp. 129–151, 2001, doi: 10.2307/2696401.

[22] J. E. Bessen and M. J. Meurer, "Patent litigation with endogenous disputes," *Amer. Econ. Rev.*, vol. 96, no. 2, pp. 77–81, Apr. 2006, doi: 10.1257/000282806777212288.

[23] J. Lerner, "The litigation of financial innovations," *J. Law Econ.*, vol. 53, no. 4, pp. 807–831, Nov. 2010, doi: 10.1086/655757.

[24] B. C. Rudy and S. L. Black, "Attack or defend? The role of institutional context on patent litigation strategies," *J. Manage.*, vol. 44, no. 3, pp. 1226–1249, Mar. 2018, doi: 10.1177/0149206315605168.

[25] M. Corsino, M. Mariani, and S. Torrisi, "Firm strategic behavior and the measurement of knowledge flows with patent citations," *Strategic Manage. J.*, vol. 40, no. 7, pp. 1040–1069, Mar. 2019, doi: 10.1002/smj.3016.

[26] S. Chih-Yi and L. Bou-Wen, "Attack and defense in patent-based competition: A new paradigm of strategic decision-making in the era of the fourth industrial revolution," *Technol. Forecasting Social Change*, vol. 167, Jun. 2021, Art. no. 120670, doi: 10.1016/j.techfore.2021.120670.

[27] X. Yu and B. Zhang, "Obtaining advantages from technology revolution: A patent roadmap for competition analysis and strategy planning," *Technol. Forecasting Social Change*, vol. 145, pp. 273–283, Aug. 2019, doi: 10.1016/j.techfore.2017.10.008.

[28] M. Schankerman and S. Scotchmer, "Damages and injunctions in protecting intellectual property," *RAND J. Econ.*, vol. 32, no. 1, pp. 199–220, 2001, doi: 10.2307/2696404.

[29] C. Shapiro, "Injunctions, hold-up, and patent royalties," *Amer. Law Econ. Rev.*, vol. 12, no. 2, pp. 280–318, Aug. 2010, doi: 10.1093/aler/ahq014.

[30] D. Somaya, "Strategic determinants of decisions not to settle patent litigation," *Strategic Manage. J.*, vol. 24, no. 1, pp. 17–38, Jan. 2003, doi: 10.1002/smj.281.

[31] T. S. O'Connor, "Trade dress: The increasing importance of an ancient yet new form of intellectual property protection," *J. Bus. Res.*, vol. 67, no. 3, pp. 303–306, Mar. 2014, doi: 10.1016/j.jbusres.2013.05.017.

[32] D. Somaya, "How patent strategy affects the timing and method of patent litigation resolution," in *Strategy Beyond Markets*, vol. 34. West Yorkshire, U.K.: Emerald Group Publishing Limited, May 2016. [Online]. Available: https://www.emerald.com/insight/content/doi/10.1108/S0742–332220160000034014/full/html/

[33] W. Shalaby and W. Zadrozny, "Patent retrieval: A literature review," *Knowl. Inf. Syst.*, vol. 61, no. 2, pp. 631–660, Jan. 2019, doi: 10.1007/s10115-018-1322-7.

[34] S. Jones, "Graphical query specification and dynamic result previews for a digital library," in *Proc. 11th Annu. ACM Symp. User interface Softw. Technol. (UIST)*, California, SF, USA, 1998, pp. 143–151, doi: 10.1145/288392.288595.

[35] P. G. Anick, J. D. Brennan, R. A. Flynn, D. R. Hanssen, B. Alvey, and J. M. Robbins, "A direct manipulation interface for Boolean information retrieval via natural language query," in *Proc. 13th Annu. Int. ACM SIGIR Conf. Res. Develop. Inf. Retr. (SIGIR)*, Brussels, Belgium, 1990, pp. 135–150, doi: 10.1145/96749.98015.

[36] T. Russell-Rose, J. Chamberlain, and F. Shokraneh, "A visual approach to query formulation for systematic search," in *Proc. Conf. Hum. Inf. Interact. Retr.*, Mar. 2019, pp. 379–383, doi: 10.1145/3295750.3298919.

[37] S. Tiwana and E. Horowitz, "FindCite: Automatically finding prior art patents," in *Proc. 2nd Int. Workshop Pat. Inf. Retr. - PaIR*, Hong Kong, 2009, pp. 37–40, doi: 10.1145/1651343.1651352.

[38] C. G. Harris, R. Arens, and P. Srinivasan, "Using classification code hierarchies for patent prior art searches," in *Current Challenges in Patent Information Retrieval*, Berlin, Germany, 2011, pp. 287–304, doi: 10.1007/978-3-642-19231-9_14.

