Scientometrics
https://doi.org/10.1007/s11192-019-03223-8

Structural decomposition of technological domain using patent co-classification and classification hierarchy

Changbae Mun[1] · Sejun Yoon[1] · Hyunseok Park[1]

Received: 10 April 2018

Abstract
This paper proposes a new method for decomposing a technological domain (TD). Specifically, the method identifies sub-TDs at the different levels of technological hierarchy within the TD based on the characteristics of patent co-classification and classification hierarchy. We defined the smallest class, named Minimum Overlapped Class (MOC), constructed by overlaps of sub-group IPC(s) and sub-class UPC(s), and sub-TD is basically identified as a set of the MOCs. In order to cluster the MOCs, technological distances among MOCs are calculated based on patent co-classification and hierarchical structure of patent classification systems. Technologically similar MOCs are grouped by using a hierarchical clustering and the identified clusters at the different level of hierarchy show the hierarchical structure of a TD. Detailed technological content for each sub-TD is represented by extracting representative keywords through a text-mining technique. The method is empirically tested by the solar photovoltaic technology and the results show that the identified sub-TDs are reasonably acceptable by qualitative analysis.

Keywords Classification overlap method (COM) · Patent co-classification · Classification hierarchy · Sub-technologies · Sub-domain · Hierarchical class similarity

Introduction

Technologies has been considered as the most important driver for economic development. There has been much effort to analyze technologies for better understanding on technological changes and economic impacts (Schmoch et al. 2003; Evenson and Puttnam 1988; Verspagen et al. 1994; Lybbert and Zolas 2014). Patents, as one of the most important technical data, have been widely used for technological research. For patent analysis, collecting a set of patents related to a specific technology of interest is fundamentally important. Basically, the term 'technology' is flexible and so can be used with different levels of specificity for different research purposes (Benson and Magee 2015b), For example, some technologies can be considered as sub-technologies in other technological domains. Therefore, it

✉ Hyunseok Park
hp@hanyang.ac.kr

[1] Department of Information Systems, Hanyang University, Seoul, Republic of Korea

is essential to identify a set of patents for any level of technological specificity for different purposes of technological research.

In this paper, we propose a method to decompose a technology into sub-technologies. Specifically, the method objectively and reproducibly identifies sub-technologies at different levels of hierarchy and relevant patents for each of them. The decomposition starts from a technological domain (TD). The TD can be defined as 'the set of artifacts that fulfill a specific generic function utilizing a particular, recognizable body of knowledge' (Magee et al. 2016). The TD is an appropriate level for technological decomposition, because TD is not only focusing on a specific technological knowledge, but also a comprehensive boundary which includes different detailed technologies whose underlying functionality and scientific effect are basically same. The patents related to a specific TD can be isolated by using Classification Overlap Method (COM), suggested by (Benson and Magee 2013, 2015b). The COM identifies patents of interest based on the characteristics of co-classification in multiple patent classifications and the overlapped classifications between two different patent classification systems, e.g. IPC and UPC, represent a specific TD with high completeness and relevancy; an average relevancy of 44 TDs is 85.54% (Benson 2014; Benson and Magee 2016). The decomposition process starts with identifying the minimum overlapped classes (MOCs) within the TD, and then MOCs are clustered at different levels of hierarchy by the MOC distances based on the co-classification and class hierarchy. To test the proposed method, we conducted an empirical analysis using patents related to the Solar photovoltaic technology. The results show that TD is well decomposed into acceptable number of clusters, i.e. sub-TDs, with meaningful size of patents at the different levels of technological hierarchy. We qualitatively analyzed the identified clusters and their involved patents. The clusters at the same level of hierarchy has high independency and patents in each cluster are also highly relevant to each other and toward to the overall inventive knowledge for the cluster. In addition, we found that some identified sub-TDs are technologically irrelevant to the TD of interest, i.e. noise patents collected by COM, and so the method can be used for improving the accuracy of a patent set for a TD. We expect the proposed method enriches various researches related to technological changes at subdomain level, modular design, technological tree generation, or functional analysis.

The rest of this paper is structured as follows: "Theoretical background" section reviews theoretical background for this research, "Methodology" section describes the proposed method, "Empirical study: solar photovoltaic technology" section presents the empirical analysis and discussion of the results. Finally conclusion and discussions are drawn in "Conclusion and discussions" section.

Theoretical background

Patents for technological research

Patents are a powerful data source for technological research because the patents include up-to-date and reliable technological information. Most of existing technological knowledge, including software or business method, is patented and many of them are only publicly opened through the patents. Moreover, patents are evaluated their technological novelty by professionals, so every patent has novel knowledge that shows continual progress in the technological developments. Patents are basically a well-structured document. It consists of bibliographic information, such as title, date, assignee, inventors, and references, and detailed description on a specific technology, which enables to use many different types of analytic approaches for various purposes. Therefore, patents are useful not only for understanding of technological changes (Benson and Magee 2015a; Magee et al. 2016; Park and Magee 2017; Verspagen 2007; Fleming 2001; Fleming and Sorenson 2001; Jaffe and Trajtenberg 2002; Jaffe et al. 1993; Martinelli 2012; Verhoeven et al. 2016; Mina et al. 2007; Yan and Guan 2018; Guan and Liu 2016; Von Wartburg et al. 2005; Basnet and Magee 2017), but also for technical or managerial problem-solving purposes (Fu et al. 2015; Park et al. 2013a, b; Yoon and Kim 2011; Choi et al. 2018; Walter et al. 2017; Murphy et al. 2014; Moehrle et al. 2005; Carley et al. 2018; Wang et al. 2017). In particular, patents have one of the most complete technological classification systems, such as IPC (International Patent Classification), UPC (United States Patent Classification), or CPC (Cooperative Patent Classification). The patent classifications are such a backbone of the patent system and make patents more valuable and useful for various purposes. However, use of a patent classification is insufficient to identify a set of patents for a specific TD. This is the reason that many studies have focused on developing a patent search approach.

Classification overlap method (COM)

A keyword-based patent searching is the most widely adopted way to identify patents for a specific TD. Basically this approach identifies all patents containing the defined keywords in their texts. Although this might seem to be acceptable, the keyword-based searching has two critical limitations. First, this usually contains a large number of irrelevant patents to a TD (Trajtenberg 1987; Moeller and Moehrle 2015; Benson and Magee 2013). Considering that the definition of a TD, patents for other applications of the technologies related to the defined keywords should not be included in the patent set for the TD. However, these application patents usually contain the keywords and so they are inappropriately included in the patent set. For example, US9876467 (Wiring structure for solar cell roof) clearly contains the term ‘solar cell’, but this technology is about effective wiring structure on the automobile roof and the solar cells (or panels) are simply used as one component that generate electric energy. Therefore, the patents must not be included in the solar cell domain. Second, since technological developments are a continual progress by recombination of existing knowledge (Strumsky and Lobo 2015; Youn et al. 2015; Weitzman 1998; Antonelli 2011; Fleming 2001), an emerging TD, such as CT (computed tomography), MRI (magnetic resonance imaging), or 3D printing, is actually not totally new, but mainly based on many previous base technologies which are essentially identified for analyzing the TD. However, these base technologies usually do not have the terms related to the emerging TD and so a keyword-based searching is fundamentally difficult to find all relevant patents to the TD (Trajtenberg 1987; Benson and Magee 2013). For example, MRI was emerged in early 1970s and coined by Paul C. Lauterbur at that time. However, there existed many base technologies for MRI developed before 1970s, but none of them uses the term MRI. Therefore, a keyword-based searching is insufficient to collect a complete set of patents for the TD.

