Journal of Informetrics 10 (2016) 967–980

Contents lists available at ScienceDirect

Journal of Informetrics

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

Unconnected component inclusion technique for patent network analysis: Case study of Internet of Things-related technologies

Yasutomo Takano*, Cristian Mejia, Yuya Kajikawa

Graduate School of Innovation Management, Tokyo Institute of Technology, 3-3-6 Shibaura, Minato-ku, Tokyo 108-0023, Japan

ARTICLE INFO

Article history:
Received 21 January 2016
Received in revised form 8 May 2016
Accepted 24 May 2016
Available online 13 September 2016

Keywords:
Patent network analysis
Patent family analysis
Unconnected component
Clustering
Internet of Things

ABSTRACT

In this study, we propose an unconnected component inclusion technique (UCIT) for patent citation analysis. Our method generates a cluster solution that includes unconnected and connected components of a direct citation network, enabling a more complete analysis of the technology fields. Case studies of Internet of Things-related technologies were conducted to test the effectiveness of our proposed method. We observed that UCIT increased the number of nodes especially in relatively small networks. Additionally, we analyzed how the clusters changed by adding unconnected patents to the citation network and identified four types of clustering phenomenon. Our method can be used by patent officers, R&D managers, and policy makers when they want to understand the technology landscape better.

1. Introduction

To understand technological trends and model innovation processes, researchers are increasingly using patent data because they are easily available and full of valuable information. By analyzing patent data, we can measure research and development activity, technological progress, and inventor productivity. Analytical methods include text mining (Hu, Fang, & Liang, 2014; Wang, Liu, Ding, Liu, & Xu, 2014), analysis of international patent classification (IPC) frequency and co-occurrence (Leydesdorff, Kushnir, & Rafols, 2014), and citation-based analysis (Karki & Krishnan, 1998; Verspagen, 2007).

Among them, patent citations have been used as indicators of forecasting technology success (Altuntas, Dereli, & Kusiak, 2015), emerging technologies (Daim, Rueda, Martin, & Gerdsri, 2006), and technology selection (Marra, Emrouznejad, Ho, & Edwards, 2015). Citations have also demonstrated that they improve the accuracy of patent retrieval techniques (Lopez & Romary, 2010; Mahdabi & Crestani, 2014). Because citations naturally point to previous patents, mapping citation network is useful in creating a landscape of a particular technology, in which the flow of innovation can be observed and technology fronts can be detected (Ogawa & Kajikawa, 2014).

Citation network analysis originated in the work of De Solla Price (1965), in which a citation network of papers was studied. Since then, network analysis on papers has flourished and the methodology has been translated to the patent world. This is justified because the bibliometric properties of papers and patents have been shown to be similar (Narin, 1994).

* Corresponding author.
E-mail address: takano.y.ac@m.titech.ac.jp (Y. Takano).

http://dx.doi.org/10.1016/j.joi.2016.05.004

When a citation network is created for the purpose of studying technology landscapes, a group of patents related to a selected technology is retrieved and then the citation connections within that group are established. Some patents might be connected to others, forming subgroups, whereas others may remain without any connection. Special attention is given to the largest subgroups, namely, the largest or maximum component, whereas the remaining subgroups are neglected. Networks represented by the largest component, especially of patent citation networks, might have a disadvantage in that they could be considered not large enough to represent a technology field adequately.

In contrast to networks of academic papers, patent networks may lose a considerable amount of nodes by neglecting the unconnected components (Shibata, Kajikawa, & Sakata, 2010). Our objective is to reduce the amount of information lost in patent network analysis. To achieve this, we applied an expansion strategy at the dataset level to establish connections between the largest connected component and the relevant unconnected components. Thus, we tackle the situation when a patent citation network is regarded as small compared to the size of the dataset. Our proposed methodology is called unconnected component inclusion technique (UCIT), which supports the inclusion of relevant, but originally unconnected, components in citation network analysis.

To illustrate and evaluate the advantages of our proposed method, we applied UCIT to the citation networks of the Internet of Things (IoT) and its three main hardware components. IoT is a concept related to the "pervasive presence around us of a variety of things or objects which, through unique addressing schemes, are able to interact with each other and cooperate with their neighbors to reach common goals" (Atzori, Iera, & Morabito, 2010). The hardware of IoT consists of prior technologies, specifically, radio-frequency identification (RFID), near field communication (NFC), and sensor networks (SN) (Whitmore, Agarwal, & Da Xu, 2014). These four networks present different characteristics in size and age, which is useful for a comprehensive assessment of the method.

The rest of the paper is as follows: Section 2 discusses the methods for network creation and the concept of patent family as a unit of analysis; here, we discuss the preference of citation networks over other methodologies. Section 3 explains the UCIT for increasing the coverage of the network, and Section 4 discusses the corresponding results. Practical implications are described in Section 5, where our four case studies are analyzed in depth, and limitations of the study are discussed in Section 6. Finally, in Section 7, we present our conclusion, showing the benefits and future directions in expansion strategies.

