This paper has been accepted by *Journal of Informetrics*. Please cite as: Zhang, Y., Shang, L., Huang, L., Porter, A. L., Zhang, G., Lu, J. & Zhu, D. 2016, A hybrid similarity measure method for patent portfolio analysis, *Journal of Informetrics*, 10(4), 1108-1130, DOI: 10.1016/j.joi.2016.09.006.

A hybrid similarity measure method for patent portfolio analysis

Yi Zhang[1], Lining Shang[2], Lu Huang[2, *], Alan L. Porter[3], Guangquan Zhang[1], Jie Lu[1], Donghua Zhu[2]

[1]Decision Systems & e-Service Intelligence research Lab, Centre for Quantum Computation & Intelligent Systems, Faculty of Engineering and Information Technology, University of Technology Sydney, Australia;

[2]School of Management and Economics, Beijing Institute of Technology, Beijing, P. R. China;

[3]Technology Policy and Assessment Centre, Georgia Institute of Technology, Atlanta, USA;

Email Addresses: yizhangbit@gmail.com; sln_work@163.com; huanglu628@163.com (corresponding email address); alan.porter@isye.gatech.edu; jie.lu@uts.edu.au; zhudh111@bit.edu.cn.

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.

Keywords Patent analysis; Similarity measure; Text mining; Bibliometrics;

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 (Intarakumnerd & Charoenporn 2015), identifying potential business opportunities or development trends (Fabry et al. 2006; Zhou et al. 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 et al. 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 et al. 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 et al. 2011), or co-word analyses (Nakamura et al. 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 et al. 2016b), 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 sub-categories 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 et al. 2016b). 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 et al. 2014b).

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.

2.1. Bibliometric similarity measures and related techniques

Measuring similarities among bibliometric units (e.g., journals, patents, authors, or words) is a central task in bibliometrics (Klavans & Boyack 2006). Despite a series of techniques that have been introduced to investigate similarity measures in text mining and related information technology (IT) fields, e.g., corpus- and knowledge-based approaches, and ontology (Lau, Tsui & Lee 2009; Wu, Lu & Zhang 2011; Sánchez et al. 2012), the main analytic approach for bibliometric similarity measures is still vector-based – a data corpus is represented by a vector (Boyack, Klavans & Börner 2005). For the past few decades, Salton's cosine (Salton & Buckley 1988) and Jaccard's index (Braam, Moed & Van Raan 1988) have become two popular techniques for deriving similarity measures. Although a number of studies compared their performance (Hamers et al. 1989; Leydesdorff 2008; Moehrle 2010), debate over which is best continues. Significantly, a boom of research into science maps and the exponential growth of bibliometric data introduce both challenges and opportunities for measuring bibliometric similarity. Based on the inter-citation/co-citation distributions in a 1-million-record dataset, Klavans & Boyack (2006) compared six approaches for measuring similarity in science maps – raw frequency, cosine, Jaccard's index, Pearson's coefficient, average relatedness factor (Pudovkin & Garfield 2002), and their proposed normalized frequency measure, K50. Cosine and K50 with inter-citations was recommended as performing better in the experiments. Boyack et al. (2011) also compared the accuracy of several text-based approaches to similarity measures, e.g. tf-idf, latent semantic analysis, and topic models, and from a technical perspective this comparison provides insights for further topic analysis in bibliometrics.

2.2. Indicators for bibliometrics and patentometrics

General indicators for bibliometrics include citations, co-citations, terms and phrases, bibliographic coupling, etc. Ahlgren & Colliander (2009) compared each of indicators in a small dataset of only 43 papers. Their dataset is too small to easily rank the indicators or the related approaches in any general situation.

Over the last several decades, citation and co-citation analyses have become the most representative indicators, and have been widely applied to various bibliometric studies, e.g., topic analysis (Chen, Ibekwe SanJuan & Hou 2010), science maps (Braam, Moed & Van Raan 1991; Boyack, Klavans & Börner 2005; Klavans & Boyack 2009), and trend analysis (Garfield, Paris & Stock 2006; Lucio-Arias & Leydesdorff 2008; Chen et al. 2012). As a significant subsequent study, Klavans & Boyack (2016) specifically compared the accuracy of direct-citation and co-citation for representing knowledge taxonomy, and clarified that direct-citation analysis shows better performance (Klavans & Boyack 2006).

Text elements (e.g., words, terms, and phrases) are another important bibliometric resource that plays an active role in topic analysis (Wang et al. 2014; Zhang et al. 2016a), bibliometric maps (Noyons & van Raan 1998; Zhu & Porter 2002; van Eck et al. 2010), and trend analysis (Zhou et al. 2014; Zhang et al. 2016b).

However, since most bibliometric data is unlabeled, it is always not easy to evaluate the accuracy of term-based analysis. Existing validation measures include: 1) constructing labeled training sets based on expert knowledge (Harman & Voorhees 2006) or manual validation (Zhou et al. 2013; Huang et al. 2015); 2) introducing specific information as the label, e.g., subject categories in the Web of Science's databases (Yau et al. 2014), grant acknowledgements of the Medline database (Boyack et al. 2011), and program categories of academic proposals (Zhang et al. 2016a); and 3) focusing on the ratio of similarity within a cluster and the similarity between clusters, in which a higher ratio denotes better performance (Kassab & Lamirel 2008). Although this option provides a good design for unsupervised environments, it has not been widely used.

Similar to bibliometrics, quantitative studies with patent information can be called patentometrics (Stock & Stock 2013). Patent similarity measures follow the traditions of bibliometric similarity measures. However, considering that patents usually focus on technologies and products, the similarity between patents is usually defined in terms of technological distance or technological proximity (Jaffe 1986). IPC is a highly distinctive indicator for measuring the technological similarity in patentometrics. The co-occurrence distributions in IPCs are specified as co-classifications, and bind with vector-based similarity measures (Boyack & Klavans 2008). The basic rule is that patents in a given category should be more similar to each other than to those in other categories (Jaffe 1986). There is a trend to blend IPCs with other indicators in patentometrics. For example, Kay et al. (2014) and Leydesdorff, Kushnir & Rafols (2014) both constructed a kind of overlay map with IPCs and co-citation/citation distributions respectively. Furthermore, IPC is also popular in studies on automated patent classification (Fall et al. 2003; Chen & Chang 2012). Text elements are usually combined with certain criteria, such as TRIZ principals, subject-action-object structures, or problem/solution patterns (Cong & Tong 2008; Hu, Fang & Liang 2013; Yoon, Park & Kim 2013). Applications for patent classification in diverse languages have also been addressed (Fall et al. 2004; Kim & Choi 2007).

3. Methodology

We aim to construct a hybrid similarity measure method for patent portfolio analysis that includes both IPCs and text elements. The method includes: a categorical similarity measure model based on IPC categories and sub-categories; a semantic similarity measure model based on tree structures, and a weighting model. The framework of our method is shown in Fig. 1.

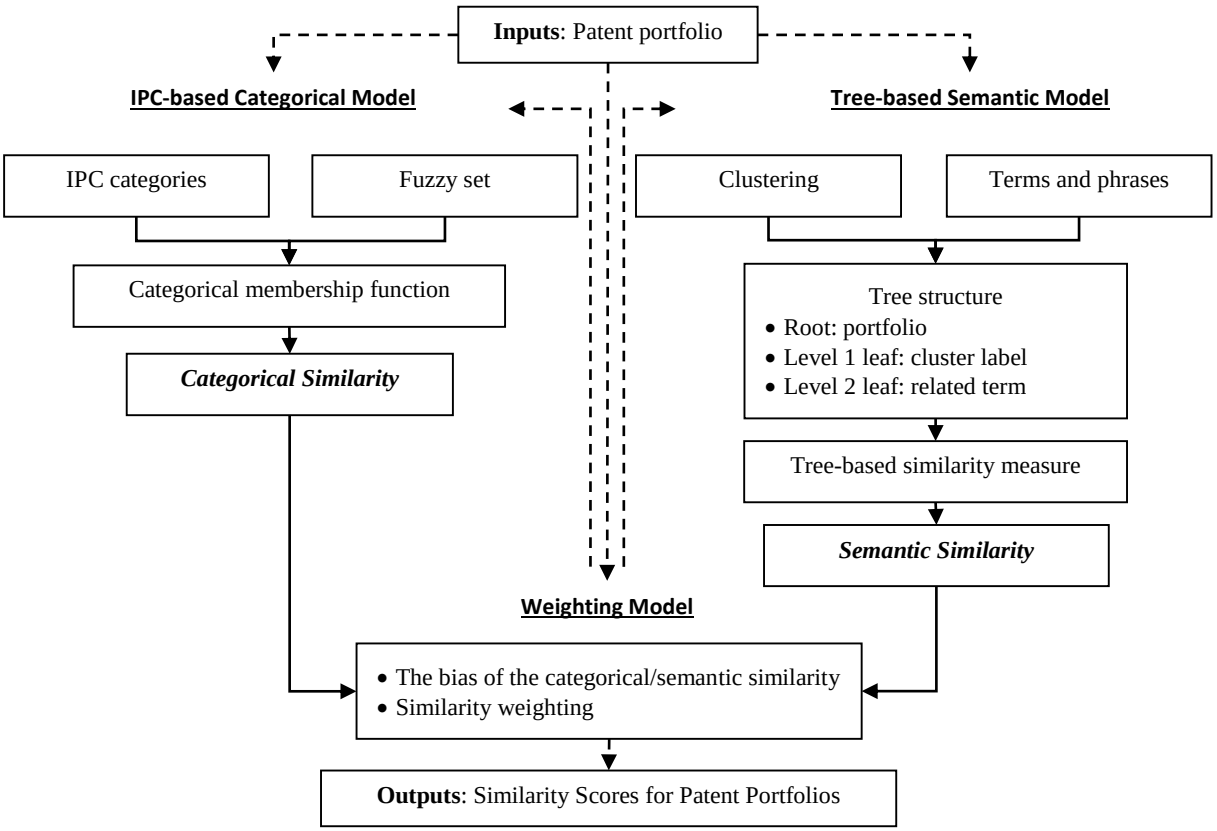

Fig. 1. The framework of the hybrid similarity measure method for patent portfolio analysis.

3.1. Inputs: Patent portfolio

This paper defines a patent portfolio as a group of patents with similar features, i.e., patents that involve similar technical topics, or belong to the same entity – individual, organization, region, country, etc. Here we simply assume a patent portfolio is a patent corpus.

3.2. The IPC-based categorical model

IPC is a core element in this section, so we first take effort to justify the question: Which kind of IPCs is the best for our method: 3-digit, 4-digit, or 7-digit? Early studies of patent maps mostly used 3- or 4-digit IPCs for classification, since they can act as the basic categories or sub-categories of a classification system (Boyack & Klavans 2008; Leydesdorff, Kushnir & Rafols 2014). Kay et al. (2014) proposed a multi-level aggregation process involving all 3-digit, 4-digit, and 7-digit IPCs to ensure broad data coverage for patent maps. This approach makes sense for research that focuses on a wide range of scientific subjects; however, patent portfolio analysis sometimes concentrates on a specific subject or technology, e.g., a comparison between two companies in a given technological area. In that situation, almost all patents would belong to the same one or two 3- or 4-digit IPCs and thus they would be too general to describe detailed technological components. Our method therefore uses 7-digit IPCs as its default setting with the aim of capturing multiple dimensional features. It is worth noting that 3 or 4-digit IPCs could be an option for multi- or interdisciplinary studies.

No matter what kind of IPCs are chosen – 3, 4, or 7 digits – the classifications they provide are still somewhat vague, and will not precisely represent the detail in technology. For example, if portfolio *P* contains IPC *A* (with frequency 10) and IPC *B* (with frequency 1), it is easy to say that portfolio *P* investigated the technologies of IPC *A*, however the low frequency of IPC *B* makes a simple 'involved' or 'not involved' classification difficult. It would make better sense to say, "IPC *B* was investigated a little bit", but that leaves users with the trouble of deriving a numeric value to describe "a little bit". Fuzzy sets may be an efficient solution to this problem (Zadeh 1965), and they have already been applied in a wide range of studies in the

fields of information science, decision science, and management science (Ma, Zhang & Lu 2012; Zhang et al. 2013). We introduce fuzzy set routines to deal with this kind of ambiguity and translate nuanced human knowledge into defined numeric values.

We denote all patent portfolios as the universe $X = \{x_1, x_2, \dots, x_i, \dots, x_{n-1}, x_n\}$, where n is the total number of portfolios. Given that each IPC is defined as a fuzzy set A_j, where $j \in [1, m]$ and m is the total number of IPCs, the membership function $A_j(x_i)$ represents the degree to which portfolio x_i investigates IPC A_j. Generally, there are many approaches to define membership functions – e.g., machine learning techniques or manual configuration based on expert knowledge – but using machine learning techniques to train a membership function, given the complexity of ST&I data, is still an unsolved task in the field of computer science and information systems. In this model, we use expert knowledge to decide the membership function based on training sets and experiments.