The COM is a relatively new patent searching approach which overcomes the two limitations (Fig. 1). The main idea behind the COM is patent examiners classify patents into different classification systems by using different criteria and the overlapped classes by two different classification systems provide more specific boundary than classes in each system. (Benson and Magee 2013, 2015b) showed that the COM provides a high performance in identifying patents related to a specific TD. The procedure of the COM

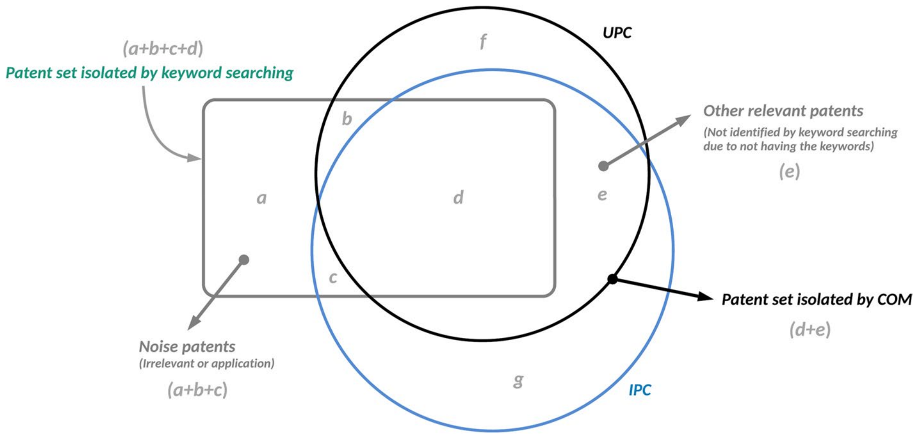

Fig. 1 Conceptual relationship between keyword-based search and COM. Note: alphabets (*a–g)* denote set of patents only belongs to specific area

is as follows. First, patents related a TD is first searched by using simple keywords related to the TD. Second, the representative IPCs and UPCs for the TD are identified by calculating *recall (# patents in the pre-search within the patent class/# collected patents in the pre-search*), *precision (# patents in the pre-search within the patent class/# patents in patent class*), and then *mean-precision-recall* ((*precision* + *recall)/*2). Third, IPCs and UPCs combinations that show high *mean-precision-recall* are identified and the identified set is evaluated by manual reading of patents by technically-knowledgeable people.

Hierarchical structure in patent classification system

Patent classification systems are constructed based on hierarchical structures. Basically, upper level classes delineate a specific technology and lower level classes, subsumed in upper level classes, usually delineate more specific technologies, such as processes, and structural and functional features, within the scope of the upper level classes. The classes at the same level have an equivalent depth in the hierarchy and technological independency between them.

This paper used IPC and UPC as two different patent classification systems for the COM and the proposed method. IPC basically has five hierarchies: section, class, subclass, main-group and sub-group. In sub-groups, there exist one more hierarchical structure and it is represented by the number of dot (*n*-dot sub-group). The deepest hierarchy in the sub-group for IPC (Rev. 2017) is the 9-dot sub-group. UPC has broadly classified into two hierarchies: class and sub-class. Sub-classes inherit the basic characteristics of the parent classes and the hierarchies within the sub-class is represented by dots. Each sub-class begins with mainline sub-class which has no dot and then depth is expressed by the number of dots (*n*-dot indent sub-class) under the mainline. The deepest hierarchy in the sub-class for UPC (Ver. 2013) is the 11-dot sub-class. Figure 2 shows examples of specific hierarchy in IPC and UPC.

IPC (International Patent Classification)

Section:	H		Electricty
Class:	H01		Basic Electric elements
Subclass:	H01F		Magnets
Main group:	H01F	1/00	Magnets or magnetic bodies characterized by the magnetic materials therefor
One-dot sub-group:		1/01	● Of inorganic materials
Two-dot sub-group:		1/03	● ● Characterized by their coercivity
Three-dot sub-group:		1/032	● ● ● of hard-magnetic materials
Four-dot sub-group:		1/04	● ● ● ● Metals or alloys
Five-dot sub-group:		1/047	● ● ● ● ● Alloys characterized by their composition
Six-dot sub-group:		1/053	● ● ● ● ● ● Containing rare earth metals
⋮		⋮	⋮

UPC (United States Patent Classification)

Class:	210		Liquid purification or separation
Mainline subclass:		600	Processes
One-dot incent subclass:		601	● Treatment by living organism
Two-dot incent subclass:		602	● ● Including plant or animal of higher order
Two-dot incent subclass:		603	● ● Including collecting or storing gas
Three-dot incent subclass:		604	● ● ● And reusing oxidant
Two-dot incent subclass:		605	● ● Anaerobically, with subsequently aerobically treating liquid
⋮		⋮	⋮

Fig. 2 Hierarchical structure of IPC (H10F) and UPC (210)

Methodology

The proposed method decomposes a specific TD identified by COM into multiple sub-TDs at different levels of hierarchy. Specifically, the method identifies sub-TDs by the patent co-classification and classification hierarchy based distance measurements. The overall procedure of the method is as follows (Fig. 3). First, MOCs existing within the TD are identified. Second, technological distance between MOCs are calculated based on patent co-classification and hierarchical structure of patent classifications. Third, clusters whose size is acceptably large and independency is relatively high are identified

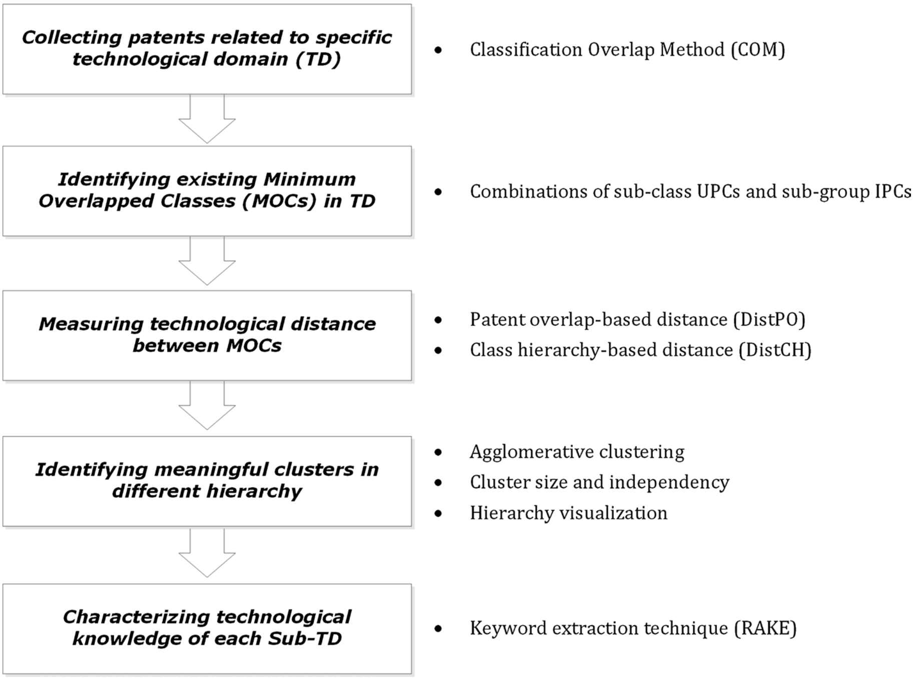

Fig. 3 Overall process of the method

at different levels of hierarchy. Lastly, overall technological knowledge of each cluster is characterized using a keyword extraction technique.

Identifying existing MOCs in TD

The proposed method first identifies the existing MOCs in a TD. We defined MOC as the smallest classification based on the concept of the COM. IPC and UPC can have about 1.2 billion MOCs [73,270 of IPCs (Rev. 2017) times 165,998 of UPCs (Rev. 2013)] and therefore patents classified into the same MOC can be considered to be technologically very similar inventions. MOCs related to a specific TD are the combinations of sub-classes of IPC and sub-groups of UPC related to the TD. For instance, the IPC and UPC combination for Solar PV technology is H01L and 136. MOCs related to the TD are combinations of sub-classes under H01L and 136. H01L-21/02 and 136/201 can be one example of MOC in the solar PV domain. We consider that MOCs having at least one patent as the existing MOCs in TD.

Measuring technological distance between MOCs

To identify sub-TDs in the TD, the technological distance between MOCs are calculated. We developed two metrics based on the characteristics of patent co-classification and hierarchical structures of patent classification systems.