2. Related works

In this section, we review papers that analyze the different forms of clustering or document classification through networks. Some of them focus on networks of academic articles, but their methodologies apply to patent datasets as well, given the similarities they share. As the network is the basis for the clustering technique, we want to focus primarily on its construction methods. We also discuss the usefulness of the patent family as the unit of analysis.

2.1. Network construction

Basically, networks are created from a collection of documents retrieved from a database; the quality of the database and the information retrieval strategy may have an impact on the size and connectivity of the network. Thus, a noisy dataset (i.e., documents with little relatedness between them) might result in sparse networks that do not capture the full characteristics of the field in analysis. This concern is addressed by the information retrieval research, which looks for methods to increase the relatedness of documents retrieved. Specifically, research studies on patent retrieval have focused on the improvement of the algorithms in the search engines (Bache, 2011) or on strategies for query refinement (Mahdabi & Crestani, 2014). Therefore, the starting point for patent network analysis is the retrieval of the adequate dataset. Once the corpus of documents is obtained, several strategies can be used to establish the links between the documents. These construction methods are mainly based on IPC codes, text contents, and citations.

IPC codes are the standard for patent classification. In network analysis, IPC codes have been used to create networks by linking two or more patents whenever they share an IPC code. Leydesdorff (2008) used this strategy to analyze a dataset of 138,751 patents, mapping networks at different levels of the IPC code for the whole dataset and for country subsets. The results, however, were not positive. Networks based on IPC made a poor representation of the technology space, and the use of other methods such as text-based networks was suggested.

When it comes to text-based methods for network construction, researchers have mainly focused on two trends, one relying on text similarity measures and the other applying probabilistic models. In a large-scale experiment, Boyack et al. (2011) evaluated the accuracy of nine text similarity measures for clustering 2.15 million biomedical articles. The PMRA (PubMed Related Articles) measure proved to be the best, followed by BM25 and a topic model-based measure. PMRA is a similarity measure specific to the PubMed interface; thus, its application in other fields is limited. The other two are common methods in text mining literature. Even though the aforementioned experiment was comprehensive, there was no general agreement. For instance, Hamedani and Kim (2014) used a dataset of more than 1 million computer science articles to compare several similarity measures; their experiment showed that there was no significant difference between BM25 and the classic cosine similarity of tf-idf weighting. In another experiment, document classification using the tf-idf yielded better performance when the unit of comparison was phrases instead of single terms (Zhang et al., 2016).

Other text-based methods rely on probabilistic models that can be estimated from text corpus. Yau, Porter, Newman, and Suominen (2014) explored the possibilities of classifying academic articles from seven research domains using topic models. Four topic model algorithms were compared, and the hierarchical Dirichlet process scored the highest performance. Gretarsson, O Donovan, Asuncion, and Newman (2011) also exploited topic models for network creation, but their focus was more on improving the analysis of the document through visualization than on performance evaluation or clustering.

The performance of text-based methods depends on the selection of text fields, using words or phrases; the cleaning strategy of the text corpus; and the tuning of the parameters in the algorithms. Therefore, there is still no significant evidence that will make researchers choose one method over the other. On the other hand, citation-based methods for network analysis and clustering have reached consensus.

Three types of citation networks can be constructed. The simplest one is direct citation, which refers to the establishment of the linkage from a document to all the cited references in it; this is repeated for each document in the dataset, resulting in a network of cited references (De Solla Price, 1965). The other two methods are bibliographic coupling and co-citation. Each possible pair of documents in a dataset is compared and bibliographically coupled when they share one or more cited references. The number of references they share is the strength of this connection (Kessler, 1963). Finally, co-citation connects all references cited within a document; the more two documents are cited together (co-occur) in other documents, the stronger their relation gets (Small, 1973). Details on the construction process and characteristics of these methods can be found in Shibata, Kajikawa, Takeda, and Matsushima (2009). Once a network is created by any of the three methods, tightly connected groups of documents can be identified and clustered.

As stated by Waltman and Van Eck (2012), intuitively, direct citation networks may convey the strongest relationship between the documents compared to co-citation and bibliographic coupling because they need a third intermediary document to establish the linkage, being indirect methods for network creation. To know which method represents research domains more accurately, previous research studies have compared the performance of these networks under different conditions. Shibata et al. (2009) compared the three methods when analyzing research domains of gallium nitride, complex networks, and carbon nanotubes. In all cases, direct citation performed best based on different evaluations such as size of clusters, speed, and topological relevance. Clusters observed from the direct citation network suggested a high content similarity. Klavans and Boyack (2015), who studied rather large datasets, recently conducted the most comprehensive citation network analysis to date by using a corpus of 48.5 million documents, which basically included all indexed articles in Scopus from 1996 to 2012 and other sources, and obtained the same results as those of Shibata et al. Although most of their works have favored co-citation and bibliographic coupling in the past, they now motivate future studies on citation networks to focus on direct citation.

Hybrid methods combining citations and text have also been proposed in the literature, but they inherit the same issues as those of text-only methods, and their effectiveness depends on the selection of parameters, etc. Therefore, we decided to build our methodology based on direct citation networks, given the straightforward way of construction and the demonstrated superior performance for large and domain-specific networks.

