A Method for Assessing Patent Similarity Using Direct and Indirect Citation Links

Hsiao-Chun Wu†, Hung-Yi Chen‡, Kung-Yen Lee‡, and Ying-Chieh Liu†
* Department of Business Administration, Chaoyang University of Technology, Taichung, Taiwan (R.O.C.)
Email: sc524@cyut.edu.tw
† Department of Information Management, Chaoyang University of Technology, Taichung, Taiwan (R.O.C.)
Email: hychen39@cyut.edu.tw, allanliu@cyut.edu.tw
‡ Department of Engineering Science and Ocean Engineering, National Taiwan University, Taipei, Taiwan, ROC
Email: Kungyenster@gmail.com

Abstract—Assessing patent similarity is a fundamental and critical step in patent citation analysis. When evaluating a similarity among two patents, considering both direct and indirect citation links leads to more precise similarity assessment. This study proposes a method for assessing patent compound similarity that includes direct and indirect similarities. Given a direct similarity matrix that represents a patent citation network, the method calculates indirect similarity matrices and then obtains a compound similarity matrix. Keyword analysis in the text mining is employed to obtain a similarity for a pair of patents. In addition, two criterion are proposed for validating the compound similarities for the patent citation network.

Keywords—Citation link, keyword analysis, patent citation, patent similarity.

I. INTRODUCTION

The patent citation analysis uses patent citation links to extract useful information for technology management. The applications of the patent citation include patent classification, technology diffusion, technological changes and so on. In general, a similarity for each patent pair is evaluated according to their citing or cited prior arts in a citation analysis. Then, other procedures or methods, such as clustering or factor analysis, employ the similarities to extract other useful information.

Evaluating patent similarity is a critical task in patent citation analysis. As academic papers cite previous works for support, so do patents reference prior arts. Therefore, methods in bibliometrics have been widely applied to evaluate the similarity of patents. Bibliography coupling and co-citation are two popular methods. Bibliographic coupling is a method for measuring the similarity between academic papers. When a set of papers reference more than one of the same articles, they are similar in a certain degree. Co-citation is another method in bibliometrics for measuring the similarity between academic papers. When two papers appear simultaneously in many papers, there is a degree of relationship between the two papers. Examples of using co-citation analysis include [1], [2], [3], [4], and [5]. The main advantage of using bibliographic coupling or co-citation methods is that the citation information can discern the track of technological changes. However, the main shortcoming is that the two methods require lots of patents to evaluate the patent similarity. These two methods might not be appropriate for a technology area with only few patents. Keyword analysis in text mining provides another way to evaluate the similarity between patent pairs. Keywords are extracted from each patent. Then, the sets of keywords are used for evaluating similarity for each patent pair.

Some methods of patent citation analysis consider both direct and indirect citation links. An indirect citation link means two patents are connected by more than one intermediate patent. While direct citation links reveal related recent prior arts, indirect citation links reveal tracks of technological change. Considering direct and indirect citation links provides more information for assessing patent similarities.

This study proposes a method for evaluating patent similarity based on direct citation links and multi-stage indirect citation links. Keyword analysis is employed to assess the similarity of a patent pair with a direct citation link. Based on the similarities for direct citation links, matrix operations are used to compute indirect similarities for patents connected by indirect citation links. Adding the direct and indirect similarities together can identify compound similarities for patent pairs. This study also proposes two criterion to validate the resultant compound similarities. The advantages of the proposed method are twofold. Firstly, this method can be applied to technology areas with only a few patents. Instead of analyzing the citation links, the proposed method adopts keyword analysis to analyze keywords in the titles and abstracts of two patents to evaluate their similarity. The method does not require a large citation network to evaluate the similarities of patent pairs. Secondly, the direct and indirect similarities are considered in the proposed method to provide further information for patent classification and other applications.

The rest of the paper is organized as follows. Section two reviews the methods used to compute indirect similarities. Section three presents a method to calculate the compound similarities for patent pairs in which direct and indirect similarities are both considered. Section four provides criterion for validating the resultant compound similarities. Finally, conclusions are drawn in the last section.
II. LITERATURE REVIEW

Wartburg et. al. [5] proposed multi-stage patent citation analysis to measure inventive progress. When measuring the similarity (or proximity) between a patent and its set of directly cited prior patents, their method assumes that the new patent is based equally on the set of cited prior patents. So, if the new patent cites \( n \) prior patents, the similarity between the new patent and any cited prior patent is \( 1/n \), as shown in Figure 1a. Two patents connected through other patents form an indirect citation relationship. For example, patent \( A \) cites patent \( B \) and patent \( B \) cites patent \( C \). Then, patent \( A \) has an indirect citation relationship with patent \( C \). To measure the similarity between two patents with an indirect citation relationship, the strengths of the direct relationships connecting the two patents are multiplied together. For example, the similarity between patent \( A \) and \( B \) is \( 1/n \), and the similarity between patents \( B \) and \( C \) is \( 1/m \). Then, the similarity between patents \( A \) and \( C \) is \( 1/n \times 1/m \), as shown in Figure 1b.