Once the membership function is decided, each patent portfolio is represented by an m-dimensional vector $V(x_i) = \{\vartheta_{1,i}, \vartheta_{2,i}, \dots, \vartheta_{j,i}, \dots, \vartheta_{m-1,i}, \vartheta_{m,i}\}$, in which $\vartheta_{j,i}$ is the membership grade that portfolio x_i belongs to the fuzzy set of IPC A_j. We then use the cosine measure (Salton & Buckley 1988) to calculate the categorical similarity between two patent portfolios x_i and x_k as follows:

$$CS(x_i, x_k) = \frac{V(x_i) \cdot V(x_k)}{|V(x_i)||V(x_k)|}$$

where $|V(x_i)|$ is the norm of the vector $V(x_i)$ and can be calculated below:

$$|V(x_i)| = \sqrt{(\vartheta_{1,i})^2 + (\vartheta_{2,i})^2 + \cdots + (\vartheta_{m,i})^2}$$

3.3. The tree-based semantic model

Traditional text similarity measures (e.g., vector-based approaches) simply use the frequency of raw terms to calculate the similarity between records. Sometimes the tf-idf analysis is used to weight terms. There is no doubt that weighting terms is an effective way to highlight important topics, but it is critical to directly identify the terms with a high tf-idf value terms as important terms (Zhang et al. 2014a). This section aims to propose another way: a 3-level tree to distinguish the importance of textual elements in a hierarchical structure, as shown in Fig. 2.

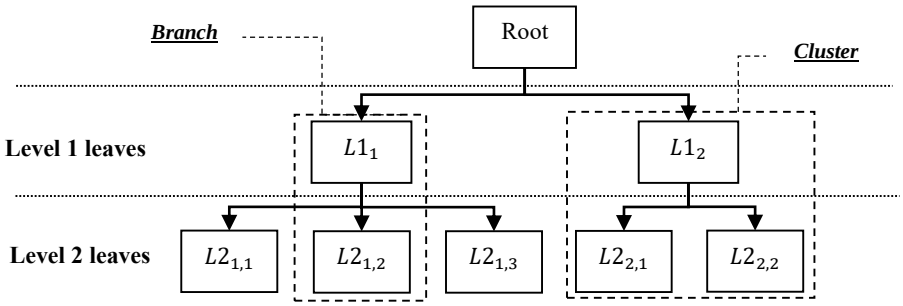

Fig. 2. The 3-level tree structure for the semantic model.

A brief outline of the stepwise process to construct a 3-level tree structure is given below. More details can be found in Appendix A.

1) Pre-processing: Retrieve core technological terms from a corpus via a term clumping process (Zhang et al. 2014a) and construct a portfolio-term matrix to link patent portfolios with contained core terms. One patent portfolio is represented as a tree, and its involved terms are the leaves of the tree.

2) Identifying Level 1 and Level 2 leaves: Apply a clustering algorithm to each patent portfolio separately. Group the core technological terms of a portfolio into several clusters, and identify the term with the

highest prevalence value in one cluster as a Level 1 leaf (Zhang et al. 2014b). Remaining terms in the cluster are linked to the Level 1 leaf as Level 2 leaves.

3) Tree construction: One tree represents one patent portfolio. It includes a root, a number of Level 1 leaves linked directly to the root, and a number of Level 2 leaves that are linked as normal to the Level 1 leaves.

As an example, given that a 3-level tree structure T contains p clusters and the p-th cluster has q_p terms. We denote the tree T as

$$T = \{B(L1_1, L2_{1,1}), B(L1_1, L2_{1,2}), \dots, B(L1_1, L2_{1,q_1}), \dots, B\left(L1_p, L2_{p,q_p}\right)\}$$

where $B(L1_i, L2_{i,j_i})$ $(i \in [0,p], j \in [1,q_i])$ is one branch of the tree. It consists of one Level 1 leaf $L1_i$, and one Level 2 leaf $L2_{i,j_i}$.

Note that $L1_i \neq L2_{i,j_i}$, and in some extreme situations $L1_i$ can be null and then $L2_{i,j_i}$ is linked directly to the root. An example of the 3-level tree T is given in Table 1.

Table 1

An example of the 3-level tree *T*

Patent Portfolio (Root)	*Level 1 Leaf*	*Level 2 Leaf*
Tree *T*	$L1_1$	$L2_{1,1}$
		$L2_{1,2}$
		…
		$L2_{1,q_1}$
	$L1_2$	$L2_{2,1}$
		$L2_{2,2}$
		…
		$L2_{2,q_2}$
	…	…
	$L1_p$	$L2_{p,1}$
		…
		$L2_{p,q_p}$

The semantic model follows the 3-level tree structure to measure the similarity between patent portfolios. We calculate the similarity by comparing the branches of the tree structures representing these portfolios. We traverse and compare all branches in both trees and exhaustively list the six matching types $R_x(B,B'), x = \{1,\dots,6\}$ in Table 2. Typically, it is easy to definitively determine that a match across all the leaves of an entire branch in Level 1 and Level 2 is a priority. Since we always highlight the representativeness of a Level 1 leaf in one cluster, we also imagine that a Level 1 leaf-based match is better than a Level 2 leaf-based match. We specifically discuss bias in Section 3.4 and Section 4.

Table 2

The six matching types between two trees

No.	*Matching Type*	*Description*	*Weight*
1	$R_1(B,B'):\{L1_i = L1'_{i'}, L2_{ij} = L2'_{i'j'}\}$	The entire branches are the same;	w_1
2	$R_2(B,B'):\{L1_i = L1'_{i'}, L2_{ij} \neq L2'_{i'j'}\}$	Only the Level 1 leaves are the same;	w_2
3	$R_3(B,B'):\{L1_i \neq L1'_{i'}, L2_{ij} = L2'_{i'j'}\}$	Only the Level 2 leaves are the same;	w_3
4	$R_4(B,B'):\{L1_i \neq L1'_{i'}, L2_{ij} \neq L2'_{i'j'}, L1_i = L2'_{i'j'}, L2_{ij} \neq L1'_{i'}\}$	Only a Level 1 leaf matches a Level 2 leaf in the other tree;	w_4
5	$R_5(B,B'):\{L1_i \neq L1'_{i'}, L2_{ij} \neq L2'_{i'j'}, L1_i \neq L2'_{i'j'}, L2_{ij} = L1'_{i'}\}$	Only a Level 2 leaf matches a Level 1 leaf in the other tree;	w_5
6	$R_6(B,B'):\{L1_i \neq L1'_{i'}, L2_{ij} \neq L2'_{i'j'}, L1_i = L2'_{i'j'}, L2_{ij} = L1'_{i'}\}$	A Level 1 leaf matches a Level 2 leaf in the other tree, and its Level 2 leaf matches the Level 1 leaf in the other tree.	w_6

Note. The six matching types include all possible matches between two trees.

Note that our method follows the basic rule of traditional tree-based similarity measurement, i.e., tree traversal, but we replace node-based comparisons with branch-based comparisons to maximize the use of significant terms. We also take term frequency and the number of branches into consideration. Referring to Table 2, we define $SS(P1,P2)$ as the semantic similarity value between portfolio $P1$ and $P2$. $SUM(R_k)$ is the total number of the matching type R_k between $P1$ and $P2$. $\Gamma(R_k)$ is the frequency of the terms involved in matching type R_k. Φ_{P1} and Φ_{P2} are the total number of the branches that portfolio $P1$ and $P2$ have respectively, and $\Phi_{P1,P2}$ is the total number of matched branches between the two portfolios. Therefore, $SS(P1,P2)$ can be calculated as follows:

$$SS(P1,P2) = SS(T_{P1},T_{P2}) = \frac{2\Phi_{P1,P2}}{\Phi_{P1}+\Phi_{P2}} \sum_{k=1}^{6} \left(w_k \times \frac{SUM(R_k)}{\sum_{k=1}^{6} SUM(R_k)} \times \frac{\Gamma(R_k)}{\sum_{k=1}^{6} \Gamma(R_k)}\right)$$

$$R_k = \begin{cases} 1, \text{if } R_k(B,B') \text{ matches} \\ 0, \quad \text{else} \end{cases}$$

An example of two trees and their branches is given in Fig. 3. The two trees can be described as: $T1 = \{B(L1_1, L2_{1,1}), B(L1_1, L2_{1,2}), B(L1_2, L2_{2,1})$ and $T1' = \{B(L1'_{1'}, L2'_{1'1'}), B(L1'_{2'}, L2'_{2'1'}), B(L1'_{3'}, L2'_{3'1'})\}$, and each leaf consists of a term in the set {*t1*, *t2*, *t3*, *t4*, *t5*, *t6*}.

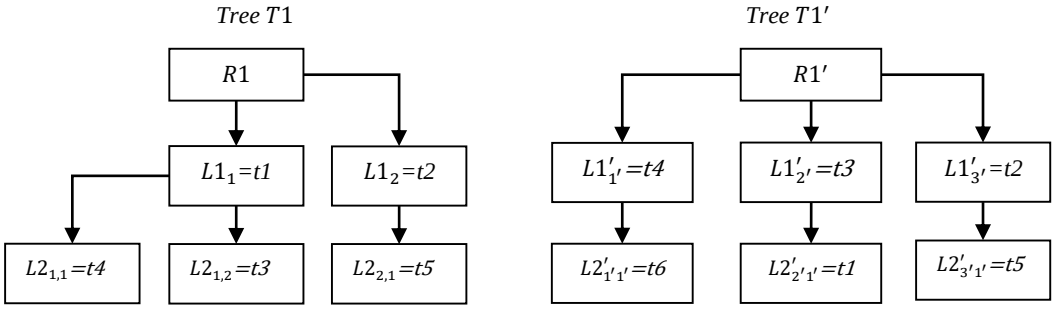

Fig. 3. Samples of the two trees and their branches.

We list all matching types between $T1$ and $T1'$ in Table 3. Note that: 1) if there is more than one match between two branches, we only choose the match with the highest weight; and 2) if two or more matches share the same weight, we prefer the match with a higher term frequency.

Table 3

The matches between $T1$ and $T1'$ in Fig. 3

No	*Branch of* $T1$	*Matching Branch of* $T1'$	*Matching Type*	*Note*
1	$B(L1_1, L2_{1,1})$	$B(L1'_{1'}, L2'_{1'1'})$	R_5	Only the best match is chosen, i.e., the matching type with higher weights.
		$B(L1'_{2'}, L2'_{2'1'})$	R_4	
2	$B(L1_1, L2_{1,2})$	$B(L1'_{2'}, L2'_{2'1'})$	R_6	-
3	$B(L1_2, L2_{2,1})$	$B(L1'_{3'}, L2'_{3'1'})$	R_1	-

Compared to traditional tree-based approaches, which simply depend on term frequency, we focus on the proportion of different matching types and introduce expert knowledge to weight these types. This can improve its adaptability to diverse practical requirements. However, while we retain term frequency as an important indicator, we also introduce several effective approaches to remove noise and consolidate synonyms. These efforts further improve the performance of term-based analyses.

We also take the frequency of branches and the proportion of matched branches into consideration. One branch usually contains two distinct terms, and we set up three options to calculate the frequency of a branch. 1) We use the frequency of the Level 1 leaf to represent the frequency of the branch, which highlights the importance of Level 1 leaf. 2) We aggressively use the larger frequency between the Level 1 leaf and the Level 2 leaf; and 3) to conservatively use the smaller frequency between the two leaves. Generally we chose the first option as a default setting. The proportion of matched branches is to highlight the importance of the matched branch in the entire tree structure, and it is also able to distinguish the following situation: as given in Fig. 3, there is one more tree $T1''$, which contains $T1'$ but has much more branches, so the engagement of the proportion of matched branches can make sure $SS(T1, T1')$ is larger than $SS(T1, T1'')$, since we believe $T1'$ concentrates on this technology more than $T1''$.

3.4. Weighting model

We use the weighting model for flow design and weight configuration, but emphasize that this configuration needs to take actual requirements into consideration. The two main objectives of this section include: the weights of the match types in the comparison between tree structures, and the biases between categorical and semantic similarities.