2.2. Patent family

Citation networks of patents can be constructed by using any of two units of analysis: the patent or the patent family. A patent family is defined as a collection of patents that are filed in different countries but refer to the same invention (Thomson Innovation, 2016). Depending on the country, the citation behavior changes. The United States Patent and Trademark Office asks patentees to be as comprehensive as possible in citing all state-of-the-art technologies related to the invention, whereas the European Patent Office and the Japan Patent Office focus on those citations strictly related to the invention (Michel & Bettels, 2001). Moreover, some patentees may strategically choose to cite, or not, patents from competitors, or they just failed to find similar patents by the date they filed the patent in one office, but found new similar technologies when filing in other countries (Wilson, 1995). Then, the use of a single patent authority database for patent networks may diminish the analysis owing to the incompleteness of the citations. To overcome this issue, we can use patent family networks.

To create patent family networks, we need to merge the members of the family (i.e., the same patent in different countries) and aggregate their citations. By combining the citation information, we can increase the connectivity of the network (Nakamura, Suzuki, & Kajikawa, 2015). Despite this benefit, the use of patent family networks has been scarce in the literature. Some of the few examples are Fajardo-Ortiz, Ortega-Sánchez-de-Tagle, and Castaño (2015) and Hu and Liu (2015); the former analyzed patent families of technologies related to Ebola research, and the latter mapped patent families of microelectromechanical systems (MEMS).

However, citation networks of patents may show an undesired characteristic regardless of whether they were built using patents from a single authority or from patent families. They are usually sparse. For instance, the aforementioned networks connected only 20%[1] of the 102 patent families of Ebola's research technologies and 41% of the 27,152 patent families of MEMS technology. Other papers have explicitly reported this as an issue. When comparing the gap between academic research and patents of solar cell technologies, Shibata et al. (2010) observed that, even though an increasing number of patents were filed through the years, the largest component of the patent network did not grow and remained small, connecting only 30% of the 333,207 patents retrieved. Nakamura et al. (2015) noted this problem and measured how the patent family network improves the connectivity over patents from a single authority. They compared the networks of

[1] Percentages shown in this section were calculated based on network and dataset sizes reported in those papers.

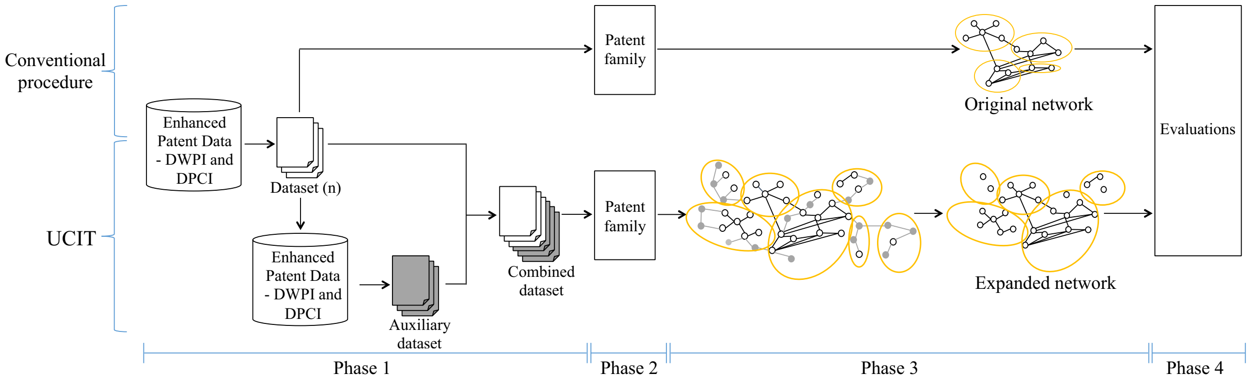


Fig. 1. Overview of our research.

five drive train technologies, validating that, by merging patent families, the connectivity increased. Even so, the maximum improvement was observed by connecting 51% of patent families. When the network is not large enough, we cannot regard it as a comprehensive representation of a technology field.

We do not claim that small networks lead to wrong conclusions. The frequency of citations in patents has a relation to their value, and it might be possible to weight each patent by its number of citations (Hall, Jaffe, & Trajtenberg, 2005; Harhoff, Narin, Scherer, & Vopel, 1999). Thus, even though citation networks are sparse, they still capture mainstream trends of the technological field. Nonetheless, they might fail to represent all its features. Therefore, our methodology aims to increase the participation of unconnected components when the lack of connectivity is a concern. Instead of clustering only the network, our methodology generates clusters that include both the network and the relevant unconnected components. This brings the possibility of a more complete analysis of the technologies and is achieved by direct citation alone.

