Journal of Informetrics 12 (2018) 259–270

Contents lists available at ScienceDirect

Journal of Informetrics

journal homepage: www.elsevier.com/locate/joi

Missing links: Timing characteristics and their implications for capturing contemporaneous technological developments

Chung-Huei Kuan [a], Mu-Hsuan Huang [b], Dar-Zen Chen [c,*]

[a] Graduate Institute of Patent, National Taiwan University of Science and Technology, No. 43, Sec. 4, Keelung Rd., Taipei, 10607, Taiwan
[b] Department of Library and Information Science, National Taiwan University, No. 1, Sec. 4, Roosevelt Rd., Taipei, 10617, Taiwan
[c] Department of Mechanical Engineering and Institute of Industrial Engineering, National Taiwan University, No. 1, Sec. 4, Roosevelt Rd., Taipei, 10617, Taiwan

A R T I C L E I N F O

Article history:
Received 2 August 2017
Received in revised form 15 January 2018
Accepted 15 January 2018
Available online 22 February 2018

Keywords:
Missing link
Citation
Bibliographic coupling
Patent citation network
Main path analysis

A B S T R A C T

A *missing link* in this study refers to a pair of patents whose relatedness is not manifested by one citing the other but implied by their strong bibliographic coupling. By analyzing empirical data, this study discovers that the occurrence of missing links is not coincidental but arises systematically; patent pairs with missing links usually have highly overlapped application processes, whereas those with direct citations more frequently have successive or less overlapped application processes. The missing links thus may capture relatedness between patents that direct citations fail to detect. By applying main path analysis to a network containing 34,083 patents, 155,076 citations, and 9,213 missing links designed to simulate direct citations, this study further finds that the missing links—accounting for only approximately 5% of all connections—identify patents embodying contemporaneous technological developments, which may evade detection if only direct citations are considered.

1. Introduction

Investigating relatedness between the technological content of patents is central to the study of technological development. Such study may include mapping the technology landscape and monitoring technology evolution, evaluating the interaction between and impact of science and technology, observing technology spillover into other geographical areas or industry segments, and shaping patent owners' competitive goals and merger and acquisition strategies.

Direct citation (DC) (Hall, Jaffe, & Trajtenberg, 2005; Jaffe, Fogarty, & Banks, 1998; Trajtenberg, 1990) is perhaps the most widely applied tool in detecting patent relatedness. Taking its application in technology spillover as an example, a cited patent is regarded as containing the pieces of technical information that the citing patent is constructed upon; these pieces of technical information are deemed to flow from the cited to the citing patent holder's affiliated geographical areas (Figueiredo, Guimarães, & Woodward, 2015; Li, 2014; Murata, Nakajima, Okamoto, & Tamura, 2014) or industry segments (Karvonen & Kässi, 2013; Kim, Lee, & Sohn, 2016; Noailly & Shestalova, 2017).

Another popular relatedness tool is bibliographic coupling (BC) (Kessler, 1963), in which two patents are bibliographically coupled if they respectively cite at least one common document. The two patents' degree of relatedness is measured by the total number of commonly cited documents, referred to as the bibliographic coupling strength (BCS).

* Corresponding author.
E-mail addresses: maxkuan@mail.ntust.edu.tw (C.-H. Kuan), mhhuang@ntu.edu.tw (M.-H. Huang), dzchen@ntu.edu.tw (D.-Z. Chen).

https://doi.org/10.1016/j.joi.2018.01.005
1751-1577/

Both DC and BC may be utilized individually or with other relatedness tools. For example, Leydesdorff, Kushnir, and Rafols (2014) integrated DC with patent classification codes, and Nakamura, Suzuki, Sakata, and Kajikawa (2015) combined DC and co-word analysis. Kuusi and Meyer (2007) employed BC alone to cluster some related patents and identify an emerging technological paradigm in the field of carbon nanotubes. Lo (2007) also employed only BC to identify technological connections between major research organizations in the field of genetic engineering. Conversely, Von Wartburg, Teichert, and Rost (2005) combined DC and BC in a multistage analysis to reveal the technological change. Chen, Huang, Chen, and Lin (2012) used both DC and BC to construct citation networks among smart grid patents and observed the evolution of clusters of patents through a number of overlapping snapshots. Park, Jeong, Yoon, and Mortara (2015) used BC and patent text semantic analysis to locate potential research and development collaboration partners in the field of fuel cell membrane electrode assembly technology.