The weights of the matching types in the tree comparison reflect bias in the leaves at different levels. It is promising that a match with a Level 1 leaf is better than a match with a Level 2 leaf, and a match with one entire branch is the best. Therefore, one strategy for weighting the six matching types is: $w_1 > w_2 > w_3$ and $w_6 > w_4 \geq w_5$. Note that similarity measures are usually indirect, but if they are direct, i.e., the similarity between portfolio A and B is different from the similarity between B and A, it is reasonable that $w_4 > w_5$, otherwise we prefer to make $w_4 = w_5$. However, the strategy for comparing a match type and its inverse (e.g., w_1 and w_6) is notable. It is better to leave space for discussion depending on actual cases and situations, but, based on our experience and the general requirements of patent analysis, we provide the default setting: $w_1 > w_2 \sim w_6 > w_4 = w_5 > w_3$. Following this priority, an analytic hierarchy process (AHP) (Saaty 1990) can be a feasible option to engage expert knowledge to decide the weights with consideration on actual needs.

The bias in categorical and semantic similarities is a complicate issue, and varies from case to case. Categorical similarity, derived from IPCs, indicates the technological map of a patent portfolio, which can be a part of the portfolio's R&D strategies. Terms can be used effectively to explore technologies and their components in detail, which is promising for specific technology-oriented topic analysis, evaluation, and

forecasting studies. However, a cleaning process is important to ensure accurate semantic similarity measures. Thus, the categorical model performs better in a raw dataset, while the semantic model makes great sense with deeply-cleaned data situation, and we expect benefits if both methods can be combined. One strategy for using both models to complement each other within the flow is to apply categorical similarity measures to pre-processing and filtering, and use semantic similarity measures for further accurate similarity identification. The formula below calculates technological similarities using both methods and is extended in Section 4.

$$S(A,B) = S\big(CS(A,B), SS(A,B)\big)$$

Note that, although both $CS(A,B)$ and $SS(A,B)$ have been normalized separately, and the values of both are within the interval [0, 1], we aim to ensure that the integration stands on the same stage, so further normalization in this model may be required.

4. Empirical Study

Approaches to similarity measures hold a strong capability to identify technological similarities that assist in a wide range of patent portfolio analyses, e.g., patent maps (Kay et al. 2014; Leydesdorff, Kushnir & Rafols 2014), technology mergers and acquisitions (Makri, Hitt & Lane 2010; Park, Yoon & Kim 2013; Yoon & Song 2014), and general topic analysis for technical intelligence (Fabry et al. 2006; Zhou et al. 2014). This paper applies our hybrid similarity measure method to analyze the technological similarities between selected firms in China's medical device industry, where each firm is represented by a patent portfolio. This empirical study serves to demonstrate the feasibility, reliability, and performance of our method.

4.1. Data

Based on a list of firms in China's medical device industry from the Wind Financial Terminal financial database[1], we divided a total of 709 firms into 3 groups as shown in Table 4. Group A contains 15 firms (1 billion RMB to +∞). Group B contains 56 firms (1 million RMB to 1 billion RMB). Group C contains 638 firms (0 to 1 million RMB). The purpose of this scenario is to consider technological similarity in technology mergers and acquisitions (Makri, Hitt & Lane 2010), where the level of an acquirer's capital cannot be ignored. Note that although the empirical study concentrates on the methodology, the standard concerns that arise during technology mergers and acquisitions are a way to further illustrate the reliability and practical implementation of the method. Since it is unrealistic to expect all firms to have hundreds or thousands of patents, we selected firms based on the following criteria: 1) we limited the total number of firms within the interval [50, 100], since experts will be invited to mark the technological similarities between selected firms, and such scale can be appropriate; 2) we basically followed the share of the three groups, and randomly selected firms holding with more than one patent; and 3) based on patent analysis, it is easier to seek the technological similarities between large firms than small ones. Thus, we increased the share of Groups A and B to better demonstrate the benefits of our method (but the share of Group C was still significant larger than the sum of the other groups). At this stage, we randomly selected 10, 15, and 40 firms from the three groups respectively as shown in Table 4.

Table 4

Firm information for China's medical device industry and our empirical dataset

Group	*Total Assets (RMB)*	*Num. of Firms (%)*	*Num. of Selected Firms (%)*
A	more than 1 billion	15 (2.1%)	10 (15.4%)
B	between 1 million and 1 billion	56 (7.9%)	15 (23.1%)
C	less than 1 million	638 (90.0%)	40 (61.5%)

[1] http://www.wind.com.cn/Default.aspx

We collected patents from the Derwent Innovation Index (DII) database[2] in Web of Science. A search on assignee codes, assignee names, and IPCs was performed. Assignee codes are available for most firms, assignee names were used when a standard assignee code did not exist, and variations in assignee name were consolidated. We referred to the *Australian Medical Devices: A Patent Analytics Report*[3] and selected 4-digit IPCs in the following range: A61B, A61C, A61D, A61F, A61G, A61H, A61J, A61K, A61L, A61M, A61N, A62B, B01L, G01N, G03B, G06F, G06Q, and H04R. As of March 1, 2016, we retrieved 1632 patent records which constitute our corpus.

4.2. Data Pre-processing

We pre-processed the data using: 1) a term clumping process (Zhang et al. 2014a) to identify the core technological terms; and 2) expert knowledge to label the data, which was used for measuring validation.

4.2.1. Term clumping process

The term clumping process was designed to clean, consolidate, and cluster scientific and technological terms, and both association rules and expert knowledge are required (Zhang et al. 2014a). We specifically applied the term clumping process to identify the core technological terms, which constitute the data foundation of the semantic model. The stepwise results of the term clumping process are shown in Table 5.

Table 5

Stepwise results of the term clumping process

No	*Step*	*# Terms*
1	Natural language processing (NLP) via *VantagePoint* (VantagePoint 2016)	20,423
2	Basic cleaning with thesaurus	18,781
3	Fuzzy matching	16,258
4	Pruning (remove terms appearing only in one record)	3,653
5	Association rules-based consolidation (combine low-frequency terms to high-frequency terms that appear in the same records)	2,204
6	Association rules-based consolidation (combine terms with more than 2 or 3 sharing words)	1,072
7	Tf-idf weighting	1,072
8	Expert knowledge-based selection	512

Note that we applied an NLP technique to a combined content of abstract and title fields rather than the full text. We are fully aware that the accuracy of further similarity measures might be lower than that conducted by the full text, but considering new issues introduced by the full text (e.g., noisy terms and techniques of processing images or PDF files), we only focused on the combined content of abstracts and titles in this paper. In addition, we used terms derived by the NLP technique rather than individual words, which can provide more specific semantic meanings and reduce some negative influence due to the use of abstracts and titles.

The tf-idf weighting ranked terms via tf-idf values, and the classical tf-idf formula (Salton & Buckley 1988) was introduced. We denote: the entire dataset (i.e., the 1632 patent records) as a corpus D, i.e., $D = 1632$; the

[2] http://thomsonreuters.com/en/products-services/scholarly-scientific-research/scholarly-search-and-discovery/derwent-innovations-index.html

[3] More detail can be seen at the website: http://www.industry.gov.au/industry/IndustrySectors/PharmaceuticalsandHealthTechnologies/Pages/Australian-Medical-Devices-Patent-Analytics-Report.aspx

frequency of a term i as t_i; the number of the patent records that contain the term i as D_i; and the total term frequency of the patent records that contain the term i as T_i. Hence, the tf-idf value of term i can be calculated as follows:

$$TFIDF(i) = \frac{t_i}{T_i} \times log \frac{D}{D_i}$$

Based on 1072 terms with tf-idf weights, the expert panel helped us establish 512 core terms – more detail is provided in Section 4.2.2. The 512 core terms were used as basic inputs for the semantic model, and an IPC-record matrix was used for the categorical model. The statistical information, including the number of patents, the number of IPCs, the total IPC frequency, the number of terms, and the total term frequency for the three groups is given in Table 6. Detailed information about the 65 firms is provided in Appendix B.

Table 6

The statistical information of the three groups

Group	*Variable*	*Min*	*Max*	*Mean*	*Standard Deviation*
A	Number of patents	2	239	79	67.2
	Number of IPCs	1	18	10.1	5.3
	Total IPC frequency	3	269	82.3	71.0
	Number of terms	0	124	46.2	33.9
	Total term frequency	0	374	126.6	105.6
B	Number of patents	2	97	27.9	26.2
	Number of IPCs	2	15	6.6	3.5
	Total IPC frequency	12	135	44.5	34.8
	Number of terms	0	26	15.1	8.6
	Total term frequency	0	74	39.5	24.6
C	Number of patents	1	56	11.0	13.4
	Number of IPCs	1	10	3.7	3.0
	Total IPC frequency	1	98	14.0	19.9
	Number of terms	0	20	9.5	4.8
	Total term frequency	0	59	26.6	15.3

Generally, a larger firm will have more patents, but several firms in Groups A and B only had few patents while some in Group C had a large number of patents. As shown in Table 6, all variables in Group C had the smallest values of standard deviation. In addition, since we applied a relatively strict term-cleaning and identification process (i.e., the term clumping process), it is reasonable that several firms holding few patents had no core terms. This real-life situation highlights our efforts to integrate the categorical and semantic models for patent portfolio analysis – IPC-based similarity measures may be noisy, but sometimes one portfolio might have no terms and be unsuitable for term-based similarity measures.

4.2.2. Validation measures

Validation measures are the way to test the performance and reliability of our methodology. In particular, we aimed to avoid possible over-fitting issues, and hence divided the data set into a training set and a test set. The

training set consisted of 50 portfolios with 1315 patents, which were randomly selected from the entire portfolio list. The remaining 15 portfolios, with 317 patents, were grouped in the test set. We used the training set to decide the parameters and weights, and the test set was designed to reveal the adaptability of our method.

An expert panel was organized to help label data. Two technical experts from General Electric Medical Systems (China) Co., Ltd. and one Master student from the School of Management and Economics at the Beijing Institute of Technology, whose research concentrated on technology mergers and acquisitions in China's medical device industry, were involved. Based on their technical background and understanding of the selected 65 firms, the three experts scored the technological similarities between the 2080 firm-pairs with values. The criterion for labelling data was: if the experts thought the technological foci of two patent portfolios were similar, they would score the similarity value of the firm-pair as *1*; if they were not sure, *0.5* would be given; and if irrelevant, they would score it as *0*. The inter-rater agreement in scores between the three experts is shown in Table 7.

Table 7

The inter-rater agreement in scores between the three experts

	Expert 1	*Expert 2*	*Expert 3*
Expert 1	1	0.89	0.79
Expert 2		1	0.76
Expert 3			1

As shown in Table 7, Expert 1 and Expert 2 shared close correlation, In fact, Experts 1 and 2 are the technical experts we invited from General Electrical Medical Systems (China) Co., Ltd. Although the scores given by Expert 3, the Masters student, were not as similar to the other two, these correlation values are still acceptable as a validation measure. Furthermore, considering that some experts issued a score of '1' conservatively, while others were liberal, we weighed the experts' scores in the following manner. Given the total number of 1 that the i-th expert scored is λ_i ($i \leq 3$), the weight of the expert can be calculated as:

$$\varepsilon_i = \frac{1/\lambda_i}{\sum_{k=1}^{3} \frac{1}{\lambda_k}}$$

We collected the expert scores for the 2080 pairs of firms, but found it extremely difficult to use the similarity values calculated by our method to directly match the scores given by the experts, even across a small range. Therefore, we used the validation rankings, and the 2080 firm-pairs were extended to 45,760 rankings. As an example, the firm-pair of Portfolio $P1$ and $P2$ will become the rankings (i.e., higher, lower, or equal) derived by the comparisons between the similarity of Portfolio $P1$ and $P2$ [the similarity can be either $S(P1, P2)$, or $CS(P1, P2)$, or $SS(P1, P2)$] and the similarities of Portfolio $P1$ and other 63 portfolios.

We introduced an *F Measure* as an indicator of performance, which can be calculated below. The *Precision* value measures how many rankings calculated by our method are the same as those given by the experts. For example, if the results in our similarity of Portfolio $P1$ and $P2$ were larger than that of $P1$ and $P3$, and the experts' scores were the same as our method, we can say that ranking is correct. The *Recall* value measures how many rankings of "higher" can be correctly calculated by our method.

$$F\ Measure = \frac{2 \times Precision \times Recall}{Precision + Recall}$$

4.3. Experiments with the training set

4.3.1. The categorical similarity measures

The entire corpus contains 41 4-digit IPCs and 119 7-digit IPCs, at this stage, we set the universe as $X = \{x_1, x_2, \dots, x_i, \dots, x_{64}, x_{65}\}$, where x_i is the i-th patent portfolio, and denoted each 7-digit IPC j as a fuzzy set A_j, where $j \in [1,119]$. A basic membership function $A_j(x_i)$ is set below, where $PN(x_i)$ represents the number of the patents held by portfolio x_i, and $PN(j|x_i)$ is the number of the patents that were held by portfolio x_i and belonged to IPC j.

$$A_j(x_i) = \frac{PN(j|x_i)}{PN(x_i)}$$

It is clear that $A_j(x_i)$ calculates membership grades in a linear way and, we tested several non-linear functions, based on the training set, to reconstruct $A_j(x_i)$ to seek the most suitable function for the best performance.