3. Methodology

An overview of our research is given in Fig. 1. In this paper, we use the word "conventional" to refer to direct citation networks without any further processing or expansion strategy, which is the normal way of creating patent citation networks in the literature. To compare the performance of UCIT to that of the conventional procedure in a patent network analysis, we need to complete four phases. In Phase 1, datasets of patents are obtained. In Phase 2, datasets are processed to include patent family information, which enables networks of patent families to be generated instead of networks of single patents. In Phase 3, the networks are created and clustered. The following two clustered networks from the same dataset are compared:

- the *original network*, which is the maximum component of the network formed by the conventional procedure, and
- the *expanded network*, which is the same original network plus the set of unconnected nodes that share citation relationships outside the dataset.

In the final phase, both the original and the expanded networks are compared and evaluated.

3.1. Conventional and proposed procedure

To collect patent data for the four cases, we used the Derwent World Patent Index (DWPI) database by Thomson Innovation. We chose this database because it is known as a comprehensive source of patent information. It contains patent information from more than 80 patent authorities worldwide, collecting the family members of each patent. It also includes the Derwent Patent Citation Index (DPCI), which contains citation references for each patent family. This information is needed to construct networks of patent families. Moreover, this database provides concise abstracts written in English by the DWPI editorial team for patents issued in more than 30 foreign languages, which are useful in evaluating our results by means of text analysis.

The conventional procedure for building a patent citation network is shown in Fig. 2. Step (1) involves the collection of a patent dataset. According to the analysis targets, appropriate query is used and bibliographic records of patents are retrieved. In step (2), the dataset is processed to include patent family information. All citations of the family are aggregated into a single patent representative of that family that will act as a node in the network. In addition, the aggregated citations are the edges. In Step (3), the direct citation network is created. In Step (4), the directions of the edges are removed and only the largest connected component (the maximum component) is selected for analysis to focus on those patents directly connected. Finally, in Step (5), a clustering algorithm is applied to the largest connected component. With this procedure, edges can be established only among patents in the dataset. It is worth noting that Step (2), the aggregation of patent families,

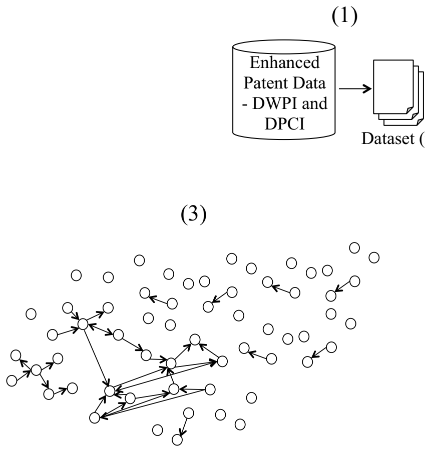


Fig. 2. Conventional procedure for creating a patent citation network.

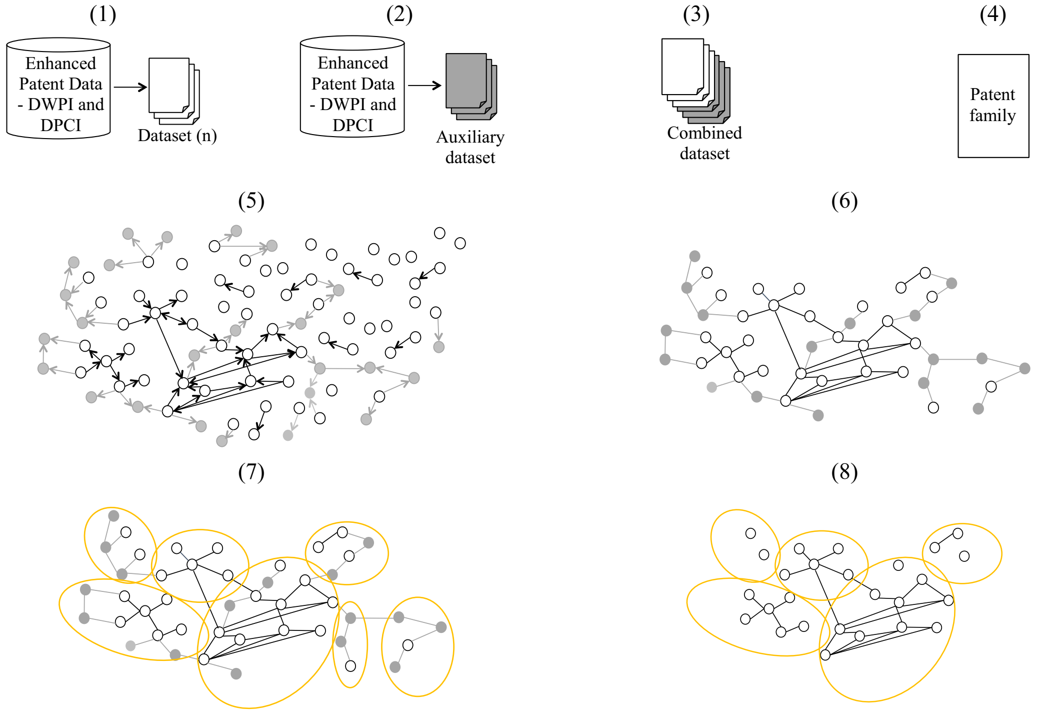


Fig. 3. Procedure of UCIT for creating patent citation networks.

is optional, but as discussed in Section 2 this step has the benefit of increasing the connectivity of the network and we want to compare UCIT to the most comprehensive scenario.

