Journal of Informetrics 10 (2016) 1108–1130

Contents lists available at ScienceDirect

Journal of Informetrics

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

A hybrid similarity measure method for patent portfolio analysis

Yi Zhang [a], Lining Shang [b], Lu Huang [b,*], Alan L. Porter [c], Guangquan Zhang [a], Jie Lu [a], Donghua Zhu [b]

[a] Decision Systems & e-Service Intelligence research Lab, Centre for Quantum Computation & Intelligent Systems, Faculty of Engineering and Information Technology, University of Technology Sydney, Australia
[b] School of Management and Economics, Beijing Institute of Technology, Beijing, PR China
[c] Technology Policy and Assessment Centre, Georgia Institute of Technology, Atlanta, USA

ARTICLE INFO

Article history:
Received 9 December 2015
Received in revised form
25 September 2016
Accepted 25 September 2016

Keywords:
Patent analysis
Similarity measure
Text mining
Bibliometrics

ABSTRACT

Similarity measures are fundamental tools for identifying relationships within or across patent portfolios. Many bibliometric indicators are used to determine similarity measures; for example, bibliographic coupling, citation and co-citation, and co-word distribution. This paper aims to construct a hybrid similarity measure method based on multiple indicators to analyze patent portfolios. Two models are proposed: categorical similarity and semantic similarity. The categorical similarity model emphasizes international patent classifications (IPCs), while the semantic similarity model emphasizes textual elements. We introduce fuzzy set routines to translate the rough technical (sub-) categories of IPCs into defined numeric values, and we calculate the categorical similarities between patent portfolios using membership grade vectors. In parallel, we identify and highlight core terms in a 3-level tree structure and compute the semantic similarities by comparing the tree-based structures. A weighting model is designed to consider: 1) the bias that exists between the categorical and semantic similarities, and 2) the weighting or integrating strategy for a hybrid method. A case study to measure the technological similarities between selected firms in China's medical device industry is used to demonstrate the reliability our method, and the results indicate the practical meaning of our method in a broad range of informetric applications.

1. Introduction

Patent statistics serve as an important indicator of the activities and outcomes of research & development (R&D) (Tseng, Lin, & Lin, 2007). Analyzing patents and patent portfolios is increasingly contributing to academic research, public policy, and business intelligence. Such analysis can: reveal emphasis in science, technology, & innovation (ST&I) endeavours across fields of research (Porter & Detampel 1995); determine who is engaging in what research and to what extent (e.g., organizations, regions, and countries), and add value to collaborative relationships (Porter & Newman 2011); and provide further insights into a wide range of applications, e.g., evaluating the impact of national patent regimes on technology transfer

* Corresponding author.
E-mail addresses: yizhangbit@gmail.com (Y. Zhang), sln_work@163.com (L. Shang), huanglu628@163.com (L. Huang), alan.porter@isye.gatech.edu (A.L. Porter), jie.lu@uts.edu.au (J. Lu), zhudh111@bit.edu.cn (D. Zhu).

http://dx.doi.org/10.1016/j.joi.2016.09.006
1751-1577/

(Intarakumnerd & Charoenporn 2015), identifying potential business opportunities or development trends (Fabry et al., 2006; Zhou, Zhang, Porter, Guo, & Zhu, 2014), mapping the R&D landscape and monitoring technological structures (Choi & Park 2009), and pinpointing patent strategies that may help shape overall business goals (Su, Lai, Sharma, & Kuo, 2009).

Similarity measures are fundamental tools for identifying relationships within or across patent portfolios. Many bibliometric indicators are used to investigate such analyses; for example, bibliographic coupling (Kessler, 1963), citation (Garfield, Sher, & Torpie, 1964), co-citation (Small, 1973), and co-word distribution (Callon, Courtial, Turner, & Bauin, 1983). Additionally, combining several indicators in patent analysis is currently popular, e.g., blending citations with international patent classification (IPC) codes (Kay et al., 2014; Leydesdorff, Kushnir & Rafols 2014), bibliographic coupling (Chen, Huang, Hsieh, & Lin, 2011), or co-word analyses (Nakamura, Suzuki, Sakata, & Kajikawa, 2015). As a traditional mainstream bibliometric indicator, citations and co-citations connect scientific documents via forward and backward links. These direct relationships can easily identify similarities between documents (Zhang, Zhang, Zhu, & Lu, 2016), but not all patent databases provide citation information. Usually, patents only cite references that are directly relevant, and some of them are non- patent documents (Rip 1988). Therefore, patent citations will take patents and scientific publications into consideration, and related analysis can be more complex than expected.