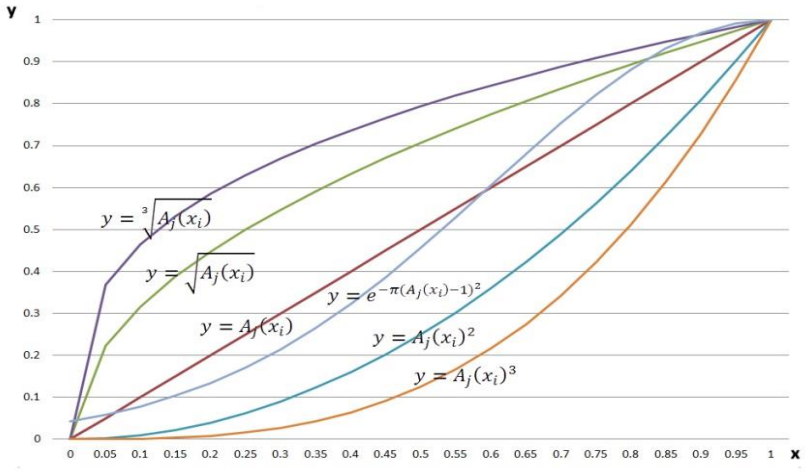

Fig. 4. Fuzzy set A_j with different membership functions

This model uses the cosine measure (given in Section 3.2) for vector-based similarity measures; however, we allocated a 0.5 weight when a 7-digit IPC did not match but its superordinate 4-digit IPC was the same. For example, portfolio A contains A61B001, while portfolio B does not but it has A61B005, which shares the same superordinate 4-digit IPC A61B with A61B001. At this stage, the cross product[4] of A61B001 $cp(A,B)$ is calculated as below:

$$cp(A,B) = \vartheta_A(\text{A61B001}) \times (0.5 \times \vartheta_B^{Min})$$

where $\vartheta_A(A61B001)$ is the membership grade of A61B001 in vector $V(A)$, and ϑ_B^{Min} is the minimum membership grade in vector $V(B)$.

We also compared with traditional cosine and Jaccard approaches with raw IPC frequency, where the cosine approach follows the definition given in Section 3.2 and the Jaccard approach is further described as follows:

$$Jaccard(x_i, x_k) = \frac{V(x_i) \cdot V(x_k)}{|V(x_i)| + |V(x_k)| - V(x_i) \cdot V(x_k)}$$

The results of the experiments are given in Table 8.

Table 8

Performance of the categorical model vs. traditional cosine & Jaccard approaches

[4] Note that the norm of the related vector will also be changed.

No	*Membership Function*	*Precision*	*Recall*	*F Measure*
1	A traditional Jaccard approach with raw IPC frequency	0.659	0.690	0.674
2	A traditional cosine approach with raw IPC frequency	0.659	0.691	0.675
3	$y = A_j(x_i)$	0.674	0.711	0.692
4	$y = \sqrt{A_j(x_i)}$	0.684	0.728	0.705
5	$y = \sqrt[3]{A_j(x_i)}$	0.686	0.731	0.708
6	$y = A_j(x_i)^2$	0.666	0.699	0.682
7	$y = A_j(x_i)^3$	0.664	0.698	0.681
8	$y = e^{-\pi(A_j(x_i)-1)^2}$	0.679	0.718	0.698

Several interesting observations are summarised as follows: 1) the comparison between Approaches 1 and 2 endorse the argument given by Leydesdorff (2008), in which the cosine measure demonstrates tiny advantages over the Jaccard approach; 2) the engagement of fuzzy sets provided a benefit to the categorical model, and non-linear deformations further improved the performance; and 3) the reconstruction with the cube root function (Approach 5) obtained the highest *F Measure* (with the highest values in both *Precision* and *Recall*), which illustrated the best performance in the experiments with the training set.

4.3.2. The semantic similarity measures

Based on the 512 core terms derived by the term clumping process and their co-occurrence relationships, we generated a term co-occurrence map (shown in Fig. 5) via VOSviewer (van Eck & Waltman 2009; Waltman, van Eck & Noyons 2010). According to Fig. 5, we could easily address several “hot” topics, and this figure provided an overview of the landscape of China’s medical device industry, which could be a reference to help understand the diverse technical foci of the corpus and better explore their similarities.

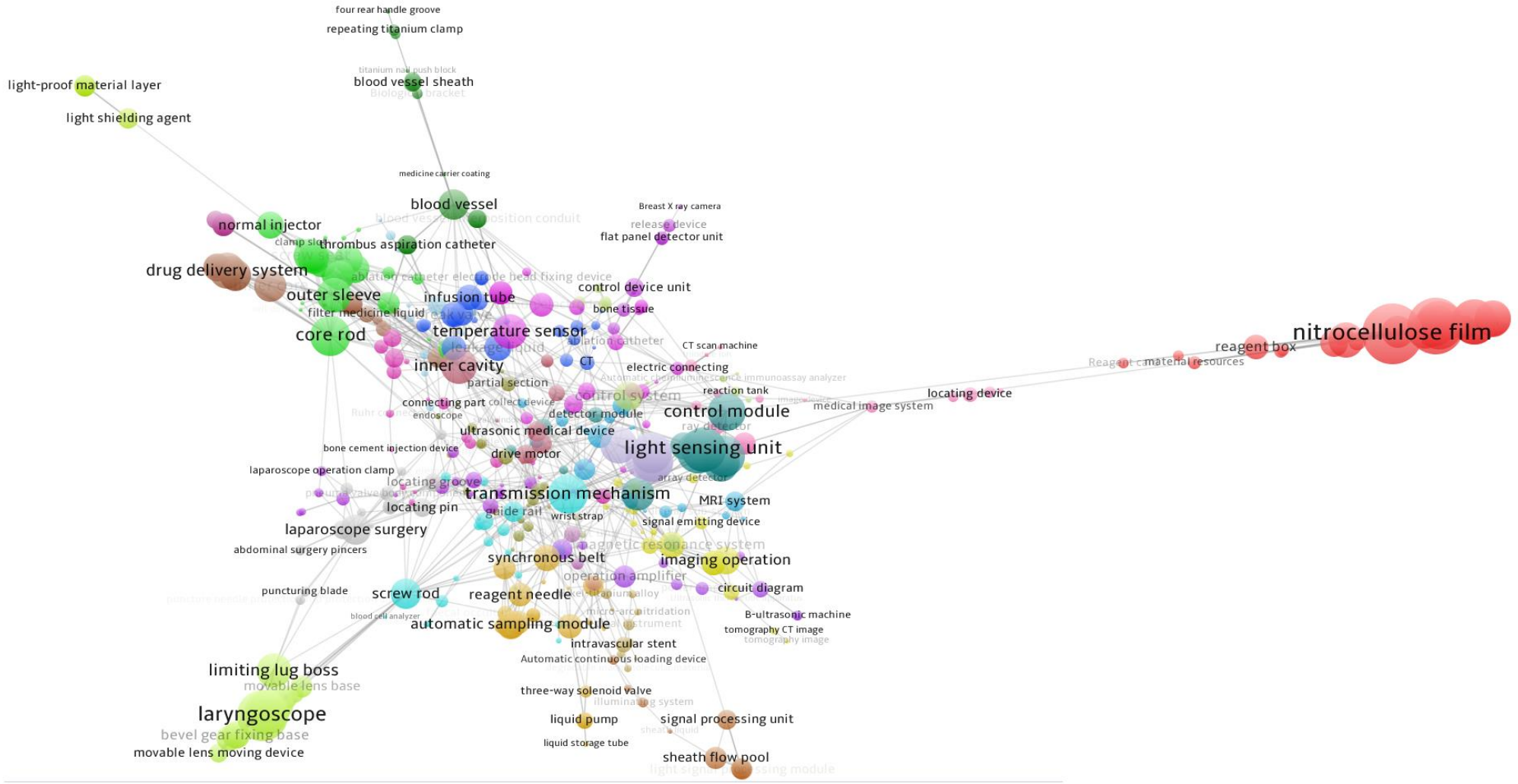

Fig. 5. Term co-occurrence map of China's medical device industry (generated by VOSviewer).

An improved K-means algorithm (Zhang et al. 2016a) was introduced to cluster terms in each patent portfolio. We used raw term frequency directly and returned a local optimum K value of 5 in the interval [3, 10]. In this context, we created a 3-level tree structure for each patent portfolio. We set the cluster label as a Level 1 leaf and the remaining terms in the cluster were set as Level 2 leaves. If the total number of terms in a portfolio was fewer than 5, we treated them all as Level 2 leaves. A sample of the tree structure is shown in Table 9.

Table 9

Sample tree structure of a patent portfolio

Patent Portfolio (Root)	*Level 1 Leaf*	*Level 2 Leaf*
Firm X	magnetic resonance	x ray photography system
		control unit
	medical image device	CT
		magnetic resonance image
		magnetic resonance system
		medical image system
		MRI system

Before we calculated the semantic similarity, we introduced AHP to set the weights of the six matching types (given in Table 2). As discussed, we chose $w_1 > w_2 > w_6 > w_4 = w_5 > w_3$ as the basis of our case: an entire match is best ($w_1 > w_2 \sim w_6$); matches with Level 1 leaves are always better than matches with Level 2 leaves ($w_2 > w_6$ and w_4 & $w_5 > w_3$); and there is no direction for the similarity measure ($w_4 = w_5$). At this stage, following the basic steps and the AHP fundamental scale proposed by Saaty (1990), we compared the pairwise values and constructed the matrix in Table 10. Since the consistency ratio (CR) was less than 0.1, the estimate of the pairwise comparison matrix was acceptable. Therefore, we set the priority vector as the vector of the weights, where $W = \{w_1, \dots, w_6\} = \{0.43, 0.21, 0.04, 0.08, 0.08, 0.16\}$.

Table 10

The pairwise comparison matrix

	R1	*R2*	*R3*	*R4*	*R5*	*R6*	*Priority Vector*
R1	1	3	6	5	5	4	0.43
R2	1/3	1	5	3	3	2	0.21
R3	1/6	1/5	1	1/3	1/3	1/4	0.04
R4	1/5	1/3	3	1	1	1/3	0.08
R5	1/5	1/3	3	1	1	1/3	0.08
R6	1/4	2	4	3	3	1	0.16
							λ_{max}=6.25, CR=0.04

The semantic similarity values $SS(P1, P2)$ were then calculated. The performance of our semantic similarity measure approach compared with the raw frequency-based cosine and Jaccard approaches is shown in Table 11.

Table 11

Performance of the semantic model vs. traditional cosine/Jaccard approaches

No	*Method*	*Precision*	*Recall*	*F Measure*
1	Raw frequency & cosine	0.729	0.819	0.771
2	Raw frequency & Jaccard	0.726	0.815	0.768
3	Tree	0.751	0.841	0.793

As shown in Table 11, the tree-based sematic model improved the performance in both *Precision* and *Recall*, and demonstrates the efficiency of this experiment. Delving into the reasons why, we conclude: 1) the tree structure provided a way to effectively retrieve and identify the most significant concepts of patent portfolios, which could be core technologies or the technological components of related companies. The comparison between these structures further highlights such concepts and makes the results accurate; 2) the basic elements of the trees in this case were the core terms identified by our expert panel, and it is clear that the influence of these deep-cleaned terms cannot be ignored; and 3) despite not being the main focus of this paper, the comparison between the cosine and Jaccard approaches is a hotspot in bibliometrics. In our experiments, the results of both the categorical and semantic models support the arguments stated by previous studies, although we found the advantage of cosine approaches to be extremely weak.

4.3.3. Similarity score calculation

Before calculating similarity scores, we normalized the results derived from the categorical and semantic similarity measures to make sure they were within the same scale for further integration. A min-max normalization approach was applied, and the formula is described as follows:

$$x' = \frac{x - min(x)}{max(x) - min(x)}$$

where x is an original value, x' is the normalized value, and $min(x)$ and $max(x)$ is the minimum and maximum value of the feature x respectively.

Actually, this case showed no clear bias between IPCs and text elements, but did leave concern that the scope was concentrated on a relatively limited technology (i.e., medical device), and text elements might be preferable. However, aiming to seek the best way to combine the results of the two models, we first chose the best results of the two models, i.e., Approach 5 in Table 8 and Approach 3 in Table 11, and applied a traditional weighting approach to functionally integrate the results. The similarity score $S(A,B)$ can be calculated as follows:

$$S(A,B) = S\big(CS(A,B), SS(A,B)\big) = w_{cs} \times CS(A,B) + w_{ss} \times SS(A,B)$$

where w_{cs} and w_{ss} is the weight of $CS(A,B)$ and $SS(A,B)$ respectively, and $w_{cs} + w_{ss} = 1$.