The procedure of UCIT is shown in Fig. 3. The assumption of this technique is that two unconnected patents may still share a citation relationship with other patents outside the dataset. Then, an auxiliary patent serves as a bridge to connect them. As in the conventional procedure, in Step (1) of the UCIT, a dataset of patents is obtained. We call this dataset (n). In Step (2), an auxiliary dataset of patents is obtained based on any of the following citation information:

(a) Cited references (backward citations) from all patents in (n)
(b) Citing references (forward citations) to all patents in (n).

The citation information of these auxiliary datasets is used to establish the connections between the maximum and the unconnected components of (n). Furthermore, both (a) and (b) can be used in combination or in a recursive fashion (e.g., the cited of the cited). In Step (3), the original dataset (n) is combined with the selected auxiliary datasets. In Step (4), the data

Table 1
Auxiliary datasets.

Label	Description
$(n-1)$	Set of cited patents from all patents in (n)
(n+1)	Set of citing patents to all patents in (n)
$(n-2)$	Set of cited patents from all patents in $(n-1)$

Table 2
Combined datasets.

Method	Cited references		Citing references
Conventional procedure		(n)	
UCIT	$(n-1)+$	(n)	
UCIT		(n)	$+(n+1)$
UCIT	$(n-2)+(n-1)+$	(n)	
UCIT	$(n-1)+$	(n)	$+(n+1)$

Table 3
Network totals derived from the conventional procedure.

	Average year	Size of dataset (n)	Size of maximum component	Coverage
RFID	2007.7	49,910	25,116	50.3%
NFC	2010.9	9908	2800	28.3%
IoT	2010.8	5435	76	1.4%
SN	2007.7	607	17	2.8%

are processed to include patent family information. The processed file is used to create a network in Step (5). The network created by this procedure has the following characteristics:

- The direction of new edges can be observed

1) from patents of dataset (n) to those of auxiliary datasets
2) from patents of an auxiliary dataset to those of dataset (n)
3) from patents of an auxiliary dataset to other patents of an auxiliary dataset

- Some patents of dataset (n) remain disconnected because they do not have cited references and have yet to receive citations.

In Step (6), the direction of edges is removed and the maximum component is obtained. This maximum component includes patents from the original dataset (n) and from auxiliary datasets. Clusters are obtained in Step (7). Because the maximum component has changed, the number of clusters is expected to be different from those in the original network. Finally, in Step (8), patents from the auxiliary dataset, including their edges, are removed. This is necessary to retain only the patents in the original dataset (n), which are the subject of interest. Here, connected and unconnected components are present. Clusters containing a single patent are also removed.

3.2. Datasets and their basic characteristics

In our study, UCIT was applied to four cases: patents of IoT and its three hardware components, as described by Whitmore et al. (2014). The following queries were used to retrieve and collect the bibliographic data: “radio frequency identification” OR “RFID” for RFID, “near field communication” OR “NFC” for NFC, “Internet of Things” OR “IoT” for IoT, and “sensor networks” for sensor network. The dataset (n) for the four cases contains 49,910, 9908, 5435, and 607 patents, respectively. The dataset covered patents until March 20, 2013, and were retrieved on August 20, 2015, using the search condition “enhanced patent data–DWPI and DPCI.” After the data retrieval, we collected the lists of cited and citing patents not present in the dataset and obtained auxiliary datasets based on that information.

To determine whether cited or citing references establish better connections to unconnected components, we combined different auxiliary datasets with the original dataset (n). Tables 1 and 2 show the composition of those datasets. For each technology, five networks were created: one based on the conventional procedure and four that employ UCIT with the combined datasets, as shown in Table 2. A topological clustering algorithm based on modularity maximization (Clauset, Newman, & Moore, 2004) was used to cluster all networks.

Table 3 provides the basic information about the network created using the conventional procedure. The coverage refers to the percentage of patents in the dataset used to form the citation network. IoT and SN can be regarded as sparse and highly unconnected, whereas RFID and NFC coverage can be improved.

To determine the combination of auxiliary datasets that most improves the coverage ratio, we conducted a cost-performance analysis. If we expand the dataset, it is clear that the coverage improves. However, it takes time to collect

Table 4
Cost according to the combined datasets.

	(n)	$(n)+(n+1)$	$(n-1)+(n)$	$(n-1)+(n)+(n+1)$	$(n-2)+(n-1)+(n)$
RFID	49,910	176,904	257,036	384,030	1,677,339
NFC	9908	24,147	49,459	63,698	405,124
IoT	5435	6845	8859	10,269	41,213
SN	607	4081	5576	9050	87,837

Table 5
Performance according to the combined datasets.

	(n)	$(n)+(n+1)$	$(n-1)+(n)$	$(n-1)+(n)+(n+1)$	$(n-2)+(n-1)+(n)$
RFID	25,116	30,255	33,903	34,383	–
NFC	2800	3628	4806	5143	6431
IoT	76	154	537	550	785
SN	17	229	316	372	417

Table 6
Cost performance according to the combined datasets.