Patent overlap-based distance ($\mathrm{Dist_{PO}}$)

Since patents are classified into multiple patent classifications, MOCs including same patents at the same time can be considered as technologically similar MOCs. To calculate the degree of the distances, we adopted a vector space model (Salton et al. 1975) to transform each MOC into vector (Fig. 4). The distance between MOCs are calculated by cosine distance measurement:

$$\mathrm{CosDist}\left(\mathrm{MOC}_i, \mathrm{MOC}_j\right) = 1 - \frac{M_i \cdot M_j}{M_i M_j}, \tag{1}$$

where M_i (and M_j) is a vector for the MOC_i (and MOC_j), $M_i \cdot M_j$ is a dot product of two vectors M_i and M_j, and the range of the distance is [0,1]. However, the dimensionality of a MOC vector, i.e. # of patents in a TD, is usually very high and this may lead to little difference in distance among MOCs. The logistic function is applied to give a clear difference in MOC distances and the formulation for $\mathrm{Dist_{PO}}$ (Patent overlap-based distance) is as follows:

$$\mathrm{Dist_{PO}}\left(\mathrm{MOC}_i, \mathrm{MOC}_j\right) = \frac{1}{1 + \mathrm{e}^{-10\left(\mathrm{CosDist}\left(\mathrm{MOC}_i, \mathrm{MOC}_j\right) - 0.5\right)}}. \tag{2}$$

Class hierarchy-based distance ($\mathrm{Dist_{CH}}$)

As mentioned in "Hierarchical structure in patent classification system" section, patent classification is constructed based on a technological hierarchy between the upper and lower classes. This structure is very similar to a conceptual hierarchy for semantic analysis, such as WordNet. Therefore, we applied a concept hierarchy-based similarity measurement to calculate technological distance between two classes. First, patent classification is transformed into a hierarchical network, like a concept hierarchy, whose node is a class and edge is a hierarchical relationship between two classes (Fig. 5). Then, technological distances are calculated by a path length-based similarity measurement. The path length is a score representing a count of edges between two classes in the shortest path and, thus, the shorter the path between two nodes, the more technologically similar they are. Specifically, we used Wu and Palmer's (Wu and Palmer 1994) approach and the equation for technological distance is as follows:

$$\mathrm{Dist_{WP}}\left(C_i, C_j\right) = 1 - \frac{2 \cdot d\left(\mathrm{LCS}\left(C_i, C_j\right)\right)}{d\left(C_i\right) + d\left(C_j\right)}, \tag{3}$$

	Pat_1	Pat_2	Pat_3	Pat_4	Pat_4	⋯	Pat_m
MOC_1	1	0	0	1	1	⋯	1
MOC_2	0	1	1	0	0	⋯	1
⋮							
MOC_N	1	0	0	0	1	⋯	0

Fig. 4 Vector space modeling of MOCs

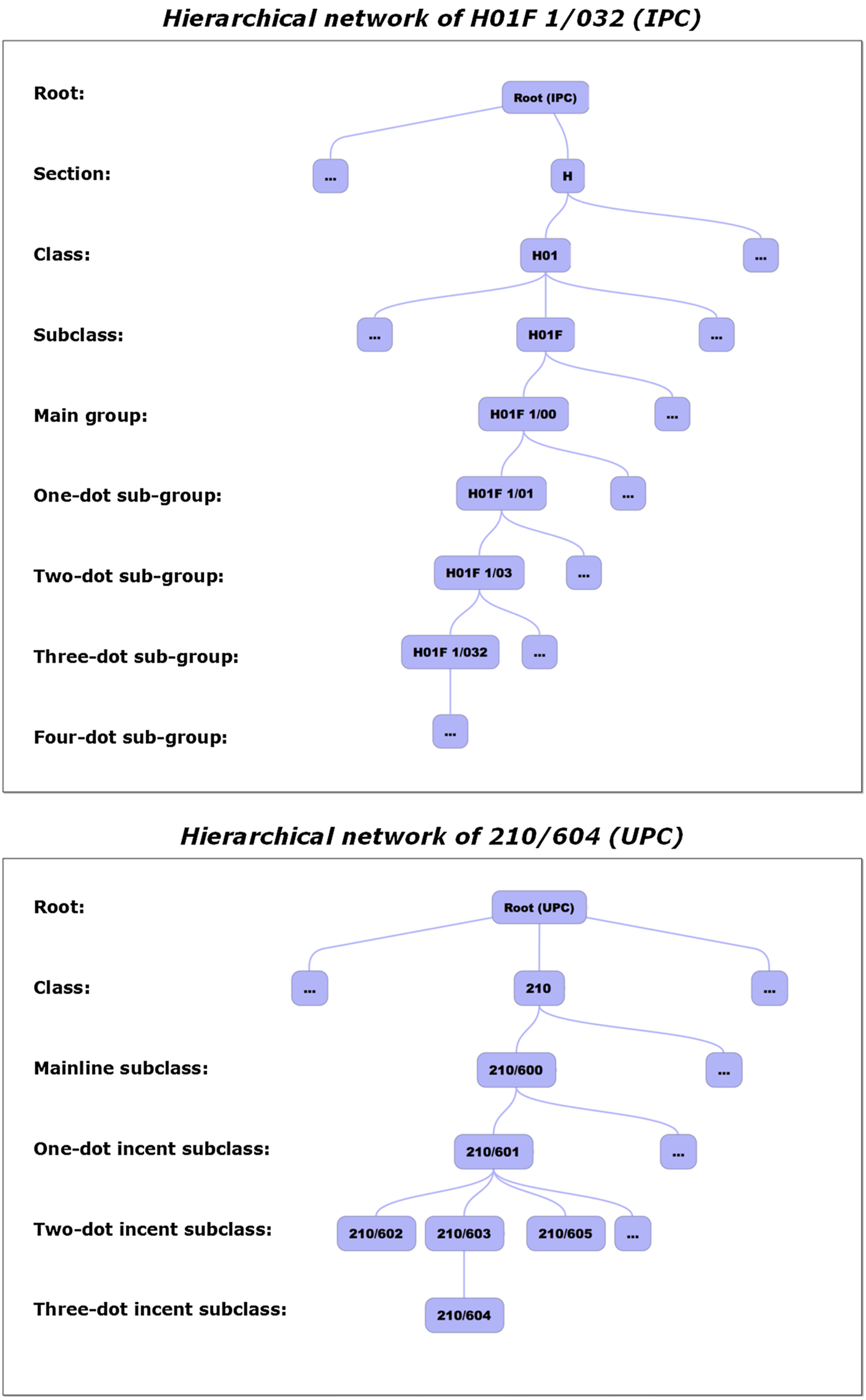

Fig. 5 Visualization of IPC and UPC to hierarchical network

where C_i (and C_j) is a specific class in a patent classification system, d (C_n) is the path length from the root to C_n, LCS (C_i, C_j) is the most specific common ancestor of C_i and C_j in the hierarchy, and the range of the score is [0,1]. Since each MOC is a combination of two different patent classifications, the class hierarchy-based distance ($\mathrm{Dist}_{\mathrm{CH}}$) is calculated based on hierarchical distances of two patent classifications, IPC and UPC. Therefore, $\mathrm{Dist}_{\mathrm{CH}}$ between MOCs can be calculated by:

$$\mathrm{Dist}_{\mathrm{CH}}\left(\mathrm{MOC}_i, \mathrm{MOC}_j\right) = \frac{\mathrm{Dist}_{\mathrm{WP}}\left(\mathrm{IPC}_i, \mathrm{IPC}_j\right) + \mathrm{Dist}_{\mathrm{WP}}\left(\mathrm{UPC}_i, \mathrm{UPC}_j\right)}{2}, \tag{4}$$

where IPC_i and UPC_i (or IPC_j and UPC_j) are IPC and UPC of MOC_i (or MOC_j), and the range is [0,1].