[39] S. Wang, Z. Lei, and W.-C. Lee, "Exploring legal patent citations for patent valuation," in *Proc. 23rd ACM Int. Conf. Conf. Inf. Knowl. Manage.*, Nov. 2014, pp. 1379–1388, doi: 10.1145/2661829.2662029.

[40] S. Arts and R. Veugelers, "Technology familiarity, recombinant novelty, and breakthrough invention," *Ind. Corporate Change*, vol. 24, no. 6, pp. 1215–1246, Dec. 2015, doi: 10.1093/icc/dtu029.

[41] O. Couteau, "Forward searching—A complement to keyword- and class-based patentability searches," *World Pat. Inf.*, vol. 37, pp. 33–38, Jun. 2014, doi: 10.1016/j.wpi.2014.01.007.

[42] I. von Wartburg, T. Teichert, and K. Rost, "Inventive progress measured by multi-stage patent citation analysis," *Res. Policy*, vol. 34, no. 10, pp. 1591–1607, Dec. 2005, doi: 10.1016/j.respol.2005.08.001.

[43] H. Small, "Co-citation in the scientific literature: A new measure of the relationship between two documents," *J. Amer. Soc. Inf. Sci.*, vol. 24, no. 4, pp. 265–269, Jul. 1973, doi: 10.1002/asi.4630240406.

[44] B. Gipp and J. Beel, "Citation proximity analysis (CPA)—A new approach for identifying related work based on co-citation analysis," in *Proc. Scientometrics Informetrics (ISSI)*, Rio de Janeiro, Brazil, 2009, pp. 571–575. [Online]. Available: https://www.sciplore.org/wp-content/papercite-data/pdf/gipp09a.pdf/

[45] H. Gurulingappa, B. Müller, R. Klinger, H. T. Mevissen, M. Hofmann-Apitius, C. M. Friedrich, and J. Fluck, "Prior art search in chemistry patents based on semantic concepts and co-citation analysis," in *Proc. TREC*, 2010, pp. 1–9. [Online]. Available: https://trec.nist.gov/pubs/trec19/papers/fraunhofer-scai.chem.rev.pdf/

[46] E. Schiebel, "Visualization of research fronts and knowledge bases by three-dimensional areal densities of bibliographically coupled publications and co-citations," *Scientometrics*, vol. 91, no. 2, pp. 557–566, May 2012, doi: 10.1007/s11192-012-0626-8.

[47] X. Wang, Y. Zhao, R. Liu, and J. Zhang, "Knowledge-transfer analysis based on co-citation clustering," *Scientometrics*, vol. 97, no. 3, pp. 859–869, Jul. 2013, doi: 10.1007/s11192-013-1077-6.

[48] H. J. No, Y. An, and Y. Park, "A structured approach to explore knowledge flows through technology-based business methods by integrating patent citation analysis and text mining," *Technol. Forecasting Social Change*, vol. 97, pp. 181–192, Aug. 2015, doi: 10.1016/j.techfore.2014.04.007.

[49] H. Nakamura, S. Suzuki, I. Sakata, and Y. Kajikawa, "Knowledge combination modeling: The measurement of knowledge similarity between different technological domains," *Technol. Forecasting Social Change*, vol. 94, pp. 187–201, May 2015, doi: 10.1016/j.techfore.2014.09.009.

[50] K. Sanguri, A. Bhuyan, and S. Patra, "A semantic similarity adjusted document co-citation analysis: A case of tourism supply chain," *Scientometrics*, vol. 125, no. 1, pp. 233–269, Jul. 2020, doi: 10.1007/s11192-020-03608-0.

[51] A. Cammarano, F. Michelino, M. P. Vitale, M. La Rocca, and M. Caputo, "Technological strategies and quality of invention: The role of knowledge base and technical applications," *IEEE Trans. Eng. Manag.*, early access, Mar. 2, 2020, doi: 10.1109/TEM.2020.2973861.

[52] D. Korobkin, S. Fomenkov, A. Kravets, S. Kolesnikov, and M. Dykov, "Three-steps methodology for patents prior-art retrieval and structured physical knowledge extracting," *Commun. Comput. Inf. Sci.*, vol. 535, pp. 124–136, Dec. 2015, doi: 10.1007/978-3-319-23766-4_10.

[53] 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, doi: 10.1002/smj.2699.

[54] D. M. Korobkin, S. A. Fomenkov, and A. G. Kravets, "Methods for extracting the descriptions of sci-tech effects and morphological features of technical systems from patents," in *Proc. 9th Int. Conf. Inf., Intell., Syst. Appl. (IISA)*, Zakynthos, Greece, Jul. 2018, pp. 1–4, doi: 10.1109/IISA.2018.8633624.