Our considerations follow. 1) Since the performance of the semantic model was better than the categorical model, we set several observation points to detect change in performance. 2) Despite a continuity interval of between 0 and 1, the most interesting thing to test was whether the semantic model and the categorical model could work in a complementary fashion to achieve better performance than applying the results individually. The performance of the traditional weighting approach-based similarity calculation is shown in Fig. 6.

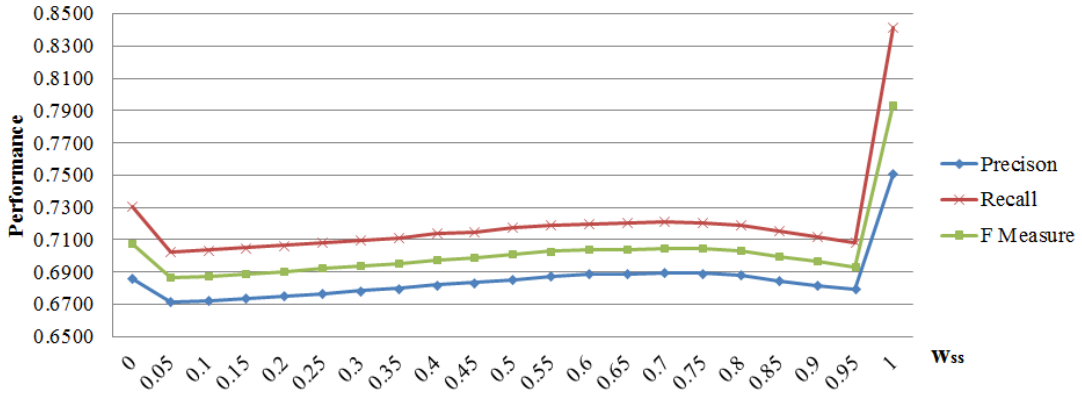

Fig. 6. Performance of the traditional weighting approach-based similarity calculation.

Unfortunately, the categorical similarity measures and the semantic similarity measures did not complement each other if we simply combined their results with certain weights. Performance of the combinations was even worse than when applied individually. One argument to explain the reason is: we validated the results by the rankings, but such combinations broke basic rules in the two models despite normalization. Therefore, we designed several assembled sets to integrate the two types of similarity values in a milder way, and their performance is given in Table 12.

Table 12

Performance of the assembled sets for the categorical and semantic similarity measures

No	*Method*	*Precision*	*Recall*	*F Measure*
1	$S_1(A,B) = \begin{cases} CS(A,B) & if\ CS(A,B) > 0 \\ SS(A,B) & else \end{cases}$	0.65	0.66	0.66
2	$S_2(A,B) = \begin{cases} SS(A,B) & if\ SS(A,B) > 0 \\ CS(A,B) & else \end{cases}$	0.67	0.69	0.68
3	$S_3(A,B) = \begin{cases} SS(A,B) & if\ SS(A,B) > CS(A,B) \\ CS(A,B) & else \end{cases}$	0.68	0.71	0.69
4	$S_4(A,B) = \begin{cases} SS(A,B) & if\ 0 < SS(A,B) < CS(A,B) \\ CS(A,B) & else \end{cases}$	0.79	0.92	0.85

The four assembled sets hold diverse priorities, e.g., $S_1(A,B)$ is prior to the results of the categorical similarity measures while $S_2(A,B)$ emphasizes the semantic model, and $S_3(A,B)$ prefers a larger similarity value but $S_4(A,B)$ uses smaller values. The first three assembled sets did not achieve our expectation, but the performance of $S_4(A,B)$ was good and illustrates the complementarity of the categorical and semantic similarity measures. However, it is intriguing that smaller values worked better than larger values. We consulted the expert panel on this issue, and one reasonable explanation was provided: it is difficult for experts to use a numeric value to evaluate the technological similarity between portfolios, and a large number of portfolios were marked as having limited or no similarity to the others. Thus, in some senses, the similarity derived from terms or IPCs might be deeper and broader than the marks given by experts. In another words, using smaller values to increase the threshold of exploring similarity might be able to cater to expert scores. However, one remaining concern here is, whether similarities derived by either of the models that are different from expert knowledge can be meaningful or not. We specifically focus on this question in Section 4.5.

4.4. *Experiments with the test set*

Based on the results of the experiments with the training set, the parameters for the case were decided as follows: 1) the membership function $y = \sqrt[3]{A_j(x_i)}$ and cosine measure for the categorical model; 2) the tree structure with the weights of the six matching types $W = \{w_1, \dots, w_6\} = \{0.43, 0.21, 0.04, 0.08, 0.08, 0.16\}$ for the semantic model; and 3) the assembled set $S_4(A, B)$ to integrate the results of the two models. The performance of the test set with the parameters above is given in Table 13.

Table 13

Performance of the test set with the given parameters

No	*Method*	*Precision*	*Recall*	*F*
1	The categorical model	0.75	0.81	0.78
2	The semantic model	0.84	0.91	0.87
3	The hybrid similarity measure method	0.84	0.95	0.89

Shown in Table 13, the performance of the semantic model was better than that of the categorical model, and the hybrid similarity measure method further improved performance. Similar to the results of the experiments with the training set, Table 13 demonstrates the adaptability of our methodology. Considering the risk of over-fitting, we used a training set to choose the parameters, and the test set acted as an independent dataset for evaluating the performance of our method. Clearly, the results of the test set were consistent with those from the training set, which supports the point that our methodology is not only suitable for this case study but can be adaptable to a wide range of related data sources and cases. However, such adaptability does not clarify that our method can take the place of any other approaches for measuring similarities. Instead, we emphasize that our method had the best performance in this case and holds the capability to work well in other situations.

4.5. Case study on Shanghai United Imaging Healthcare Co.

Even though the performance of our methodology has been demonstrated by experiments with training and testing sets, it is still interesting to delve into the case of China's medical device industry. The foci of this case study in Section 4.5 are: 1) to visualize the technological relationships between the selected portfolios in China's medical device industry based on the results of the hybrid similarity measure method; and 2) to address concerns on certain discrepancies between the calculated method and the expert knowledge, and explore possible industrial implications from the perspective of technology mergers and acquisitions.

We applied the parameters decided by the training set to the entire dataset and obtained 389 links between the 65 portfolios. We used VOSviewer (van Eck & Waltman 2009; Waltman, van Eck & Noyons 2010) to generate a portfolio correlation map of China's medical device industry, shown in Fig. 7 (comparably, the portfolio correlation maps respectively generated by the results of the categorical model and the semantic model are given in Appendix C). Note that, based on the hybrid similarity measure method, there are nine portfolios that do not share links to any other portfolios, thus, Fig. 7 only contains 53 nodes, which represents the 53 remaining portfolios. In addition, the size of node is used to describe the total link strength of a node.

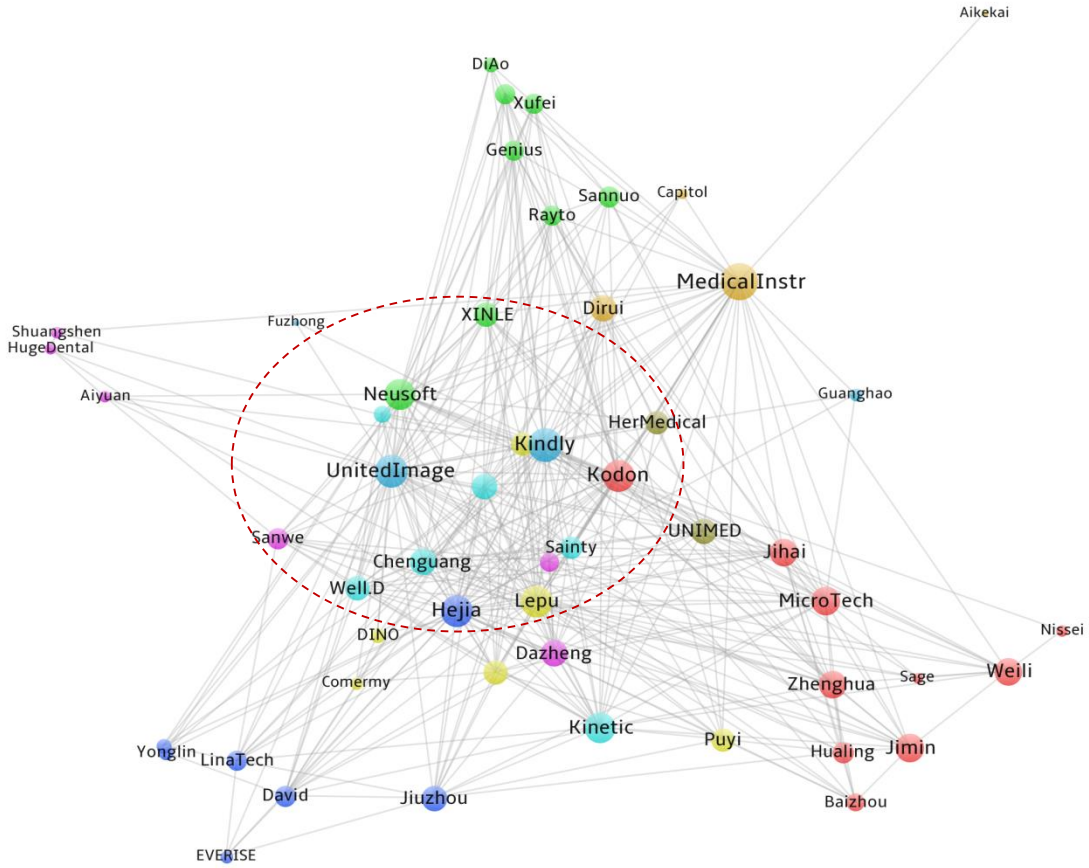

Fig. 7. Portfolio correlation map of China's medical device industry (based on hybrid similarity measure method with IPCs and terms derived from DII database).

The portfolios within the red circle all concentrate on medical imaging devices, and UnitedImage is the leading one among them. Known as United Imaging Healthcare (UIH) Co., UIH is a high-tech company in Shanghai, South China, and dedicates to the development, manufacturing and sales of innovative medical imaging equipment and technologies. Its major business units include: components, computed tomography, molecular imaging, magnetic resonance, radiotherapy, X-ray, mobile health, and software[5]. Although some of the other portfolios in the circle are also listed companies, medical imaging is only one of their major products, e.g., Neusoft is involved with medical imaging equipment and information systems[6]; Lepu mainly focuses on cardiac therapy, but angiographic equipment is one of its major products[7]; the major business direction of Hejia is minimally invasive treatment, and digital subtraction angiography is one important auxiliary product[8]. Similarly, small and medium size companies such as Well.D, Sanwe, Sainty, and Chengguang are also within this scope. These results were calculated by our hybrid similarity measure method and visualized in Fig. 7, and they exactly match the scores given by the expert panel, and the reliability of our method in identifying technological similarities from patent documents is effectively demonstrated.

It is definitely more interesting to discuss the unexpected results – those that were different from the expert scores. Concentrating on UIH, the remaining portfolios in the circle, i.e., Xinle, Kindly, Kodon, and HerMedical, are such examples. Our expert panel thought the four firms should have no technological similarities to UIH, since their main business sectors are not within the scope of medical imaging, but the method's analysis of the patent corpus and the websites of those four companies revealed intriguing results. 1) Shanghai Kindly Co. is a

[5] More details can be seen on the website: http://www.united-imaging.com/
[6] More details can be seen on the website: http://medical.neusoft.com/en/
[7] More details can be seen on the website: http://en.lepumedical.com/
[8] More details can be seen on the website: http://en.hokai.com/

listed company in Group A and is famous for medical appliances. It has no apparent direct relationship with UIH. However, "angiography catheter" is one of the items included in the category of intervention accessories[9] and linked both companies. Considering they have similar geographic locations, enterprise size, and diverse divisions for angiography, UIH and Shanghai Kindly Co. could seek opportunities to cooperate in the sector of angiography. This potential relationship belongs to certain of technological similarity, and represents a promising recommendation that may result from this method. 2) HerMedical is a small company in Wuhan, Central China and belongs to Group C in our set. This company started with patents in optical imaging, but has been dedicated specifically to imaging for gynaecological oncology[10]. By contrast, mammography supported by previous patents is new business for UIH, and new methods for detecting lung cancer in early stages have already been commercialized[11]. Therefore, it is reasonable for UIH to consider HerMedical as the target of technology mergers or acquisitions if UIH plans to advance to the field of gynaecological oncology and holds interest to the market of Central China. 3) Hebei Xinle Sci & Tech Co., in Group B, specifically focuses on the manufacturing of blood collection tube systems, which also appears to have no relationship to UIH. However, on the Chinese version of its website, a new direction for one of its major products is an intelligent blood collection management system, which could form a part of hospital or laboratory information systems[12]. This case definitely matches with UIH's focus on mobile health and software. Similarly, the main products of Kodon (also known as Tianjin Andon) are monitors for blood pressure and diabetes, but its current strategy is to promote i-health products, including a series of smart bracelets and applications in Apple Stores[13]. Xinle and Kodon have geographic advantage in collaboration, and considering the capital of Kodon and existing successful products of both companies in mobile health, we could imagine possible competition between UIH and Kodon if UIH attempts to enter the mobile health market in North China – Both Xinle and Kodon are located in the Bohai rim, the most famous economic region surrounding Beijing.