MOC distance ($\mathrm{Dist}_{\mathrm{MOC}}$)

Finally, technological distance between MOCs is calculated. Two MOCs which have high both $\mathrm{Dist}_{\mathrm{PO}}$ and $\mathrm{Dist}_{\mathrm{CH}}$ have and so we defined the $\mathrm{Dist}_{\mathrm{MOC}}$ as follows:

$$\mathrm{Dist}_{\mathrm{MOC}} = \mathrm{Dist}_{\mathrm{PO}} \times \mathrm{Dist}_{\mathrm{CH}}. \tag{5}$$

Identifying meaningful clusters in different hierarchy

Based on the distance scores from "Measuring technological distance between MOCs" section, potential sub-TDs are identified by clustering MOCs at different level of the distance (Fig. 6). Specifically, we used an agglomerative clustering algorithm with an average link method. Since the range of the average distance between clusters is [0,1], we set ten levels of the distance in 0.1 scale. Each level from 1.0 to 0.1 means the depth of hierarchy in a TD and the identified clusters at each level are potential sub-TDs at the level of hierarchy.

The sub-TDs are boundaries for more specific technologies within the TD. Since a sub-TD also have characteristics of TD, they should have an appropriate size which is a meaningful entity to be analyzed and should be relatively independent from other clusters. Therefore, size and independency of a cluster are two critical factors to identify sub-TDs.

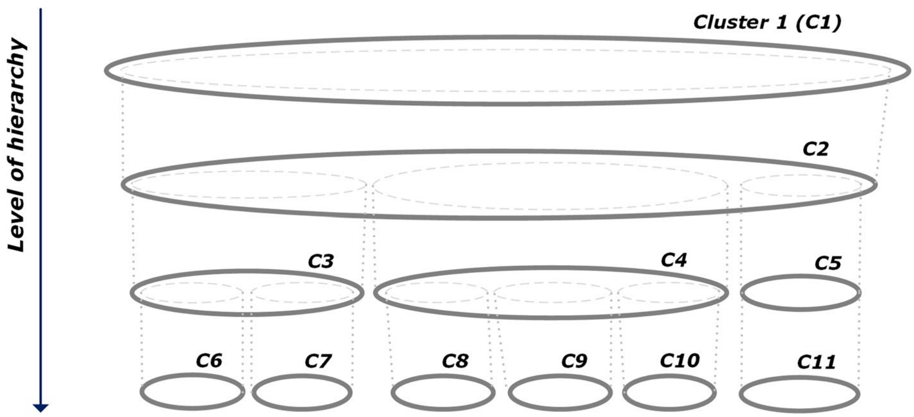

Fig. 6 Example of identified sub-TDs at different depth of hierarchy

We considered the clusters whose size (=# of patents in the cluster/# of patents in the TD) is ≥ 0.1 and an average independency at a specific level of hierarchy is ≥ 0.8 as the sub-TDs: this paper sets relatively strict cut-off values to effectively show the technological structure in the TD, but these cut-off values can be flexibly defined by considering analytic purposes. The average independency (AI) of clusters at the level of hierarchy l is calculated based on the jaccard coefficient and the formulation is as follows:

$$\mathrm{AI}_l = \frac{2}{n(n-1)} \sum_{j}^{n} \sum_{i}^{n} \frac{\left|C_i \cup C_j\right| - \left|C_i \cap C_j\right|}{\left|C_i \cup C_j\right|}, \quad (i < j) \tag{6}$$

where l is a specific level of hierarchy (from 0.1 to 1.0), n is the number of the existing clusters at the level l, and C_n is the n-th cluster at the level l. If clusters in a specific level has low independency (<0.8), all clusters at the level cannot be considered as the sub-TDs.

Characterizing technological knowledge in each cluster

Although sub-TDs, i.e. clusters, at different hierarchies are identified, it is basically difficult to qualitatively represent the specific technological knowledge in each sub-TD. Therefore, we adopted a text mining to extract the representative keywords, including key-phrases, that are mainly occurred only in a sub-TD. The extraction process is as follows: First, candidate keywords in each cluster are extracted by using RAKE (Rapid Automatic Keyword Extraction) algorithm (Rose et al. 2010). RAKE as one of the keywords extraction methods is popular due to its extraction quality and efficiency in any application. Although RAKE is not perfectly suitable for analyzing technical documents, the method, at least, seems to provide enough performance for our purpose, because RAKE extracts not only single words, but also compound nouns or phrases as 'keywords', and its algorithm shows better performance in extracting representative keywords for multiple entities. Second, keywords of a cluster are extracted based on the ranking of essentiality measure (ESS):

$$\mathrm{ESS}(k) = \frac{edf(k)^2}{rdf(k)}, \tag{7}$$

where the k is the focal keyword, *rdf(k)* as the referenced document frequency of k means the number of document in which the keyword k occurred as a candidate keyword and the candidate keywords are all identified noun chunks from a document, and *edf(k)* is the extracted document frequency of k. Third, the top ranking keywords for each cluster are selected by ESS score. The ESS scores for the selected keywords inside and outside a specific cluster are calculated to identify keywords characterizing the cluster from other clusters at the same level (Fig. 7). The representative keywords for a specific cluster i is determined based on the score calculated by cluster-specific score (CSS):

$$\begin{aligned} \mathrm{CSS}_{ij}(k) &= \frac{\mathrm{ESS}_i(k)/n_i}{\mathrm{ESS}_{\bar{i}}(k)/n_{\bar{i}}}, \quad (i \neq j) \\ \mathrm{CSS}_{ij}(k) &= \mathrm{ESS}_i(k)/n_i, \quad (i = 1 \text{ or } i = j), \end{aligned} \tag{8}$$

where j is the least superset of i, $\bar{i}$ is a complementary set of i under the j, k is a candidate keyword, ESS_i*(k)* is ESS of k in the cluster I, and n_i is the number of patents included in

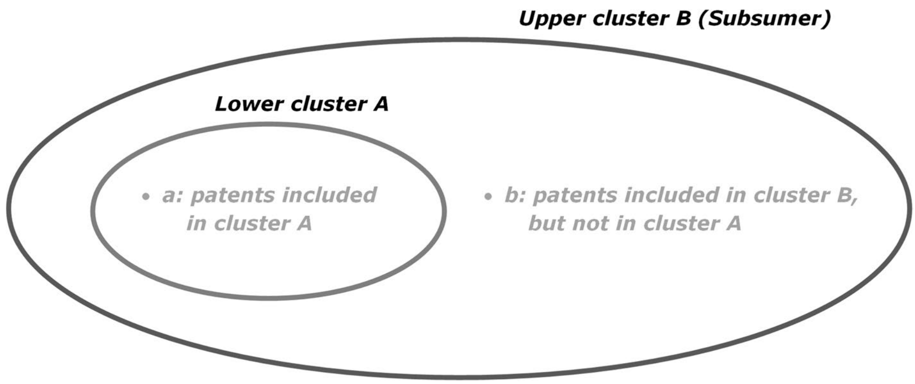

Fig. 7 Target clusters for ESS comparison of keywords

i. Finally, we filtered out common words in technical documents or non-technical words, such as at least, present invention, invention relates and etc., and selected the top 10–20 CSS keywords as the representative keywords for the cluster *i*.

Empirical study: solar photovoltaic technology

In this section, the proposed method is applied to Solar photovoltaic technology to test the practical usability of the method.

Introduction to solar photovoltaic

Solar photovoltaic (PV) is a power generation method that converts solar light directly into electricity. PV has been paid great attention for last few decades as one of the most promising type of clean energy technologies. It broadly consists of three sub-TDs: solar cell, module and panel, and mounting systems. PV cell is a core component that generates electricity. PV module (or panel) is a bundle of PV cells (that practically used for various applications). Mounting systems are related to installation and control of PV systems. Benson and Magee (2013) shows that IPC and UPC combination for solar PV by COM is H01L and 136, and the relevancy of the isolated patents by COM is 85%; this means 15% of the patents might not be closely related to the solar PV technology. The empirical study will focus on identifying the hierarchical structure of the solar PV technology and an irrelevant set of patents, about 15% of whole patents in the TD.

Data

The patents for the solar PV technology was obtained by COM specifically using the overlap between IPC H01L (Semiconductor devices; Electric solid-state devices) and UPC 136 (Batteries: thermoelectric and photoelectric). The number of patents in the set is 5203, from Jan 1, 1976 to July 1, 2013, and the technological relevancy of the patent set is 0.85 (Table 1).