	(n)	$(n)+(n+1)$	$(n-1)+(n)$	$(n-1)+(n)+(n+1)$	$(n-2)+(n-1)+(n)$
RFID	50.3%	17.1%	13.2%	9.0%	–
NFC	28.3%	15.0%	9.7%	8.1%	1.6%
IoT	1.4%	2.3%	6.1%	5.4%	1.9%
SN	2.8%	5.6%	5.7%	4.1%	0.5%

auxiliary datasets and to calculate the expanded network. On the basis of the cost performance, we intended to determine the tradeoff between extent of the expansion and performance improvement. Table 4 lists the sizes of each combined dataset. The size, i.e., the number of patents, is regarded as the cost. We observed that, by obtaining auxiliary datasets in a recursive manner, the size increased exponentially. In particular, the RFID dataset growing rate was too large for $(n-2)$ to be included in this analysis.

Table 5 shows the performance, which we defined as the increment in coverage. This resulted in the inclusion of unconnected components of (n). IoT and SN showed a sharp increment when any combined datasets were used. NFC and RFID also increased, but not as dramatically. This suggests that UCIT works best for networks that are highly sparse and unconnected. In addition, the greater the number of auxiliary datasets combined, the greater the number of unconnected patents included with the maximum component of (n). However, these increments are associated with their respective cost. Table 6 shows the ratio of cost to performance. The best ratio was obtained when using auxiliary dataset $(n-1)$ for the IoT and SN. For NFC and RFID, the ratio decreased in all cases, meaning that, although the auxiliary datasets were large, they were able to generate only a small increment in performance. Such behavior was expected because NFC and RFID were better connected than the other datasets and the size of their maximum component was already sufficiently large. Therefore, to maintain the same conditions in the evaluation phase, we extended the use of auxiliary dataset $(n-1)$ to RFID and NFC even though the cost performance was not high. Nevertheless, an increase in performance was also observed.

We assessed not only the increase in coverage, but also the quality and effect of adding the unconnected components. We demonstrated that UCIT not only added more patents, but also improved the quality of the textual coherence in the cluster. We evaluated the overall performance of UCIT by comparing the *semantic coherence* of the text contents in the clusters and by comparing the IPC distributions of the patents that were considered reachable against those that were not.

3.3. *Methods of evaluation*

By applying the aforementioned steps, we increased the number of patents to be considered in a patent network analysis. The best option for expanding the original dataset (n) is to combine it with dataset $(n-1)$, which is the set of backward citations from all patents in (n). To evaluate our claim that using unconnected components in the analysis improves not only the coverage but also the semantic coherence of tightly knit groups of patents in the citation networks, we compared the composition of clusters created using the conventional procedure to those created by UCIT. One measure commonly used to evaluate the quality of clusters in networks is *modularity*, which measures the appearance of densely connected groups of nodes (Newman 2006). However, our technique creates clusters that contain both unconnected and connected nodes. This is because after connecting the nodes with the expanded corpus and clustering the network we eliminated the patents from the auxiliary dataset (see Step (8) in Fig. 3). Thus, using modularity was not an option for evaluation in this study.

Another method of evaluation includes precision and recall, which requires a ground truth or reference data inexistent for our datasets. We could have created such reference by taking into account human-based partitions from expert opinions, which is time consuming and can be regarded as subjective. Instead, we exploited the natural characteristics offered by the network; we conducted two types of evaluations. *Evaluation* 1 measures intra-cluster text similarity and the similar-

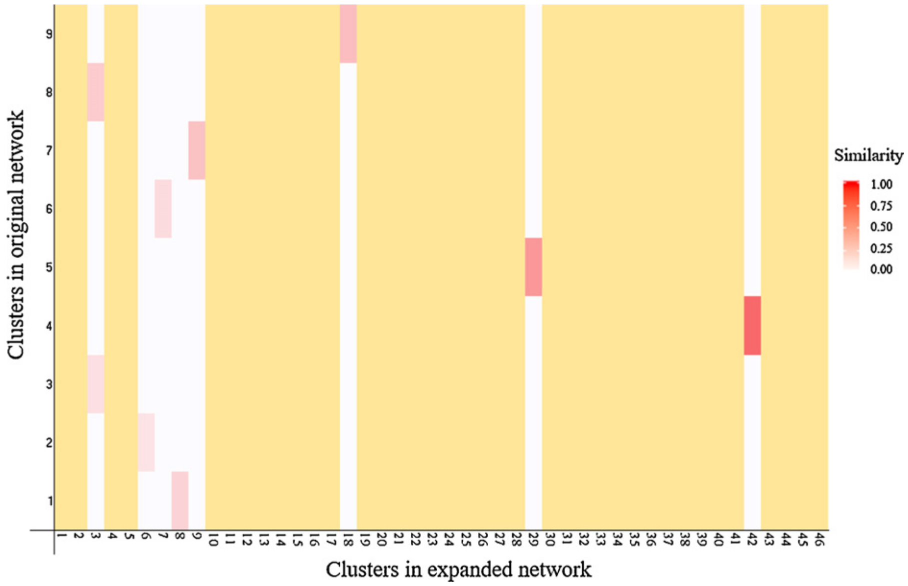


Fig. 4. Heatmap representing the similarity between clusters in both networks for IoT. (For interpretation of the references to colour in the text, the reader is referred to the web version of this article.)

ity between clusters in both networks. *Evaluation* 2 describes the reachability of unconnected patents and compares IPC subclasses.