Chang et. al.[6] proposed another method to measure the similarity between two patents. Consider two patents \( P \) and \( Q \) and there are \( n \) links connecting from \( Q \) to \( P \). Then the similarity between the two patents is measured by the Lineral Linkage Coefficient:

\[
R(P, Q) = \sum_{i=1}^{n} \alpha^{N(i)}, 0 < \alpha \leq 1. \tag{1}
\]

In the above equation, \( N(i) \) denotes the number of intermediate patents in the \( i \)th link; \( \alpha \) is a parameter that represents the marginal decline effect of each intermediate patent. In a word, the equation implies that the proximity between two patents decreases as the number of intermediate patents increases.

The main advantage of the methods of [5] and [6] is that it is easy to calculate the direct and indirect similarity. However, the main shortcoming is that setting the similarity to a proportion of the number of cited patents or to a constant cannot measure the proximity precisely. The prior arts cited by a patent might have different degrees of direct influences on the citing patent. Then, under that condition, the marginal decline effect would not be a constant along the indirect citation links. Therefore, the direct similarity between two patents should be carefully measured to provide a basis for obtaining the indirect similarity between two patents.

III. METHODOLOGY

The proposed method for calculating the compound similarities of patent pairs contains three steps. A compound similarity of a patent pair considers all similarity measures for the direct and indirect citation links between a patent pair. Details of the proposed method and an illustration of the method are given as follows.

A. Proposed Method

a) Step 1. Create a citation matrix: A matrix is created to represent the citation network of patents. Let \( i \) and \( j \) denote two patents. Notation \( c_{ij} \in \{0, 1\} \) denotes whether or not patent \( i \) cites patent \( j \). Then, the citation network for \( n \) patents can be represented by the citation matrix:

\[
C = [c_{ij}]_{n \times n}. \tag{2}
\]

b) Step 2. Create a direct similarity matrix: This step calculates the similarity between two patents given one patent directly refers to the other. Keywords in the title and the abstract of a patent are extracted to be the keyword vector for the patent. Let \( \vec{u}_i \) and \( \vec{u}_j \) denote two keyword vectors for patent \( i \) and \( j \) respectively. Then, the similarity between patents \( i \) and \( j \) is

\[
r_{ij} = F(\vec{u}_i, \vec{u}_j), \tag{3}
\]

where \( r_{ij} \in [0, 1] \) if \( c_{ij} = 1 \) and \( r_{ij} = 0 \) otherwise; and \( F(\cdot) \) is a similarity function. As a result, a direct similarity matrix can be created

\[
R = [r_{ij}]. \tag{4}
\]

In this study, we use the cosine measure for measuring the similarity between patents \( i \) and \( j \)

\[
cos(\vec{u}_i, \vec{u}_j) = \frac{\vec{u}_i \cdot \vec{u}_j}{\|\vec{u}_i\| \|\vec{u}_j\|}. \tag{5}
\]

c) Step 3. Generate a m-stage indirect similarity matrix: A multi-stage citation link for two patents indicates that one patent indirectly refer to the other through intermediate patents. For example, patent \( i \) cites patent \( j \) and patent \( j \) cites patent \( j \). Then, patent \( i \) indirectly refers to patent \( j \), which is a two-stage citation link. The similarity between two patents with a two-stage citation link can be obtained by the following equation:

\[
r_{ik}^{(2)} = \sum_{j=1}^{n} r_{ij} r_{jk}, \forall i, k. \tag{6}
\]
where \( r_{ij} \) denotes the direct similarity between patents \( i \) and \( j \). Therefore, we can have a two-stage indirect similarity matrix by

\[
R^{(2)} = \begin{bmatrix} r_{ik}^{(2)} \end{bmatrix} = R \times R. \tag{7}
\]

Therefore, the equation for a \( m \)-stage indirect similarity matrix is

\[
R^{(m)} = \prod_{s=1}^{m} R, \ m \leq n. \tag{8}
\]

**d)** Step 4. Generate a \( m \)-stage compound similarity matrix: The compound similarity for two patents considers all direct and indirect citation links between the two patents. The compound similarity for patent \( i \) and \( k \) is

\[
r_{ik}^{m} = \sum_{s=1}^{m} r_{ik}^{(s)}, \ m \leq n. \tag{9}
\]

Hence, a \( m \)-stage compound similarity matrix can be obtained by

\[
R^{m} = \sum_{s=1}^{m} R^{(s)}, \ m \leq n. \tag{10}
\]

**B. Illustration**

An example is given to demonstrate the calculation of the compound similarity for two patents. Consider the citation networks created by 8 patterns as shown in Figure 2. The numbers in the arc show a similarity for two patents. The directed arc represents the reference direction. The arrow of an arc points to a patent that is cited. For example, patent \( A \) cites patent \( B \) and the similarity between the two patents is 0.5.