Table 1 Summary of data

Search query	Number of patents	Relevancy	Range
H01L and 136	5203	0.85	US granted patents from 1971.01.01 to 2013.07.01 (application date)

Result

There exist 1606 sub-class level IPCs and 4507 sub-group level UPCs, but only 276 IPCs and 70 UPCs among them are related to the PV domain (IPC: H01L and UPC: 136). So the PV domain has 19,320 MOCs and the number of existing MOCs which have at least one patent is 2755. $Dist_{MOC}$ for all MOCs are calculated by using patent overlaps ($Dist_{PO}$) and class hierarchies ($Dist_{CH}$) in the MOCs. The result for the hierarchical clustering of MOCs by $Dist_{MOC}$ scores is shown in Fig. 8. We considered the clusters whose size is larger than 520 patents (10% of whole patents) and AI (Eq. 6) at each level of hierarchy ≥ 0.8 as the meaningful sub-TDs in the TD (Table 2).

The hierarchical structure of PV domain is visualized in Fig. 9 and Table 3 shows the representative keywords for each cluster. The results show that the TD is clearly related to the solar or photovoltaic cell (C1–C5) and can be broadly divided into three sub-TDs at the level 0.4.

Based on the keywords in Table 3, C8 is about the thermoelectric generator and module, C9 is about solar cell, including amorphous silicon, thin film, and multi-junction cell, and C10 is about PV modules and panels. The C9 and C10 are almost same with

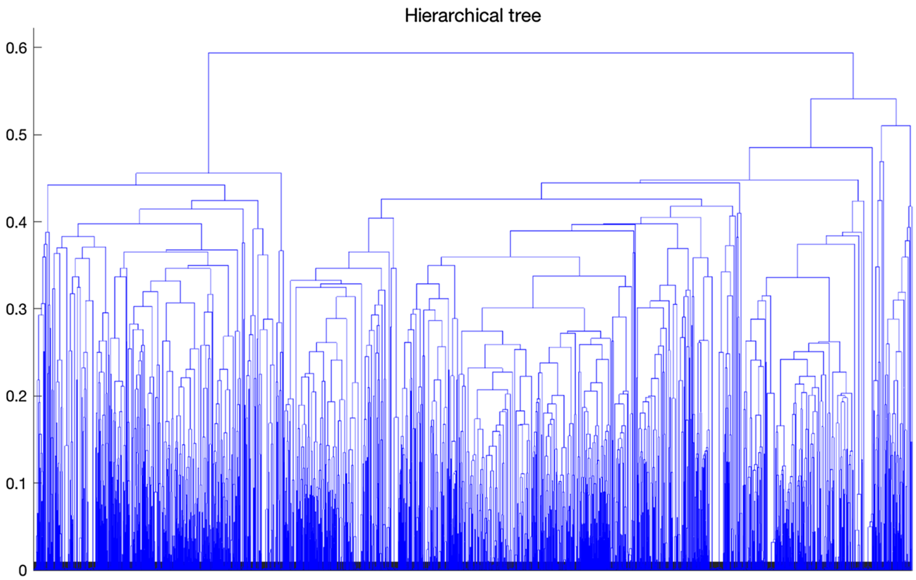

Fig. 8 Hierarchical clustering for MOCs. *Notes* Dendrogram shows how all MOCs are grouped into one cluster from the bottom, which is conceptually reverse process of TD decomposition; *x*-axis (or each branch of tree) denotes MOC and *y*-axis denotes level of structural hierarchy (or technological distance)

Table 2 Average independency at each level of hierarchy

Hierarchy level	Average independency
1.0	N/A (or 1.0)
0.9	N/A (or 1.0)
0.8	N/A (or 1.0)
0.7	N/A (or 1.0)
0.6	N/A (or 1.0)
0.5	0.964
0.4	0.845
0.3	0.819
0.2	0.896
0.1	0.986

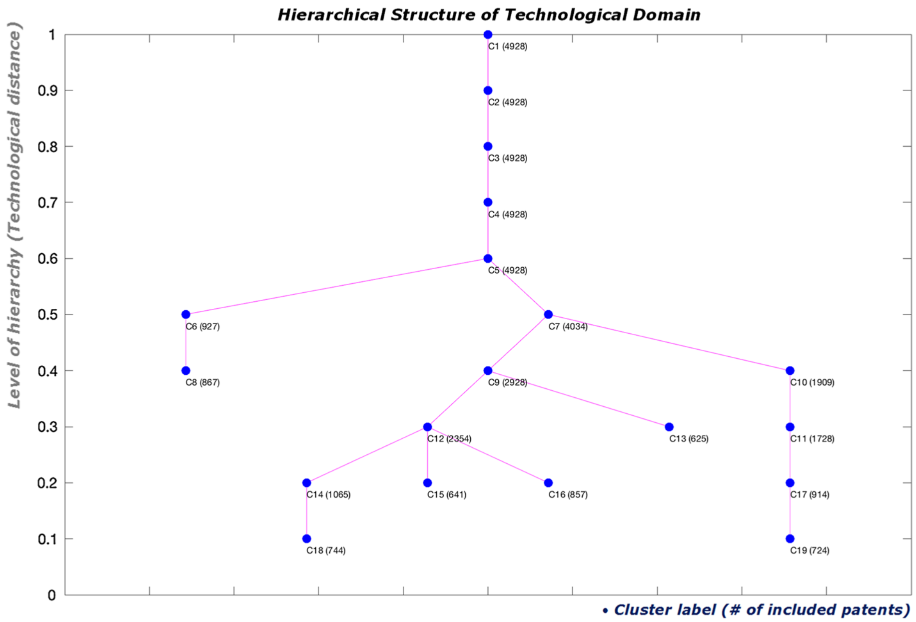

Fig. 9 Hierarchical structure of solar PV. *Notes* linkage denotes direct hierarchical relationship between upper and lower clusters

the widely-accepted classification for the solar PV domain (EPRI 2009), but C8 (and C6 as well) should be considered as the different type of solar power generation technology. Since the concept of Solar PV is based on the photovoltaic effect that converts the energy of sun light directly to electricity, the thermoelectric generation, classified in C8, seems to be an irrelevant set for solar PV; C8 could be the cause of 15% of irrelevancy in COM searching.

Under the C9, there are two sub-TDs at the level 0.3, C12 and C13. C13 is about the crystalline silicon and amorphous silicon cells. Even though C12 is difficult to be characterized by the keywords, sub-TDs of C12 seem to be clear categories: C14, C15, and C16. C14 is about surface texturing for maximizing efficiency, C15 is about