3.3.1. Evaluation 1. Comparison of clusters in the original and expanded networks

The first text similarity measure was performed on the contents of the clusters. UCIT performs well if it is able to allocate patents in clusters, which intra-cluster text similarity improves in relation to the original network. To measure this, we focused on *core patents*.

A core patent is a patent having the highest degree centrality in each cluster of the original dataset (n). Degree centrality is the sum of the edges from and to that patent. When multiple patents have the same highest degree centrality inside a cluster, all are treated as core patents. Once those patents are identified, we locate them in the network created by UCIT and compare the semantic coherence of the cluster in the original network with the semantic coherence of the cluster in the expanded network.

Semantic coherence is defined as the average of the cosine similarity of all pairs of patents in a given cluster as represented in the following equation:

$$\mathrm{CosineSimilarity}(t, s) = \frac{\overline{j_t} \cdot \overline{j_s}}{\sqrt{\sum_i j_t^{(i)} \cdot j_s^{(i)}}} \tag{1}$$

where J_t and J_s are the text vectors of patent t and patent s, respectively. Cosine similarity measures the similarity between two vectors. To calculate it, we merge the fields of “Title—DWPI” and “Abstract.” The text is then processed by converting it to lowercase; removing stop words, symbols, and punctuation marks; and stemming each word. Finally, words are treated as vectors and tf-idf is calculated. We use the cosine similarity of tf-idf because it is commonly used and has been proven as an effective way to find the relationships in a corpus of text to detect technological fronts (Shibata, Kajikawa, & Sakata, 2011a).

The second text similarity measurement was performed across networks. To measure the information added by the unconnected components, we applied the cosine similarity again. However, this time, it was used to compare the clusters in the original dataset (n) to all clusters generated by UCIT. This inter-cluster similarity can be represented as a heatmap in which the intersection is colored according to the similarity value. If the cosine similarity of two clusters is greater than a certain threshold, this suggests that both clusters contain similar information. An example of the result is shown in Fig. 4 for IoT. Similar clusters greater than a threshold empirically set at 50% are marked in red. When a cluster generated by UCIT does not share any similarity with any of the original clusters, then it contains new information, which is indicated by the yellow columns in Fig. 4. With the heatmap, we evaluated how UCIT could extract new clusters that did not appear in the original dataset.

Table 7
Number of core patents inside clusters in the expanded network that showed higher semantic coherence.

	RFID	NFC	IoT	SN	Average
Higher	575	49	9	6	159.8
All	709	74	17	6	201.5
Higher/all (%)	81.1%	66.2%	52.9%	100.0%	75.1%

Table 8
Number of clusters in the expanded network with new information.

	RFID	NFC	IoT	SN	Average
New	214	124	38	26	100.5
All	414	170	46	28	164.5
New/all (%)	51.7%	72.9%	82.6%	92.9%	75.0%

3.3.2. *Evaluation 2. Characteristic analysis of reachable and unreachable components*

Even when we expand the dataset by UCIT, some patents in the original dataset cannot be connected. The reason for the existence of such unconnected patents is that they do not have citing and cited patents. We call these as *unreachable* patents. By contrast, the unconnected patents in the original dataset that have at least one cited or citing reference are *reachable*. We then define reachability as the maximum coverage possible for a given dataset. This simply indicates the percentage of patents that have at least one citation, and they are expected to connect at some point when the appropriate auxiliary datasets are added. Then, reachability can serve to assess the performance of different citation networks, and, based on the coverage, those close to the reachability are more inclusive networks.

To determine the difference between reachable and unreachable patents, we analyzed their IPC subclasses. Patents are classified using the IPC codes, which could be up to 12 digits long. Each patent has one or more IPC codes, and they can be used to evaluate the similarity between patents or groups of patents. Patents share similar characteristics of the invention when they have the same IPC code. In this study, we compared the IPC subclasses, the first four digits of the code, within reachable and unreachable patents. We also calculated and compared the reachable patents to the maximum component of the original network to verify whether UCIT alters the frequency with which IPC codes appear. The four-digit subclasses are the third hierarchical level of the code, being the level that precisely describes and divides the technologies covered by the classification scheme of the IPC (WIPO, 2015). The remaining digits can only be interpreted under the scope of the subclass. Moreover, the four-digit code is useful for the analysis of specific technologies (Leydesdorff et al., 2014).