As discussed above, from the perspective of technology mergers and acquisitions, the results given by our hybrid method demonstrated capability to explore underlying technological similarities from patent documents in an objective way. The valuable intelligence gained from this case study also proves the benefits of our method in practical applications.

5. Discussion and Conclusions

This paper proposes a hybrid method for measuring similarity in patent portfolios. Similarity measures are separated into categorical similarities and semantic similarities to account for both IPCs and text elements. Fuzzy sets are introduced to transform vague IPC definitions into numeric values, and a semantic tree is constructed to calculate the similarities between hierarchical structures. Considering the diverse definitions of patent portfolios (e.g., patent assignee, country and region, and technical topic), the patent corpus of selected firms in China's medical device industry was used to demonstrate the reliability, adaptability, and performance of our method. We measured the technological similarities between different firms, and validated the results using an expert ranking matrix. A case study, from the perspective of technology mergers and acquisitions, was conducted to examine some unexpected results, which differed to the scores given by the expert panel. The

[9] The product information of Shanghai Kindly can be seen on the website: http://www.kdlchina.com/kdlnews.aspx?id=55, at the same time, since UIH does not directly mention angiography as its major direction of products, we locate UIH's patents focusing on this sector, and the results include CN103054580 B, CN105640583 A, etc.

[10] Unfortunately, this small company only has a website of Chinese version: http://www.heer.com.cn/index.asp

[11] News of UIH's efforts on mammography and early-stage cancer detection can be addressed at the websites: http://www.united-imaging.com/vm/company-news-article/items/uih-honored-with-three-red-dot-product-design-awards.html and http://www.united-imaging.com/vm/news-archive.html?year=2015, and such new business are supported by UIH's previous patents, such as CN103845816 A, CN104766340 A, CN 105748161 A, and CN 104182965 A.

[12] More detail can be seen on the websites: http://en.hbxinle.com/index.asp and http://www.hbxinle.com/product/ (product information in Chinese)

[13] More detail can be seen on the websites: http://www.andonhealth.com/index.htm and http://www.jiuan.com/index.php?case=archive&act=list&catid=25 (i-health products)

discussion describes the benefits in identifying underlying technological similarities between patent portfolios gained from our method.

5.1. The strengths and weaknesses of categorical and semantic similarity measure models

Similarity measures are basic tools for a wide range of bibliometric studies and, as such, detailed comparisons and discussion on the strengths and weaknesses of different approaches to similarity measures have been conducted (Leydesdorff 2008; Boyack & Klavans 2010; Moehrle 2010; Boyack et al. 2011). Yet, in-depth insights for patent-oriented similarity measures are still elusive. At this stage, a comparison between IPC- and word or term-based similarity measures can be meaningful to further patentometric studies. In particular, both categorical and semantic similarity measure models act as representations and, through experiments we compared these two models in a variety of criteria as follows.

Data Sources – both models can adapt to suit popular patent data sources, such as the United States Patent and Trademark Office (USPTO) database, the European Patent Office (EPO) database, and DII patents. However, considering similar situations in textual content, the semantic model can be used for a broader range of ST&I data sources, such as scientific publications and academic proposals.

Data Scope – data scope is one of the most important issues to influence the efficiency of both models. Generally, both show better performance on data with a relatively wide scope and low-coupled sub-domains, like multidisciplinary studies, than on data with a narrow scope with high-coupled sub-domains, like a particular technological area.

Data Size – We did not conduct direct experiments to test for scalability, but we feel that both models are suitable for large-scale datasets, since we fully considered time and space complexity when designing our algorithms. In addition, our experiments involved three groups of portfolios with a diverse number of patents, and diversity would not change significantly with an increase in data size. Therefore, it is reasonable to imagine that performance levels in large-scale datasets would be acceptable. The only concern regarding data size is that the semantic model does not work smoothly on portfolios with only a few patents. The semantic model requires a minimum threshold of terms to be effective, but the categorical model can always retrieve at least one IPC from one patent that is workable.

Expert Knowledge – Engaging sufficient experts to supervise data is not always possible, but the degree of expert engagement is a point that helps us compare the categorical and semantic similarity measure models. As shown in our experiments, the only expert knowledge required in the categorical model is for the selection of membership functions. An alternative solution is to use a training set to traverse a number of optional membership functions before deciding which one is best. By contrast, the semantic model needs relatively heavy expert engagement. On one hand, we need expert help to identify core terms for the term clumping process which significantly influence the accuracy of further analyses. On the other hand, although the weights of the six matching types can be settled and are reusable in most cases, it still makes sense to invite experts to confirm or refine these weights according to the actual requirements of research questions. In short, the categorical similarity measure model is more suitable for situations that lack expert support, while the semantic similarity measure model provides higher accuracy if sufficient expert knowledge is available.

Accuracy – IPC is a technological classification system designed and maintained by the World Intellectual Property Organization (WIPO). Patent applicants need to match their inventions with existing classification codes and in some cases such subjective judgment might add noise. Similarly, an assembled set of patents with a number of core terms could represent an invention in more than one field, and expert engagement would further improve accuracy in these situations. Given these considerations, the categorical model can perform well in relatively unsupervised environments, but will reach a bottleneck given enough noise in the IPC system. By contrast, the semantic model cannot be easily adapted to highly unsupervised environments, and the negative effects that result from meaningless terms and synonyms could serve as a fatal blow for this method. However,

the accuracy of the semantic model is rapidly improved with the engagement of experts – even limited expert knowledge – and at this stage its accuracy has the potential to be better than the categorical model, which has been demonstrated by the experiments in this paper.

Based on the discussion of the strengths and weaknesses of both models, it may be more promising to consider a strategy that integrates both. One feasible option is to use categorical similarity measure to pre-process the raw data, which would be effective for filtering noise or features, and then apply semantic similarity measures to deep-clean the data for accurate similarity measures and insight discovery. Given real-world research questions, cases, and data, another option is to decide the way to consolidate the results of the two models in $S(A, B)$, as we did in the empirical study.

5.2. Implementation and possible applications

Similarity measures are basic techniques that have been widely applied to patent analyses, and especially to patentometric studies with cluster- or classification-based analyses. Concentrating on patent portfolio analysis, there are multiple entities with which to define the portfolio (e.g., individual, organization, region, and country), and no matter which entity the portfolio reflects, it is necessary to identify relationships between the portfolios via similarity measures. In this context, we could visualize the applications of our method based on multiple needs, and we attempt to summarize as follows:

Multidisciplinary Studies – patent maps based on similarity measures are one important scholarly direction for multidisciplinary studies. Similarity measures are also fundamental parts of identifying relationships between different disciplines, which are packaged as portfolios. Related works currently emphasize IPCs (or UPCs for the patents from the USPTO) for similarity measures (Leydesdorff & Bornmann 2012; Kay et al. 2014; Leydesdorff, Kushnir & Rafols 2014), and our method supports those tasks.

Competitive Technical Intelligence (CTI) Studies – CTI emphasizes tasks in technology-related competition and collaboration analyses for a specific company, industry, or country (Porter & Newman 2011). It is easy to link patent portfolios with those entities and, compared to traditional co-occurrence analysis with bibliography couplings and IPCs, our method is able to delve into textual content to discover the underlying relationships in detail. The empirical study, in particular the case study of technology mergers and acquisitions, is within this scope.

5.3. Limitations and future research

Measuring the technological similarity between patent portfolios is not an easy task for purely quantitative computation, and there are many internal and external factors that influence the performance of our method. It is therefore meaningful to discuss such sensitive items in detail and provide a reference for applications and further studies.

The Expert Scores for Validation – We believe that technical experts know much more than machines and the scores given by experts are credible. However, we must also agree that such subjective judgement, derived from the diverse research experience of experts, is a key issue that heavily influences the validation of our methods. Despite our use of rankings to validate the results, bias still exists. In addition, technological similarity itself is a fuzzy concept, and such uncertainty also increases the difficulty of processing expert scores. We attempted to reduce the fuzziness by asking experts to categorize similarity into “similar,” “not sure,” and “irrelevant” via 1, 0.5, and 0 respectively, but it is well-known in fuzzy set theory that the term “similar” is rough. Therefore, the expert scores in our experiments are open to criticism, but do act as a fair platform to compare our methods with other traditional approaches.

The Weights and Parameters – Establishing weights and parameters is always a fundamental task. In our methods these efforts include: the selection of the membership functions; weighting the six matching types; and the method for integrating the results from the categorical and semantic models. We attempted to use the most

promising manner to decide these weights and parameters either quantitatively or qualitatively, but machine learning and optimization techniques might be required in future studies.

The Empirical Study – We chose real-world data in China's medical device industry for the empirical study, but the number of selected patents may not be sufficient for testing our method at every point. Although this case study demonstrated that our method provides insights into technology mergers and acquisitions, using real-world data also introduced unexpected challenges, such as appropriate validation measures. It would therefore be interesting to apply our methods to a broad range of cases to test their reliability and robustness.

Acknowledgements

We acknowledge support from the National High Technology Research and Development Program of China (Grant No. 2014AA015105), the Australian Research Council (ARC) under Discovery Project DP140101366, the National Science Foundation of China Yong Funds (Grant No. 71103015), and the Basic Research Foundation of Beijing Institute of Technology (Grant No. 20152142010). We acknowledge promising and helpful comments from the editor-in-chief Dr Ludo Waltman and four anonymous reviewers. We are also grateful to two technical experts from General Electric Medical Systems (China) Co., Ltd. and to Kangrui Wang from Beijing Institute of Technology for serving in our expert panel.

References

Ahlgren, P., & Colliander, C. (2009). Document–document similarity approaches and science mapping: Experimental comparison of five approaches. *Journal of Informetrics,* 3(1), 49-63.

Boyack, K. W., Klavans, R., & Börner, K. (2005). Mapping the backbone of science. *Scientometrics,* 64(3), 351-374.

Boyack, K. W., & Klavans, R. (2008). Measuring science–technology interaction using rare inventor–author names. *Journal of Informetrics,* 2(3), 173-182.

Boyack, K. W., & Klavans, R. (2010). Co‐citation analysis, bibliographic coupling, and direct citation: Which citation approach represents the research front most accurately? *Journal of the American Society for Information Science and Technology,* 61(12), 2389-2404.

Boyack, K. W., Newman, D., Duhon, R. J., Klavans, R., Patek, M., Biberstine, J. R., . . .& Börner, K. (2011). Clustering more than two million biomedical publications: Comparing the accuracies of nine text-based similarity approaches. *PLoS One,* 6(3), e18029.

Braam, R. R., Moed, H. F., & Van Raan, A. F. (1988). Mapping of science: Critical elaboration and new approaches, a case study in agricultural biochemistry. *Journal of Informetrics,* 87/88, 15-28.

Braam, R. R., Moed, H. F., & Van Raan, A. F. (1991). Mapping of science by combined co-citation and word qnalysis II. Dynamical aspects. *Journal of the American Society for Information Science,* 42(4), 252.

Callon, M., Courtial, J.-P., Turner, W. A., & Bauin, S. (1983). From translations to problematic networks: An introduction to co-word analysis. *Social Science Information,* 2(22), 191-235.

Chen, C., Ibekwe SanJuan, F., & Hou, J. (2010). The structure and dynamics of cocitation clusters: A multiple‐perspective cocitation analysis. *Journal of the American Society for Information Science and Technology,* 61(7), 1386-1409.

Chen, C., Hu, Z., Liu, S., & Tseng, H. (2012). Emerging trends in regenerative medicine: A scientometric analysis in CiteSpace. *Expert Opinion on Biological Therapy,* 12(5), 593-608.

Chen, D.-Z., Huang, M.-H., Hsieh, H.-C., & Lin, C.-P. (2011). Identifying missing relevant patent citation links by using bibliographic coupling in LED illuminating technology. *Journal of Informetrics,* 5(3), 400-412.

Chen, Y.-L., & Chang, Y.-C. (2012). A three-phase method for patent classification. *Information Processing & Management,* 48(6), 1017-1030.

Choi, C., & Park, Y. (2009). Monitoring the organic structure of technology based on the patent development paths. *Technological Forecasting and Social Change,* 76(6), 754-768.