Table 3 Representative keywords for each sub-TD

Cluster #	Representative keywords (CSS score)
C1	Solar cell (0.34), photovoltaic (0.28), module (0.13), temperature (0.13), voltaic cell (0.09), n layer (0.09), photovoltaic cell (0.09), semiconductor layer (0.08), photovoltaic device (0.08), electric conversion (0.08), thin film (0.07), photoelectric conversion (0.06), conductor material (0.05), semiconductor material (0.05)
C2	Solar cell (0.34), photovoltaic (0.28), module (0.13), temperature (0.13), voltaic cell (0.09), n layer (0.09), photovoltaic cell (0.09), semiconductor layer (0.08), photovoltaic device (0.08), electric conversion (0.08), thin film (0.07), photoelectric conversion (0.06), conductor material (0.05), semiconductor material (0.05)
C3	Solar cell (0.34), photovoltaic (0.28), module (0.13), temperature (0.13), voltaic cell (0.09), n layer (0.09), photovoltaic cell (0.09), semiconductor layer (0.08), photovoltaic device (0.08), electric conversion (0.08), thin film (0.07), photoelectric conversion (0.06), conductor material (0.05), semiconductor material (0.05)
C4	Solar cell (0.34), photovoltaic (0.28), module (0.13), temperature (0.13), voltaic cell (0.09), n layer (0.09), photovoltaic cell (0.09), semiconductor layer (0.08), photovoltaic device (0.08), electric conversion (0.08), thin film (0.07), photoelectric conversion (0.06), conductor material (0.05), semiconductor material (0.05)
C5	Solar cell (0.34), photovoltaic (0.28), module (0.13), temperature (0.13), voltaic cell (0.09), n layer (0.09), photovoltaic cell (0.09), semiconductor layer (0.08), photovoltaic device (0.08), electric conversion (0.08), thin film (0.07), photoelectric conversion (0.06), conductor material (0.05), semiconductor material (0.05)
C6	Temperature difference (172.64), heat source (53.23), thermal conductivity (36.69), temperature gradient (9.59), heat sink (7.5), thermal energy (7.5), electric power (4.32), high temperature (3.76), insulating material (3.62), electric current (3.28), electrical power (2.72), electrical energy (2.16), electrically connected (1.66), insulating layer (1.52), conductive material (1.44), type semiconductor (1.25), semiconductor material (0.93), second electrode (0.59), solar energy (0.51), thin film (0.38), electric material (0.11), thermoelectric material (0.1), thermoelectric device (0.1), thermoelectric element (0.09), electric generator (0.07), thermoelectric module (0.06), thermoelectric conversion (0.06), thermoelectric generator (0.06), hot junction (0.04), heat transfer (0.04), cold junction (0.03), thermoelectric power (0.03), thermocouple junction (0.03), heat exchange (0.03), power generation (0.03), thermocouple assembly (0.03), thermally conductive (0.03), thermoelectric semiconductor (0.03), heat exchanger (0.02), electric converter (0.02), conductor element (0.02), heat pump (0.02), thermal contact (0.02)
C7	photovoltaic device (43.66), solar cell (21.41), photovoltaic cell (15.29), semiconductor layer (12.35), conversion efficiency (3.82), solar energy (3.79), thin film (2.69), semiconductor material (1.22), voltaic cell (0.11), n layer (0.1), electric conversion (0.08), photoelectric conversion (0.08), conductor material (0.06), amorphous si (0.06), amorphous silicon (0.06)
C8	Thermoelectric device (3.11), thermoelectric material (1.61), cold junction (1.04), semiconductor material (0.46), electrical energy (0.39), electric material (0.12), thermoelectric element (0.1), electric generator (0.07), thermoelectric module (0.07), electric power (0.06), thermoelectric conversion (0.06), thermoelectric generator (0.06), thermal conductivity (0.06), electrically connected (0.05), temperature difference (0.05), heat source (0.04), hot junction (0.04), high temperature (0.04), thermal energy (0.04), thermoelectric power (0.04), heat sink (0.04), type semiconductor (0.04), heat transfer (0.03), insulating material (0.03), thermocouple junction (0.03), electrical power (0.03), thin film (0.03), thermocouple assembly (0.03), power generation (0.03), conductive material (0.03), thermally conductive (0.03), heat exchanger (0.03), electric converter (0.03), thermoelectric semiconductor (0.03), temperature gradient (0.02), heat pump (0.02), conductor element (0.02), electric current (0.02), thermal contact (0.02), heat flow (0.02), thermoelectric converter (0.02)

Table 3 (continued)

Cluster #	Representative keywords (CSS score)
C9	Conductive layer (2.5), amorphous silicon (2.44), semiconductor layer (2.4), photovoltaic element (1.94), photovoltaic device (1.88), photoelectric conversion (1.69), thin film (1.56), semiconductor material (1.39), n junction (1.31), solar cell (1.3), voltaic cell (1.12), photovoltaic cell (1.1), solar cell module (0.53), solar energy (0.4), n layer (0.11), electric conversion (0.09), amorphous si (0.07), conversion efficiency (0.06), n solar cell (0.06), n device (0.05), cell includes (0.05), electrically conductive (0.04), conversion device (0.04), conductivity type (0.04)
C10	Photovoltaic module (15.31), solar cell module (10.14), solar module (8.91), solar panel (6.61), electrical energy (4.45), solar battery (3.7), solar energy (3.36), receiving surface (2.14), electrically connected (2.08), solar radiation (1.95), photovoltaic element (1.76), electrode layer (1.25), solar cell (1.07), photovoltaic cell (1.06), voltaic cell (1.05), electrical contact (0.92), photovoltaic device (0.89), conductive layer (0.76), n layer (0.66), semiconductor layer (0.66), conversion efficiency (0.62), second electrode (0.62), first electrode (0.61), n junction (0.53), thin film (0.52), amorphous silicon (0.49), semiconductor material (0.48), cell module (0.09), electric conversion (0.07), conductor material (0.04)
C11	Solar cell array (4.82), solar collector (4.61), semiconductor material (3.35), solar energy (2.58), solar cell module (2.35), solar module (1.57), electrical energy (1.22), solar cell (1.02), electrical contact (0.96), semiconductor layer (0.91), photovoltaic device (0.86), thin film (0.47), voltaic cell (0.11), photovoltaic cell (0.11), cell module (0.1), n layer (0.08), electric conversion (0.07), photovoltaic module (0.06), solar panel (0.06), solar battery (0.05), photovoltaic element (0.04), conductor material (0.04), conversion efficiency (0.04), amorphous silicon (0.03), conductive layer (0.03), electrode layer (0.03), receiving surface (0.03), n junction (0.03), solar radiation (0.03), solar battery module (0.02)
C12	Electrical contact (2.57), conversion efficiency (2.32), semiconductor substrate (2.23), solar energy (1.91), solar cell includes (1.74), photovoltaic cell (1.63), solar cell module (1.6), solar cell (1.49), photovoltaic element (1.4), photovoltaic device (1.36), back surface (1.25), photoelectric conversion (1.24), conductive layer (1.11), semiconductor material (1.07), conductivity type (1.04), film solar cell (0.95), semiconductor layer (0.91), thin film (0.75), electrode layer (0.75), amorphous silicon (0.57), voltaic cell (0.12), n layer (0.12), electric conversion (0.09), n junction (0.05), electrically conductive (0.04), conversion device (0.04), electrically connected (0.03)
C13	Hydrogenated amorphous silicon (4.91), amorphous semiconductor (4.83), intermediate layer (4.54), transparent conductive film (4.47), crystalline silicon (4.22), crystalline semiconductor (4.09), amorphous silicon layer (4.07), silicon layer (3.78), amorphous silicon (3.68), type semiconductor layer (3.32), photoelectric conversion layer (3.07), transparent electrode (2.8), back electrode (2.76), type semiconductor (2.3), first semiconductor (2.24), semiconductor layer (2.17), transparent conductive layer (2.14), film solar cell (2.12), photoelectric conversion efficiency (2.07), photoelectric conversion device (1.98), type layer (1.89), thin film (1.87), oxide layer (1.84), glass substrate (1.84), transparent substrate (1.77), photovoltaic device (1.73), conductive layer (1.71), semiconductor device (1.61), first electrode (1.53), conversion efficiency (1.51), semiconductor material (1.46), silicon solar cell (1.46), second electrode (1.43), low cost (1.42), semiconductor film (1.37), active layer (1.3), photovoltaic element (1.29), silicon substrate (1.06), solar cell (0.88), n junction (0.82), electrical contact (0.77), voltaic cell (0.63), photovoltaic cell (0.61), electrode layer (0.06)

Table 3 (continued)