As a unique feature of patents, IPC codes provide a hierarchical taxonomy system to reflect the categories and subcategories of existing technologies. This benefit makes IPCs favorable for similarity measures, and co-classification analysis is commonly applied (Boyack & Klavans 2008). The IPC system is, however, a "vague" classification system, since it defines new and emerging technologies using existing technologies or combinations of them. But, it is not always easy to classify one invention according to existing definitions, and conservative assignments can lead to uncertainty.

For a long time, text elements (e.g., words, terms, and phrases) have acted as a supplement to citations and IPCs in patentometrics. The rapid development of natural language processing (NLP) and data cleaning techniques have enhanced the ability to retrieve precise text elements from patents. Text-based similarity measures follow the general idea of co-word analysis, in which patents are seen as similar if there is a high degree of common textual elements between two or more patents (Moehrle, 2010). However, these free text elements are much more complex than human-defined IPCs. The semantic meanings of text elements and the potential relationships among them heavily depend on the language environment. Diverse combinations of text elements also add difficulties (Zhang, Zhang, Zhu, & Lu, 2016). At the same time, traditional co-word analysis exaggerates the importance of term frequency (Peat & Willett 1991), and even the efficiency of term frequency inverse document frequency (tf-idf) analysis is debated (Zhang, Zhou, Porter, & Gomila, 2014).

In an attempt to address the above concerns, our two research questions are: 1) how should a hybrid similarity measure method for patent portfolio analysis with multiple indicators be constructed? And 2) how should significant terms be identified and weighed to improve the performance of similarity measures? This paper emphasizes both IPCs and text elements, and specifically divides the technological similarity between patent portfolios into two forms: categorical similarity and semantic similarity.

We introduce fuzzy set routines (Zadeh, 1965) to translate the rough technological categories and sub-categories of IPCs into defined numeric values, and calculate categorical similarity via vectors that consist of membership grades. In parallel, we use an algorithm to group terms into clusters, and represent a patent portfolio in a 3-level tree. The tree structure consists of the patent portfolio's terms and their clusters, and semantic similarity is determined by comparing two trees. We have also developed a model that considers the two major weighting issues in our method: bias in the two similarities and the strategy of integrating them, and also the weights of matching types in a tree-based comparison.

An empirical study to measure the technological similarities between selected firms in China's medical device industry demonstrates the feasibility and performance of our method. A specific case study that focuses on the unexpected results between expert marks and our method further endorses our methods' reliability and efficiency in helping experts discover the underlying technological relationships between patent portfolios. The results inform related patent portfolio analyses in a broad range of applications, e.g., general topic analysis for technical intelligence, patent mapping, and technology mergers and acquisitions. The main contributions of this paper include: 1) a hybrid measure method that combines categorical IPC-driven similarity and semantic text-based similarity measures; 2) an effective application of fuzzy sets to transform vague IPC categories into defined numeric values; and 3) a semantic tree structure to identify and highlight significant terms in an interactive hierarchical model for similarity measures.

This paper is organized according to the following structure. We review previous studies in Section 2. Section 3 follows and presents our hybrid similarity measure method for patent portfolio analysis. In Section 4, we use our method to measure the technological similarities between selected firms in China's medical device industry from the Web of Science's Derwent Innovation Index (DII) patent database. Finally, we provide an in- depth discussion on the strengths and weaknesses of the categorical and semantic similarity measures, possible applications, limitations, and future directions of our method in Section 5.

2. Related work

This paper reviews previous literature from two categories: bibliometric similarity measures and related techniques; and indicators for bibliometrics and patentometrics.