Cong, H., & Tong, L. H. (2008). Grouping of TRIZ Inventive Principles to facilitate automatic patent classification. *Expert Systems with Applications, 34*(1), 788-795.

Fabry, B., Ernst, H., Langholz, J., & Köster, M. (2006). Patent portfolio analysis as a useful tool for identifying R&D and business opportunities—an empirical application in the nutrition and health industry. *World Patent Information,* 28(3), 215-225.

Fall, C. J., Törcsvári, A., Benzineb, K., & Karetka, G. (2003). Automated categorization in the international patent classification. *Proceedings of ACM SIGIR Forum,* 37(1), 10-25.

Fall, C. J., Törcsvári, A., Fiévet, P., & Karetka, G. (2004). Automated categorization of German-language patent documents. *Expert Systems with Applications,* 26(2), 269-277.

Garfield, E., Sher, I. H., & Torpie, R. J. (1964). *The use of citation data in writing the history of science*. Philadelphia, Pennsylvania, USA: Institute for Scientific Information Inc.

Garfield, E., Paris, S., & Stock, W. G. (2006). HistCiteTM: A software tool for informetric analysis of citation linkage. *Information Wissenschaft und Praxis,* 57(8), 391.

Hamers, L., Hemeryck, Y., Herweyers, G., Janssen, M., Keters, H., Rousseau, R., & Vanhoutte, A. (1989). Similarity measures in scientometric research: The Jaccard index versus Salton's cosine formula. *Information Processing & Management,* 25(3), 315-318.

Harman, D. K., & Voorhees, E. M. (2006). TREC: An overview. *Annual Review of Information Science and Technology,* 40(1), 113-155.

Hu, Z., Fang, S., & Liang, T. (2013). Automatic patent classification oriented to Problems & Solutions. *Proceedings of the Conference on Artificial Intelligence and Data Mining 2013*, Sanya, China.

Huang, Y., Schuehle, J., Porter, A. L., & Youtie, J. (2015). A systematic method to create search strategies for emerging technologies based on the Web of Science: Illustrated for 'Big Data'. *Scientometrics,* 105(3), 2005-2022.

Intarakumnerd, P., & Charoenporn, P. (2015). Impact of stronger patent regimes on technology transfer: The case study of Thai automotive industry. *Research Policy,* 44(7), 1314-1326.

Jaffe, A. (1986). Technological opportunity and spillovers of R&D: Evidence from firms' patents, profits, and market value. *American Economic Review,* 76(5), 984-1001.

Kassab, R., & Lamirel, J.-C. (2008). Feature-based cluster validation for high-dimensional data. *Proceedings of the 26th IASTED International Conference on Artificial Intelligence and Applications*, 232-239.

Kay, L., Newman, N., Youtie, J., Porter, A. L., & Rafols, I. (2014). Patent overlay mapping: Visualizing technological distance. *Journal of the Association for Information Science and Technology,* 65(12), 2432-2443.

Kessler, M. M. (1963). Bibliographic coupling between scientific papers. *American documentation,* 14(1), 10-25.

Kim, J.-H., & Choi, K.-S. (2007). Patent document categorization based on semantic structural information. *Information Processing & Management,* 43(5), 1200-1215.

Klavans, R., & Boyack, K. W. (2006). Identifying a better measure of relatedness for mapping science. *Journal of the American Society for Information Science and Technology,* 57(2), 251-263.

Klavans, R., & Boyack, K. W. (2009). Toward a consensus map of science. *Journal of the American Society for Information Science and Technology,* 60(3), 455-476.

Klavans, R., & Boyack, K. W. (2016). Which type of citation analysis generates the most accurate taxonomy of scientific and technical knowledge? *Journal of the Association for Information Science and Technology*, http://arxiv.org/abs/1511.05078, under review.

Lau, A., Tsui, E., & Lee, W. (2009). An ontology-based similarity measurement for problem-based case reasoning. *Expert Systems with Applications,* 36(3), 6574-6579.

Leydesdorff, L. (2008). On the normalization and visualization of author co‐citation data: Salton's Cosine versus the Jaccard index. *Journal of the American Society for Information Science and Technology,* 59(1), 77-85.

Leydesdorff, L., & Bornmann, L. (2012). Mapping (USPTO) patent data using overlays to Google Maps. *Journal of the American Society for Information Science and Technology,* 63(7), 1442-1458.

Leydesdorff, L., Kushnir, D., & Rafols, I. (2014). Interactive overlay maps for US patent (USPTO) data based on International Patent Classification (IPC). *Scientometrics,* 98(3), 1583-1599.

Lucio-Arias, D., & Leydesdorff, L. (2008). Main‐path analysis and path‐dependent transitions in HistCite™‐based historiograms. *Journal of the American Society for Information Science and Technology,* 59(12), 1948-1962.

Ma, J., Zhang, G., & Lu, J. (2012). A method for multiple periodic factor prediction problems using complex fuzzy sets. *IEEE Transactions on Fuzzy Systems,* 20(1), 32-45.

Makri, M., Hitt, M. A., & Lane, P. J. (2010). Complementary technologies, knowledge relatedness, and invention outcomes in high technology mergers and acquisitions. *Strategic Management Journal,* 31(6), 602-628.

Moehrle, M. G. (2010). Measures for textual patent similarities: A guided way to select appropriate approaches. *Scientometrics,* 85(1), 95-109.

Nakamura, H., Suzuki, S., Sakata, I., & Kajikawa, Y. (2015). Knowledge combination modeling: The measurement of knowledge similarity between different technological domains. *Technological Forecasting and Social Change,* 94, 187-201.

Noyons, E., & van Raan, A. (1998). Advanced mapping of science and technology. *Scientometrics,* 41(1-2), 61-67.

Park, H., Yoon, J., & Kim, K. (2013). Identification and evaluation of corporations for merger and acquisition strategies using patent information and text mining. *Scientometrics,* 97(3), 883-909.

Peat, H. J., & Willett, P. (1991). The limitations of term co-occurrence data for query expansion in document retrieval systems. *Journal of the American Society for Information Science,* 42(5), 378-383.

Porter, A. L., & Detampel, M. J. (1995). Technology opportunities analysis. *Technological Forecasting and Social Change,* 49(3), 237-255.

Porter, A. L., & Newman, N. C. (2011). Mining external R&D. *Technovation,* 31(4), 171-176.

Pudovkin, A. I., & Garfield, E. (2002). Algorithmic procedure for finding semantically related journals. *Journal of the American Society for Information Science and Technology,* 53(13), 1113-1119.

Rip, A. (1988). Mapping of science: Possibilities and limitations. In A. F. J. van Raan (Ed.), *Handbook of quantitative studies of science and technology* (pp. 253-273). North-Holland: Elsevier Science Publishers B.V.

Saaty, T. L. (1990). How to make a decision: The analytic hierarchy process. *European Journal of Operational Research,* 48(1), 9-26.

Salton, G., & Buckley, C. (1988). Term-weighting approaches in automatic text retrieval. *Information Processing & Management,* 24(5), 513-523.

Sánchez, D., Batet, M., Isern, D., & Valls, A. (2012). Ontology-based semantic similarity: A new feature-based approach. *Expert Systems with Applications,* 39(9), 7718-7728.

Small, H. (1973). Co‐citation in the scientific literature: A new measure of the relationship between two documents. *Journal of the American Society for Information Science,* 24(4), 265-269.

Stock, W. G., & Stock, M. (2013). *Handbook of information science*. Berlin: Walter de Gruyter.

Su, F. P., Lai, K. K., Sharma, R., & Kuo, T. H. (2009). Patent priority network: Linking patent portfolio to strategic goals. *Journal of the American Society for Information Science and Technology,* 60(11), 2353-2361.

Tseng, Y.-H., Lin, C.-J., & Lin, Y.-I. (2007). Text mining techniques for patent analysis. *Information Processing & Management,* 43(5), 1216-1247.

van Eck, N., & Waltman, L. (2009). Software survey: VOSviewer, a computer program for bibliometric mapping. *Scientometrics,* 84(2), 523-538.

van Eck, N., Waltman, L., Noyons, E., & Buter, R. (2010). Automatic term identification for bibliometric mapping. *Scientometrics,* 82(3), 581-596.

VantagePoint. (2016). VantagePoint. Retrieved June 10, 2016, from https://www.thevantagepoint.com/

Waltman, L., van Eck, N. J., & Noyons, E. C. (2010). A unified approach to mapping and clustering of bibliometric networks. *Journal of Informetrics,* 4(4), 629-635.

Wang, X., Ren, J., Zhang, Y., Zhu, D., Qiu, P., & Huang, M. (2014). China's patterns of international technological collaboration 1976–2010: A patent analysis study. *Technology Analysis & Strategic Management,* 26(5), 531-546.

Wu, D., Lu, J., & Zhang, G. (2011). Similarity measure models and algorithms for hierarchical cases. *Expert Systems with Applications, 38*(12), 15049-15056.

Yau, C.-K., Porter, A., Newman, N., & Suominen, A. (2014). Clustering scientific documents with topic modeling. *Scientometrics,* 100(3), 767-786.

Yoon, B., & Song, B. (2014). A systematic approach of partner selection for open innovation. *Industrial Management & Data Systems,* 114(7), 1068-1093.

Yoon, J., Park, H., & Kim, K. (2013). Identifying technological competition trends for R&D planning using dynamic patent maps: SAO-based content analysis. *Scientometrics,* 94(1), 313-331.

Zadeh, L. A. (1965). Fuzzy sets. *Information and Control,* 8(3), 338-353.

Zhang, Y., Porter, A. L., Hu, Z., Guo, Y., & Newman, N. C. (2014a). "Term clumping" for technical intelligence: A case study on dye-sensitized solar cells. *Technological Forecasting and Social Change, 85*, 26-39.

Zhang, Y., Zhou, X., Porter, A. L., & Gomila, J. M. V. (2014b). How to combine term clumping and technology roadmapping for newly emerging science & technology competitive intelligence: "Problem & Solution" pattern based semantic TRIZ tool and case study. *Scientometrics,* 101(2), 1375-1389.

Zhang, Y., Zhang, G., Chen, H., Porter, A. L., Zhu, D., & Lu, J. (2016a). Topic analysis and forecasting for science, technology and innovation: Methodology and a case study focusing on big data research. *Technological Forecasting and Social Change,* 105, 179-191. doi: 10.1016/j.techfore.2016.01.015

Zhang, Y., Zhang, G., Zhu, D., & Lu, J. (2016b). Science evolutionary pathways: Identifying and visualizing relationships for scientific topics. *The Journal of the Association for Information Science and Technology*, to appear.

Zhang, Z., Lin, H., Liu, K., Wu, D., Zhang, G., & Lu, J. (2013). A hybrid fuzzy-based personalized recommender system for telecom products/services. *Information Sciences,* 235, 117-129.

Zhou, X., Porter, A., Robinson, D. K., & Guo, Y. (2013). Analyzing research publication patterns to gauge future innovation pathways for nano-enabled drug delivery. *2013 Proceedings of PICMET'13: Technology Management in the IT-Driven Services (PICMET)*, San Jose, USA.

Zhou, X., Zhang, Y., Porter, A. L., Guo, Y., & Zhu, D. (2014). A patent analysis method to trace technology evolutionary pathways. *Scientometrics,* 100(3), 705-721.

Zhu, D., & Porter, A. L. (2002). Automated extraction and visualization of information for technological intelligence and forecasting. *Technological Forecasting and Social Change,* 69(5), 495-506.

Appendix A. Details of the construction of the 3-level tree structure

In this appendix, we provide the technical details of the construction of the 3-level tree structure, which includes related algorithms and some examples.

1) The term clumping process

We proposed a term clumping process (Zhang et al. 2014a) to combine quantitative analyses and expert knowledge for the cleaning, consolidating, and clustering of scientific and technological terms. The related efforts applied in this paper include:

- Thesaurus-based meaningless term removal (e.g., pronoun, preposition, and common academic terms, such as methodology, conclusion, and comment);
- Stem-based term consolidation (e.g., singular/plural, and other variations of nouns);
- Association rule-based term consolidation: terms sharing more than two words are consolidated (e.g., dye sensitized solar cell and dye sensitive solar cell), and low-frequency terms are consolidated with the high-frequency terms that appear together in a number of records;
- Expert knowledge-based pruning.