Cluster #	Representative keywords (CSS score)
C14	Front surface (4.15), semiconductor body (3.03), back surface (2.75), transparent conductive layer (2.72), electrical contact (2.38), conductive layer (2.15), photovoltaic module (1.95), second electrode (1.88), back electrode (1.88), semiconductor substrate (1.84), oxide layer (1.68), solar battery (1.56), silicon substrate (1.37), transparent substrate (1.29), conductive material (1.26), photoelectric conversion device (1.21), transparent electrode (1.21), solar cell comprising (1.21), ohmic contact (1.21), solar cell (1.19), n layer (1.17), silicon solar cell (1.14), electrically connected (1.12), semiconductor device (1.09), photovoltaic device (1.06), semiconductor layer (0.99), solar cell module (0.97), film solar cell (0.97), thin film (0.85), conversion efficiency (0.85), photovoltaic element (0.81), photovoltaic cell (0.79), electric conversion efficiency (0.77), photoelectric conversion efficiency (0.7), semiconductor material (0.69), solar energy (0.69),n junction (0.6), amorphous silicon (0.56), electrically conductive (0.06), receiving surface (0.03), light incident (0.03)
C15	Sensitized solar cell (6.58), absorber layer (3.92), photoelectric conversion element (2.83), window layer (2.67), substrate wherein (2.51), type semiconductor (2.05), thin film (1.99), photoelectric conversion layer (1.95), electronic device (1.75), thin film solar cell (1.74), electric conversion efficiency (1.7), photoelectric conversion efficiency (1.65), metal oxide (1.64), glass substrate (1.6), semiconductor film (1.58), active layer (1.55), low cost (1.55), film solar cell (1.45), photoelectric conversion device (1.43), second electrode (1.4), semiconductor substrate (1.39), band gap (1.38), conductive material (1.34), semiconductor layer (1.3), oxide layer (1.27), conversion efficiency (1.26), conductive layer (1.18), semiconductor material (1.17), transparent substrate (1.15), silicon substrate (1.13), solar cell (1.12), crystalline silicon (0.99), voltaic cell (0.92), photovoltaic cell (0.89), back surface (0.89), photovoltaic device (0.88), electrical contact (0.81), n junction (0.7), photovoltaic element (0.56), high efficiency (0.03), electric conversion layer (0.03), cell structure (0.03), conductive film (0.02), conductive substrate (0.02)
C16	Junction solar cell (4.8), band gap (2.41), ohmic contact (2.38), n junction (2.28), silicon layer (2.08), type semiconductor (1.97), photoelectric conversion device (1.97), oxide layer (1.86), type semiconductor layer (1.82), silicon solar cell (1.75), semiconductor material (1.63), active layer (1.63), semiconductor substrate (1.61), n layer (1.5), back surface (1.48), semiconductor layer (1.46), transparent substrate (1.43), photoelectric conversion efficiency (1.22), film solar cell (1.19), electric conversion efficiency (1.18), silicon substrate (1.12), conversion efficiency (1.11), electrically connected (1.1), front surface (1.1), solar cell (1.08), amorphous silicon (1.08), photovoltaic device (1.01), conductive layer (0.98), electrical contact (0.93), transparent electrode (0.93), thin film (0.89), photovoltaic element (0.89), first electrode (0.76), photovoltaic cell (0.73), second electrode (0.73), voltaic cell (0.72), solar cell module (0.62), solar energy (0.6), photoelectric conversion (0.11), conductivity type (0.06), crystalline silicon (0.03), n region (0.03), electric conversion layer (0.03)
C17	Solar battery module (8.02), solar cell module (6.88), solar cell module includes (5.34), solar cell element (3.27), bypass diode (3.12), photovoltaic module (2.85), solar battery (2.75), solar module (2.45), photovoltaic modules (2.27), electrically connected (2.21), connected solar cell (2.16), solar panel (2.06), photovoltaic panel (1.88), transparent substrate (1.86), film solar cell (1.86), photovoltaic element (1.56), electrode layer (1.51), solar cell array (1.39), receiving surface (1.21), solar cell (1.19), back surface (1.15), back electrode (1.0), electrical contact (0.92), transparent electrode (0.89), photovoltaic cell (0.8), voltaic cell (0.77), thin film (0.77), photovoltaic device (0.72), conductive layer (0.72), electrical energy (0.67), semiconductor layer (0.49), solar energy (0.45), conversion efficiency (0.42), cell module (0.16), electrical connection (0.03), electrically connecting (0.03), amorphous silicon (0.03), surface side (0.02), top surface (0.02), cell panel (0.02), insulating material (0.02)

Table 3 (continued)

Cluster #	Representative keywords (CSS score)
C18	Photoelectric conversion layer (4.31), semiconductor material (3.97), active layer (3.02), semiconductor body (2.27), reflection coating (1.64), photovoltaic device (1.57), transparent conductive layer (1.51), conductive layer (1.42), electrical contact (1.38), first electrode (1.37), photoelectric conversion efficiency(1.37), second electrode(1.35), semiconductor layer(1.29), oxide layer(1.29), metal contact(1.29), photovoltaic cell(1.26), thin film (1.18), ohmic contact (1.17), semiconductor device (1.17), semiconductor substrate (1.15), silicon solar cell (1.15), back surface (1.11), electric conversion efficiency (1.08), solar cell (0.92), transparent substrate (0.91), front surface (0.86), film solar cell (0.86), front side (0.82), silicon substrate (0.79), photovoltaic element (0.63), n layer (0.14), electrically conductive (0.06), conversion efficiency (0.06), transparent electrode (0.03), electrically connected (0.03), photoelectric conversion device (0.03), n junction (0.03), conductive material (0.03), solar energy (0.03), conductive substrate (0.03), electric conversion layer (0.03), metal layer (0.03), n region (0.03), transparent conductive oxide (0.02), insulating layer (0.02), light incident (0.02)
C19	Solar cell module (5.82), adjacent solar cell (3.67), photovoltaic panel (2.23), photovoltaic module (1.65), solar cell array (1.57), solar battery (1.52), top surface (1.4), solar energy (1.2), solar cell element (1.18), second electrode (1.15), semiconductor layer (1.08), receiving surface (1.05), solar cell (1.01), thin film (0.9), electrically connected (0.81), back electrode (0.73), photovoltaic device (0.7), electrical connection (0.67), photovoltaic cell (0.66), solar panel (0.66), light receiving surface (0.63), transparent electrode (0.45), electrical energy (0.39), cell module (0.19), voltaic cell (0.09), photovoltaic element (0.06), solar module (0.06), solar battery module (0.05), electrode layer (0.04), photovoltaic modules (0.04),first electrode (0.03), electrically connecting (0.03), film solar cell (0.03), transparent substrate (0.03), surface side (0.02), bypass diode (0.02), end portion (0.02), conductive layer (0.02), voltaic array (0.02), solar array (0.02), photovoltaic array (0.02), conversion efficiency (0.02), low cost (0.02), covering material (0.02)

dye-sensitized solar cell and focusing on increasing its efficiency, and C16 is about multi-junction solar cells.

Conclusion and discussions

This paper proposes a new method to technologically decompose a TD. Specifically, the method identifies sub-TDs at the different level of technological hierarchy within the TD. The underlying concept of this research is based on the COM, suggested by (Benson and Magee 2013, 2015b), that shows the combinatorial classifications by two different patent classification systems are effective and reliable to identify patents related to a specific TD. The proposed method defines the smallest classes (MOC), overlapped by sub-group level IPC and sub-class level UPC and calculates technological distances between MOCs based on patent co-classification and hierarchical structure of patent classification systems. Since the distance measurement considers both perspectives, the method can be relatively independent from external factors of TD, such as size of patents in TD. Technologically similar MOCs are clustered by using a hierarchical clustering and the identified clusters at different level of hierarchy show the hierarchical structure of a TD. To identify detailed technological knowledge of each sub-TD, a text mining is applied to extract the representative keywords in each sub-TD. Our empirical study using the solar PV technology shows how a TD can be technologically decomposed by the proposed method. The identified sub-TDs are qualitatively acceptable and, moreover, some sets of patents which are irrelevant to the solar PV are also identified.

In order to improve the performance of the method, we will focus on the following issues in the further work. First, although this research used the fixed level of hierarchy, ten hierarchies by 0.1 scale, further research will use more flexible level of hierarchy by using patent citation information. Stability level of patent citation network in each cluster can be used as an indicator to identify the most meaningful sub-TDs at different level of hierarchy. Second, the quality of keyword extraction in the method seems to be good to characterize the specific technological knowledge of a sub-TD. But, qualitative filtering and some domain knowledge are still required to perfectly characterize the sub-TDs. Thus, we will focus on developing a keyword extraction algorithm that can at least filter out similar keywords among sub-TDs by using a word embedding technique. Third, this paper clusters MOCs by using agglomerative clustering algorithm. Even though this clustering seems to be effective to show all hierarchical structure within the TD, it would be ineffective if user wants to identify only the first-level sub-TDs without setting specific cut-off value. We think a community detection algorithm can be used to automatically find the optimized number of sub-TDs. Fourth, this research conducted only one empirical analysis to test the method. A number of case studies are always a good strategy to assess universality of the method. Therefore, we will provide more empirical case studies in our further research. Specifically, decomposition of whole technologies, all patents, by applying the proposed method to find all existing TDs in the patent system. Lastly, the COM can be a more powerful tool by using CPC instead of using IPC. CPC is a more specific and detailed version of IPC; CPC has almost 260,000 sub-classes. Therefore, the use of CPC enables MOCs by combinations of CPC and UPC to represent more specific technological knowledge and then the identified clusters to have higher independencies.