In the example, patents \( A \) and \( B \) have direct and three indirect citation links. The direct similarity between the two patents is \( r_{AB} = 0.5 \). The two- to four-stage similarities between the two patents are:

\[
\begin{align*}
  r_{AB}^{(2)} &= 0.6 \times 0.5 = 0.3 \\
  r_{AB}^{(3)} &= 0.3 \times 0.4 \times 0.8 = 0.096 \\
  r_{AB}^{(4)} &= 0.9 \times 0.1 \times 0.8 \times 0.2 = 0.0144
\end{align*}
\]

Then the compound similarity for patents \( A \) and \( B \) is

\[
R_{AB}^{4} = r_{AB}^{(2)} + r_{AB}^{(3)} + r_{AB}^{(4)} = 0.5 + 0.3 + 0.096 + 0.0144 = 0.9104.
\]

We use matrix operations to obtain the compound similarity for each pair of patents in the example. The direct similarity matrix for the citation network is

\[
R = \begin{pmatrix}
0 & 0.5 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0.8 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0.5 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0.1 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0.8 \\
0 & 0.2 & 0 & 0 & 0 & 0 & 0 & 0
\end{pmatrix}.
\]

The two-stage indirect similarity matrix is

\[
R^{(2)} = R \times R = \begin{pmatrix}
0 & 0.3 & 0.12 & 0 & 0.09 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0.32 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0.08 \\
0 & 0.16 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0
\end{pmatrix}.
\]

With the same operation, we can have three- and four-stage indirect similarity matrices. Finally, the four-stage compound similarity matrix is

\[
R^{4} = R + R^{(1)} + R^{(2)} + R^{(3)} + R^{(4)} = \begin{pmatrix}
0 & 0.9104 & 0.3 & 0.12 & 0.6 & 0.9 & 0.09 & 0.072 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0.32 & 0 & 0.4 & 0 & 0 & 0 & 0 \\
0 & 0.8 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0.5 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0.016 & 0 & 0 & 0 & 0 & 0.1 & 0.08 \\
0 & 0.16 & 0 & 0 & 0 & 0 & 0 & 0.8 \\
0 & 0.2 & 0 & 0 & 0 & 0 & 0 & 0
\end{pmatrix}.
\]

**IV. Criteria for Validation**

To validate the proposed method for patent classification, two criteria are employed. The first criterion is the empirical results from [7]: the similarity for two patents decreases as the distance of the grant years for the two patents increases. Dahlin and Behrens [7] studied radical innovation in tennis-racket patents. The method of the co-citation analysis is used to measure the similarity between two patents. Their empirical results show that the patent similarities are distributed as in Figure 3. In the figure, the X-axis represents the distance between the grant years of a patent pair. The Y-axis is the similarity of the patent pair. The figure shows a pattern...
Fig. 3: Similarities against the distances in grant years for patent pairs [7]

Fig. 4: Using the UPC classification system to validate a m-stage compound similarity matrix

in which a patent pair has close grant years with high similarity.

The second criterion uses the UPC classification system for validation. The similarity between two patents belonging to the same UPC category should be higher than two patents from different UPC categories. For example, Figure 4 shows a compound similarity matrix for patent pairs in which patents are grouped according to their UPC codes: pure 438 or pure 257. The similarities in block I and III should be significantly larger than in block II and IV.

The above two criteria can be used to validate the m-stage compound similarity matrix. Outliers will be investigated further to identify reasons and possible corrections.

V. CONCLUSION

Assessing patent similarity is a core task in patent citation analysis. Considering both direct and indirect citation links provides more information for assessing patent similarity. In this study, a method for evaluating patent similarity that considers direct and indirect citation links is proposed. Keyword analysis is employed to obtain the direct similarity for a pair of patents. Then, with simple matrix operations, we have a m-stage indirect similarity matrix that represents the indirect similarity for each pair of patents in the citation network. Add the direct and several m-stage indirect similarity matrices to obtain a compound similarity matrix for the citation network. This study also proposed two criteria for validating the resultant compound similarity matrix. In the future, the proposed method will be validated empirically using patents from the semiconductor manufacturing industry.

ACKNOWLEDGMENT

This research is partially supported by National Science Council, Taiwan R.O.C. (NSC 98-2410-H-324-003-).

REFERENCES