2) The portfolio-term matrix

The portfolio-term matrix describes the direct relationships between patent portfolios and the core terms. We denote the number of the patent portfolios and the core terms as n and f respectively. $\delta_{i,j}$ is the term frequency of the core term t_j $(j \leq f)$ in the patent portfolio x_i $(i \leq n)$. Thus, the portfolio-term matrix M could be described as:

$$\mathrm{M} = \begin{bmatrix} \delta_{1,1} & \cdots & \delta_{1,f} \\ \vdots & \ddots & \vdots \\ \delta_{n,1} & \cdots & \delta_{n,f} \end{bmatrix}$$

3) The ST&I data-oriented K-means clustering algorithm

This algorithm was first proposed for a high accuracy clustering performance for the United States (US) National Science Foundation (NSF) granted proposal data (Zhang et al. 2016a). We divided the features of ST&I data into general features (e.g., title and abstract terms) and specific features (e.g., program element code of the US NSF proposals, and the IPC of the patents). We also compared normal term frequency (TF) and the TFIDF-weighted values of terms. For best performance, we automatically calculated a K-value within a selected interval and assembled the best feature set. Specifically, the improved K-means clustering algorithm applied in this paper includes the following steps:

- *Build assembled sets with blended features*. Since the dataset only includes title and abstract terms, we only built 4 assembled sets (i.e., title terms + abstract terms (with normal TF or TFIDF); and inverse ratio weighted title terms + abstract terms (with normal TF or TFIDF));
- *Build the training set with expert knowledge* (in some cases several training sets would be better). We designated all the terms in one patent portfolio as the training set, which would usually number approximately 50–100, and asked technical experts to help classify these terms into several groups;
- *Establish the parameters for best performance.* This process was aided by a validation model to ensure clustering accuracy (e.g., the local optimum K-value and the best feature set);
- *Apply the algorithm to the rest of the data set*. Technical experts helped to validate the results manually, and if the results of some portfolios did not perform well, we would change the K value by plus or minus 1 and re-run the algorithm. The best feature set remained fixed.

The algorithm in this paper is applied to each patent portfolio, and the terms in each are grouped into several clusters. An extreme situation occurs if the total number of terms in one patent portfolio is smaller than a threshold. In these cases, each term is considered to be a cluster and all are set as Level 2 leaves linked directly to the root. The threshold level depends on the specific data corpus, but we set it at a default value of 5.

4) The prevalence value

We designed the prevalence value to take term frequency and the proportion that the term occupies in each related record into consideration. We also provide a way to highlight the most representative terms (Zhang et al. 2014b). We denote that the patent portfolio x_i contains a corpus with ϕ records and a number of terms. If a cluster $C = \{t_1, t_2, \dots, t_i, \dots, t_{\alpha-1}, t_\alpha\}$ includes α distinct terms that relate to the record corpus $D = \{d_1, d_2, \dots, d_j, \dots, d_{\beta-1}, d_\beta\}$, the prevalence value of the term t_i could be calculated as:

$$P(t_i) = \frac{\beta}{\phi} \times \sum_{j=1}^{\beta} \frac{\text{the frequency that } t_i \text{ appears in } d_j}{\text{the total number of terms that appears in } d_j}$$

Appendix B. Statistical information of the entire firms

Table

Statistical information of the entire firms

Group	*FirmName*	*#Patent*	*#Term*	**Term*	*#IPC*	**IPC*
C	Aikekai Technology (Beijing)	3	11	30	3	7
C	Beijing Boren Yongtai Medical	2	8	23	2	2
B	Beijing Choice Electronic Tech	97	19	60	10	108
C	Beijing Transeasy	21	18	46	7	32
B	Capitol Bio Group	11	6	12	5	12
A	Changchun Dirui	103	47	136	11	100
C	Changzhou Huawei medical supplies	4	9	25	3	4
C	DINO Medical & Rehabilitation	13	9	24	10	42
C	Foshan anle medical apparatus	1	0	0	1	1
C	Guangzhou Baizhou Medical Technology	2	8	20	1	1
C	Guangzhou Di Ao Biotechnology	8	9	29	4	59
B	Guangzhou Weili	21	14	37	7	22
B	Guangzhou wondfo biotech	8	0	0	5	58
C	HEBEI XINLE SCI&TECH	14	20	59	9	24
C	Jiangsu Aiyuan Medical	9	15	44	2	2
C	Jiangsu Sanwe Medical Science and Technology	16	14	39	2	6
C	Jiangsu sainty Medical	9	10	34	10	98
C	Jiangyin EVERISE Medical Equipment	53	11	35	1	5
A	Lepu Beijing	94	71	157	18	83
C	Liaoning Jiuzhou	10	14	51	4	10
C	LinaTech	1	12	43	5	53
B	Micro Tech (Nanjing)	74	26	73	10	135
C	Nanjing Fuzhong medical	5	13	40	1	1
C	Nanjing Xufei Medical	2	13	36	1	2
B	Ningbo David	21	7	15	7	22
B	Rayto Life and Analytical Sciences	16	22	65	10	37
C	SID Medical Technology	2	9	30	1	1
A	Sannuo Biological Sensor	16	11	25	5	17
B	Shaanxi Qinming Medical	6	15	38	2	18
B	Shandong Dazheng	5	23	52	7	25
C	Shandong Huge Dental Material	12	9	26	9	12
C	Shandong Jihai Medical	18	13	33	10	20
C	Shanghai Bai Jin medical	5	9	19	2	3
B	Shanghai Chenguang Medical Technologies	32	24	74	5	49
C	Shanghai Comermy Medical Devices	4	11	31	2	5
C	Shanghai Guanghao medical instrument	4	10	35	1	1
A	Shanghai Kindly	100	71	243	15	104
A	Shanghai Kinetic	31	23	62	4	33
B	Shanghai Medical Instr	36	21	44	15	39
C	Shanghai Puyi Medical Instruments	6	10	25	2	5
C	Shanghai Shuangshen Medical Instrument	4	12	41	2	3
A	Shanghai United Image	239	124	374	15	269

C	Shanghai yodo medical technology	8	0	0	5	7
C	Shanghai Zhenghua Medical Equipment	24	8	23	10	36
A	Shenyang Neusoft	98	47	103	12	102
C	Shenyang Xinsong	5	7	9	4	6
C	Shenyang Yonglin medical	1	12	35	1	3
C	Shenzhen new industries biomedical engineering	2	0	0	6	28
C	Shenzhen jumper medical equip	12	0	0	2	13
C	Shenzhe Genius	56	15	21	6	37
B	Shenzhen Well.D	35	23	56	3	32
C	Suzhou Nissei	6	2	2	2	4
C	Suzhou branch of medical science and Technology	5	8	19	1	1
C	The sage (Beijing) Medical Technology	7	9	21	1	2
A	Tianjin Kodon	6	34	105	7	57
C	Tianjin MEDA	12	16	46	3	7
C	Tianjin zhixin hongda medical equip dev	15	0	0	1	1
C	UNIMED Medical Supplies	4	11	37	1	3
A	Wandong Medical Equipment	2	0	0	7	20
C	Wuhan Her Medical	4	4	4	1	3
C	Zhangjiagang Hualing Medical Equipment Manufacturing	55	7	15	2	2
C	Zhangjiagang Jinxiang Medical Equipment	8	13	40	3	6
B	Zhejiang Jimin	26	12	27	2	12
A	Zhuhai Hejia	101	34	61	13	55
B	Zhejiang tiansong medical instrument	2	0	0	4	5

Note. #Patent: the number of patents; #Term: the number of terms; *Term: the frequency of terms; #IPC: the number of IPCs; *IPC: the frequency of IPCs.

Appendix C. Portfolio Correlation Maps

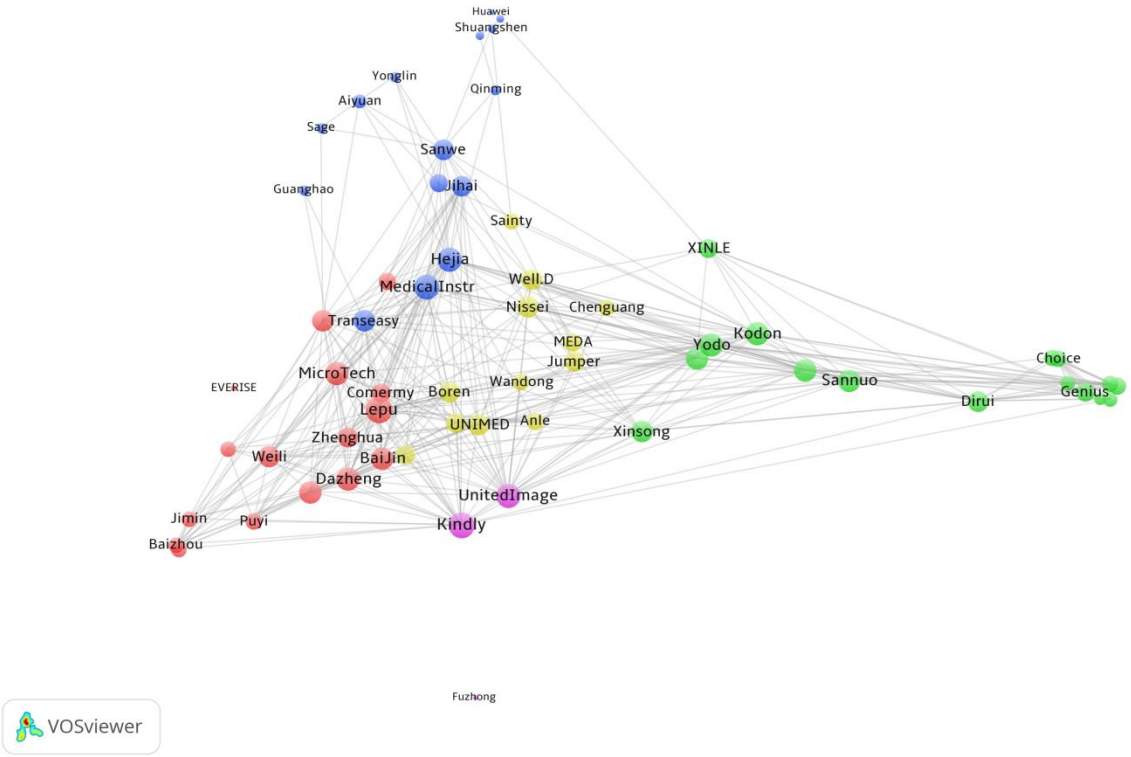

Figure D-1. Portfolio correlation map (based on the results of categorical similarity measure model).

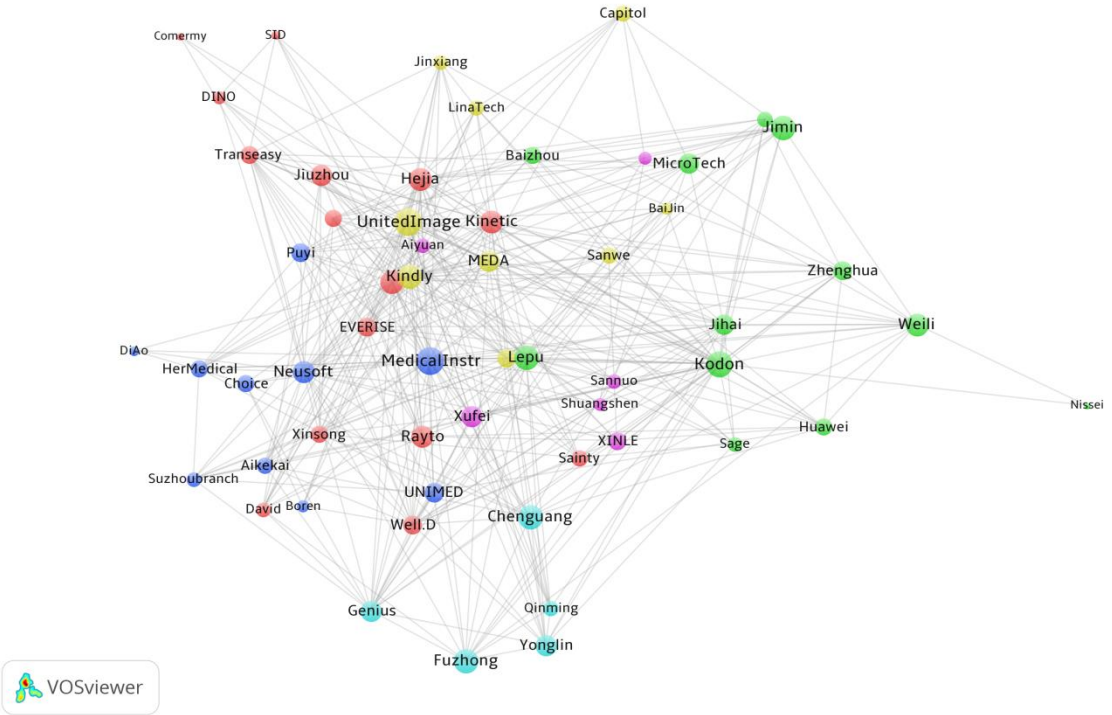

Figure D-2. Portfolio correlation map (based on the results of semantic similarity measure model).