When employing both DC and BC to capture patent relatedness, situations may occur in which one tool indicates relatedness whereas another suggests otherwise. If one patent cites another patent, the two patents are said to form a DC pair, and if they are bibliographically coupled, they are said to constitute a BC pair. Then, if both DC and BC reflect patent relatedness, as shown in the aforementioned studies, it is interesting to notice that two patents often form a BC pair but not a DC pair.

This study was triggered by this type of BC-but-no-DC pair, particularly when a pair has high BCS, a situation which is referred to as a *missing link* (ML). Two patents form an ML pair if (1) they do not cite each other, (2) they are bibliographically coupled, and (3) they have high BCS. Therefore, an ML can only occur between two bibliographically coupled patents, but two bibliographically coupled patents do not always have an ML unless conditions (1) and (3) are satisfied. In other words, for an ML pair, their relatedness is not explicitly manifested by one directly citing the other, but strongly implied by the high BCS.

An ML pair example, two US utility patents, US8,622,222 and US8,623,202, were filed by the same company, one in January 2011 and the other in October 2012. Both were granted in January 2014 by different examiners. The two patents do not cite each other but have exceptionally high BCS of 1039 (US8,622,222 cited 1063 and US8,623,202 cited 1072 domestic and foreign patents and published applications). Both patents concern membrane bioreactor technologies and it is clear, even without examining their content, that they should be highly related. Similarly, US8,585,882 and US9,586,842, both concerning water treatment technologies, were filed by different companies, one in December 2008 and the other in December 2015. They were granted by different examiners in November 2013 and March 2017, respectively. Again, the two patents do not cite each other but have high BCS of 465 (US8,585,882 cited 472 and US9,586,842 cited 535 domestic and foreign patents and published applications); their relatedness is clearly reflected by their high BCS.

The usefulness of the ML is thus that it may be utilized to discover patent relatedness that escapes detection using DC. Chen, Huang, Hsieh, and Lin (2011) considered ML pairs as "missing citations" and used them together with DC pairs to construct comprehensive clusters of patents. Yeh, Sung, Yang, Tsai, and Chen (2013), in addition to supplementing the ML pairs in a patent citation network (PCN), further considered DC pairs with BCS less than a threshold as unreliable and removed them from the PCN.

Based on this literature review, the present study intends to contribute to the discussion of patent relatedness by utilizing empirical data to address the following issues: (1) why MLs occur and whether they are simply coincidences, (2) what useful information may be derived from the relatedness captured by MLs if their occurrence is not coincidental, and (3) how MLs may be utilized to capture this useful information.

2. Data

This study selects for empirical analysis patents in the field of carbon dioxide capture, storage, recovery, delivery, and regeneration and collects a total of 34,083 US utility patents issued between 1976/1/1 and 2017/3/31 by the United States Patent and Trademark Office database. These patents contain at least one specific keyword[1] in at least one relevant field (i.e., Title, Abstract, Specification, or Claims) and at least one specific Cooperative Patent Classification symbol prefix.[2]

Among the 34,083 patents, there are 155,076 DC and 1,609,549 BC pairs. From their sheer volume, BC appears much noisier than DC. The BC pairs have a significantly skewed BCS distribution with a mean (μ) of 2.74, a standard deviation (σ) of 15.66, and a maximum of 1,123. Among the BC pairs, 72.55% (1,167,794) have the smallest BCS of 1, again suggesting that BC is relatively noisy. Among the 1,609,549 BC pairs, 75,700 are also DC pairs, and these pairs have much higher mean BCS (9.56) than the overall average. Therefore, a simultaneous DC and BC relationship indeed reflects a greater degree of relatedness between patents.

A design decision of this study is the use of a threshold to determine MLs. Swanson (1971) and Jarneving (2007) indicated that only BC pairs having BCS more than a threshold are truly related. Chen et al. (2011) used the mean BCS for pairs having a simultaneous DC and BC relationship to define a threshold, whereas Yeh et al. (2013) used the mean BCS of BC pairs without DC as a threshold. This study employs a much more conservative threshold, equal to the mean BCS plus two times the

[1] The keyword search command was '(carbon or dioxide$ or co2) AND (storage$ or captur$ or recover$ or deliver$ or regenerat$),' where '$' is the wildcard character.

[2] These CPC symbol prefixes are B63 B 35$, C01 B 3$, C01B31/20, C01 B 21/22, C02F 1$, C07C 7/10, F01N 3/10, F25J 3/02, B01J 20$, B01D 53$, and B01D 11, where '$' is the wildcard character.