[55] C. Righi and T. Simcoe, "Patent examiner specialization," *Res. Policy*, vol. 48, no. 1, pp. 137–148, Feb. 2019, doi: 10.1016/j.respol.2018.08.003.

[56] S. Arts, J. Hou, and J. C. Gomez, "Natural language processing to identify the creation and impact of new technologies in patent text: Code, data, and new measures," *Res. Policy*, vol. 50, no. 2, Mar. 2021, Art. no. 104144, doi: 10.1016/j.respol.2020.104144.

[57] N. Borodin, D. Korobkin, A. Bezruchenko, and S. Fomenkov, "The search for R&D partners based on patent data," *J. Phys., Conf.*, vol. 2060, no. 1, Oct. 2021, Art. no. 012022, doi: 10.1088/1742-6596/2060/1/012022.

[58] M. McGill. (1978). *An Evaluation of Factors Affecting Document Ranking by Information Retrieval Systems*. School of Information Studies. [Online]. Available: https://files.eric.ed.gov/fulltext/ED188587.pdf/

[59] G. Salton and M. McGill, "Introduction to modern information retrieval," in *Information Storage and Retrieval Systems*. New York, NY, USA: McGraw-Hill, 1986.

[60] J. Zhang and R. R. Korfhage, "A distance and angle similarity measure method," *J. Amer. Soc. Inf. Sci.*, vol. 50, no. 9, pp. 772–778, Jun. 1999, doi: 10.1002/(SICI)1097-4571(1999)50:9<772:AID-ASI5>3.0.CO.2-E.

[61] J. Lee, S. Park, and J. Kang, "Introducing patents with indirect connection (PIC) for establishing patent strategies," *Sustainability*, vol. 13, no. 2, p. 820, Jan. 2021, doi: 10.3390/su13020820.

[62] E. F. Moore, "The shortest path through a maze," in *Proc. Int. Symp. Switching Theory*, Cambridge, MA, USA, 1959, pp. 285–292.

[63] C. Y. Lee, "An algorithm for path connections and its applications," *IRE Trans. Electron. Comput.*, vol. EC-10, no. 3, pp. 346–365, Sep. 1961, doi: 10.1109/TEC.1961.5219222.

[64] S. S. Skiena, "Sorting and Searching," in *The Algorithm Design Manual*. London, U.K.: Springer, 2012, pp. 103–144, doi: 10.1007/978-1-84800-070-4_4.

[65] N. Sick, N. Preschitschek, J. Leker, and S. Bröring, "A new framework to assess industry convergence in high technology environments," *Technovation*, vols. 84–85, pp. 48–58, Jun. 2019, doi: 10.1016/j.technovation.2018.08.001.

[66] M. Fernández, I. Cantador, V. López, D. Vallet, P. Castells, and E. Motta, "Semantically enhanced information retrieval: An ontology-based approach," *Web Semantics, Sci., Services Agents World Wide Web*, vol. 9, no. 4, pp. 434–452, Dec. 2011, doi: 10.1016/j.websem.2010.11.003.

[67] J. Risch, N. Alder, C. Hewel, and R. Krestel, "PatentMatch: A dataset for matching patent claims prior art," 2020, *arXiv:2012.13919*.

[68] N. Reimers and I. Gurevych, "Sentence-BERT: Sentence embeddings using Siamese BERT-networks," 2019, *arXiv:1908.10084*.

[69] K. A. Bryan, Y. Ozcan, and B. Sampat, "In-text patent citations: A user's guide," *Res. Policy*, vol. 49, no. 4, May 2020, Art. no. 103946, doi: 10.1016/j.respol.2020.103946.

[70] J. Kuhn, K. Younge, and A. Marco, "Patent citations reexamined," *RAND J. Econ.*, vol. 51, no. 1, pp. 109–132, Mar. 2020, doi: 10.1111/1756-2171.12307.

**JUHYUN LEE** received the B.S. degree in statistics from Cheongju University. He is currently pursuing the Ph.D. degree in industrial management engineering with Korea University. His main research interests include developing machine learning algorithms for both structured and unstructured data, such as text, signal, and image, and applying them to solve problems, such as predictive modeling, document classification, and sentiment analysis.

**SANGSUNG PARK** received the Ph.D. degree in industrial engineering from Korea University. He is currently an Assistant Professor with the Department of Big Data Statistics, Cheongju University, Republic of Korea. His main research interests are patent big data analysis, data mining and machine learning, technology management, and evaluation that combines various industrial engineering theories.

**JUNSEOK LEE** received the Ph.D. degree in industrial and management engineering from Korea University. He is currently working with Micube Solution Inc., Republic of Korea. His main research interests include developing a machine learning algorithm for detecting abnormal manufacturing, such as fault classification and applying the machine learning algorithm for technology management.