Acknowledgements This research was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Science, ICT & Future Planning (No. 2017R1A2B4012431).

References

Antonelli, C. (2011). *Handbook on the economic complexity of technological change*. Cheltenham: Edward Elgar Publishing.

Basnet, S., & Magee, C. L. (2017). Artifact interactions retard technological improvement: An empirical study. *PLoS ONE, 12*(8), e0179596.

Benson, C. L. (2014). *Cross-domain comparison of quantitative technology improvement using patent derived characteristics*. Cambridge: Massachusetts Institute of Technology.

Benson, C. L., & Magee, C. L. (2013). A hybrid keyword and patent class methodology for selecting relevant sets of patents for a technological field. *Scientometrics, 96*(1), 69–82.

Benson, C. L., & Magee, C. L. (2015a). Quantitative determination of technological improvement from patent data. *PLoS ONE, 10*(4), e0121635.

Benson, C. L., & Magee, C. L. (2015b). Technology structural implications from the extension of a patent search method. *Scientometrics, 102*(3), 1965–1985.

Benson, C. L., & Magee, C. L. (2016). Using enhanced patent data for future-oriented. In T. U. Daim, D. Chiavetta, A. L. Porter, & O. Saritas (Eds.), *Anticipating future innovation pathways through large data analysis* (pp. 119–131). Berlin: Springer.

Carley, S. F., Newman, N. C., Porter, A. L., & Garner, J. G. (2018). An indicator of technical emergence. *Scientometrics, 115*(1), 35–49.

Choi, S., Noh, M. S., Yoon, J., Park, H., & Seo, W. (2018). Analyzing technological spillover effects between technology classes: The case of Korea Technology Finance Corporation. *IEEE Access, 6,* 3573–3584.

EPRI. (2009). *Solar Photovoltaics: Status, Costs, and Trends* (Vol. 1015804). Palo Alto, CA: EPRI.

Evenson, R., & Puttnam, J. (1988). The Yale-Canada patent flow concordance. D, Yale University.

Fleming, L. (2001). Recombinant uncertainty in technological search. *Management Science, 47*(1), 117–132.

Fleming, L., & Sorenson, O. (2001). Technology as a complex adaptive system: Evidence from patent data. *Research Policy, 30*(7), 1019–1039.

Fu, K., Murphy, J., Yang, M., Otto, K., Jensen, D., & Wood, K. (2015). Design-by-analogy: Experimental evaluation of a functional analogy search methodology for concept generation improvement. *Research in Engineering Design, 26*(1), 77–95.

Guan, J., & Liu, N. (2016). Exploitative and exploratory innovations in knowledge network and collaboration network: A patent analysis in the technological field of nano-energy. *Research Policy, 45*(1), 97–112.

Jaffe, A. B., & Trajtenberg, M. (2002). *Patents, citations, and innovations: A window on the knowledge economy*. Cambridge: MIT press.

Jaffe, A. B., Trajtenberg, M., & Henderson, R. (1993). Geographic localization of knowledge spillovers as evidenced by patent citations. *The Quarterly Journal of Economics, 108*(3), 577–598.

Lybbert, T. J., & Zolas, N. J. (2014). Getting patents and economic data to speak to each other: An 'algorithmic links with probabilities' approach for joint analyses of patenting and economic activity. *Research Policy, 43*(3), 530–542.

Magee, C., Basnet, S., Funk, J., & Benson, C. (2016). Quantitative empirical trends in technical performance. *Technological Forecasting and Social Change, 104,* 237–246.

Martinelli, A. (2012). An emerging paradigm or just another trajectory? Understanding the nature of technological changes using engineering heuristics in the telecommunications switching industry. *Research Policy, 41*(2), 414–429.

Mina, A., Ramlogan, R., Tampubolon, G., & Metcalfe, J. S. (2007). Mapping evolutionary trajectories: Applications to the growth and transformation of medical knowledge. *Research Policy, 36*(5), 789–806.

Moehrle, M. G., Walter, L., Geritz, A., & Müller, S. (2005). Patent-based inventor profiles as a basis for human resource decisions in research and development. *R&d Management, 35*(5), 513–524.

Moeller, A., & Moehrle, M. G. (2015). Completing keyword patent search with semantic patent search: Introducing a semiautomatic iterative method for patent near search based on semantic similarities. *Scientometrics, 102*(1), 77–96.

Murphy, J., Fu, K., Otto, K., Yang, M., Jensen, D., & Wood, K. (2014). Function based design-by-analogy: A functional vector approach to analogical search. *Journal of Mechanical Design, 136*(10), 101102.

Park, H., & Magee, C. L. (2017). Tracing technological development trajectories: A genetic knowledge persistence-based main path approach. *PLoS ONE, 12*(1), e0170895.

Park, H., Ree, J. J., & Kim, K. (2013a). Identification of promising patents for technology transfers using TRIZ evolution trends. *Expert Systems with Applications, 40*(2), 736–743.

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

Rose, S., Engel, D., Cramer, N., & Cowley, W. (2010). *Automatic keyword extraction from individual documents* (pp. 1–20). Text Mining: Applications and Theory.

Salton, G., Wong, A., & Yang, C.-S. (1975). A vector space model for automatic indexing. *Communications of the ACM, 18*(11), 613–620.

Schmoch, U., Laville, F., Patel, P., & Frietsch, R. (2003). Linking technology areas to industrial sectors. *Final Report to the European Commission, DG Research, 1,* 100.

Strumsky, D., & Lobo, J. (2015). Identifying the sources of technological novelty in the process of invention. *Research Policy, 44*(8), 1445–1461.

Trajtenberg, M. (1987). *Patents, citations and innovations: Tracing the links*. Cambridge, MA: National Bureau of Economic Research Cambridge.

Verhoeven, D., Bakker, J., & Veugelers, R. (2016). Measuring technological novelty with patent-based indicators. *Research Policy, 45*(3), 707–723.

Verspagen, B. (2007). Mapping technological trajectories as patent citation networks: A study on the history of fuel cell research. *Advances in Complex Systems, 10*(01), 93–115.

Verspagen, B., Morgastel, T. V., & Slabbers, M. (1994). MERIT concordance table: IPC-ISIC (rev. 2). Maastricht: MERIT Research Memorandum 2/94-004.

Von Wartburg, I., Teichert, T., & Rost, K. (2005). Inventive progress measured by multi-stage patent citation analysis. *Research Policy, 34*(10), 1591–1607.

Walter, L., Radauer, A., & Moehrle, M. G. (2017). The beauty of brimstone butterfly: Novelty of patents identified by near environment analysis based on text mining. *Scientometrics, 111*(1), 103–115.

Wang, X., Ma, P., Huang, Y., Guo, J., Zhu, D., Porter, A. L., et al. (2017). Combining SAO semantic analysis and morphology analysis to identify technology opportunities. *Scientometrics, 111*(1), 3–24.

Weitzman, M. L. (1998). Recombinant growth. *The Quarterly Journal of Economics, 113*(2), 331–360.

Wu, Z., & Palmer, M. (1994). Verbs semantics and lexical selection. In *Proceedings of the 32nd annual meeting on Association for Computational Linguistics* (pp. 133–138). Association for Computational Linguistics.

Yan, Y., & Guan, J. (2018). How multiple networks help in creating knowledge: Evidence from alternative energy patents. *Scientometrics, 115*(1), 1–27.

Yoon, J., & Kim, K. (2011). Identifying rapidly evolving technological trends for R&D planning using SAO-based semantic patent networks. *Scientometrics, 88*(1), 213–228.

Youn, H., Strumsky, D., Bettencourt, L. M., & Lobo, J. (2015). Invention as a combinatorial process: Evidence from US patents. *Journal of the Royal Society, Interface, 12*(106), 20150272.