To assess the presence of different IPC subclasses in each set (reachable, unreachable, and maximum component of the original network), we used Jaccard similarity. It is simply the ratio of the intersection over the union of IPC subclasses for a pair of sets and is not sensitive to frequency. On the other hand, we used cosine similarity, which is sensitive to weights, to determine whether highly frequent IPC codes in each set are similar despite containing other infrequent codes. Both similarity measures indicate a high similarity when they get close to 1.

4. Results

4.1. *Comparison of clusters in the original and expanded networks*

The original network was obtained by the conventional method, as shown in Fig. 2, whereas the expanded network was obtained by UCIT, as shown in Fig. 3. They were compared in our study. In Evaluation 1, we performed two analyses to evaluate the usefulness of UCIT.

First, the cosine similarity of intra-clusters that shared core patents in the original and expanded networks was calculated and compared. We refer to this comparison as semantic coherence. Core patents were tracked between the original and the expanded networks. We counted the number of core patents inside clusters in the expanded network that showed higher semantic coherence than those in the original network. The results are summarized in Table 7. On average, 75.1% had more relevant patents inside the cluster when applying UCIT. This means that UCIT effectively allocates core patents into more semantically coherent clusters.

Second, the cosine similarity between clusters in the original and the expanded networks was measured to investigate clusters containing new information. Here, a cluster with new information means that a cluster in the expanded network shares no significant similarity with any clusters in the original network. We counted the number of clusters in the expanded network that had new information, which was 75.0% on average, as shown in Table 8. These results indicate that, with UCIT, we can obtain new information from the expanded network that was not included in the clusters in the original network.

4.2. *Characteristic analysis of reachable and unreachable components*

The reachability for each case study was calculated in Evaluation 2 to determine the structure of the entire patent networks. The results are summarized in Table 9, where we can observe that UCIT got a higher coverage than the conventional

Table 9
Performance comparison of each method and reachability.

Case study	Coverage of (n) (conventional procedure)	Coverage of $(n-1)+(n)$ (UCIT)	Reachability
RFID	50.3%	67.9%	68.7%
NFC	28.3%	48.5%	61.5%
IoT	1.4%	9.9%	20.5%
SN	2.8%	52.1%	81.0%
Average	20.7%	44.6%	57.9%

Table 10
Similarities of the IPC subclass distribution for each pair of components.

		Jaccard similarity		Cosine similarity	
		Reachable	Unreachable	Reachable	Unreachable
RFID	Original maximum component	0.9322	0.7331	0.9978	0.9142
	Reachable component	1	0.7102	1	0.9324
NFC	Original maximum component	0.4655	0.4475	0.9912	0.9553
	Reachable component	1	0.4968	1	0.9758
IoT	Original maximum component	0.1342	0.0881	0.9527	0.8386
	Reachable component	1	0.4688	1	0.9387
SN	Original maximum component	0.1287	0.2051	0.9366	0.8755
	Reachable component	1	0.2500	1	0.9540

procedure, even getting close to the reachability point for RFID. For the other three technologies, there was still a gap between UCIT and reachability, but a significant increment was achieved. Regarding the maximum coverage for each technology, on average, 57.9% of patents inside the original network could be connected by using combined datasets. The reachability of IoT was especially small among the case studies, with only 20.5% of the dataset being reachable. This low percentage could be explained by the following two reasons. The first is related to terminology and taxonomy with respect to those technologies whose concepts correspond to IoT but use different terms. Second, given the newness of its technology, IoT requires additional time to acquire citations and connect the network (Marco, 2007).

Patent networks are dynamic. After a certain amount of time, some patents receive citations. However, the opposite is also true; the number of cited references in a patent family may decrease if, for example, the patentee withdraws a country-specific patent or removes a patent. Thus, the reachability is always changing and only reflects a snapshot of the moment the dataset is retrieved.

IPC subclasses were identified for three types of components: maximum, reachable, and unreachable. The frequencies of IPC subclasses within the original maximum and reachable components, as well as within the reachable components and unreachable patents, were compared to determine the technological differences. In Table 10, Jaccard similarity shows that the IPC subclasses have a low similarity compared to the cosine measure. This can be interpreted as that the three types of components are highly similar when it comes to their most frequent IPC subclasses, but that they contain several infrequent ones. From the Jaccard similarity, it can also be observed that RFID and NFC are more specific technologies because they reported higher similarity scores than IoT and SN did. One possible explanation for this difference is that IT practitioners know better the older fields of RFID and NFC, whereas they are looking for applications of newer IoT and SN.

The results of the cosine similarity show that the technological characteristics of reachable and unreachable components can be nearly the same. This finding is consistent with those of previous studies in which IPC classifications do not provide a clear structure. In other words, IPC frequencies overlap within patent networks (Leydesdorff 2008). Therefore, we cannot neglect unreachable components when patent analysis is conducted.

5. Discussions and implications

In Evaluation 1, semantic coherence was compared before and after UCIT was applied, and clusters containing new information were extracted. However, the evaluations were from a macro point of view, and their values were the average of all clusters. In this section, we provide a detailed analysis of the difference between the original and the expanded network.

By including unconnected patents with UCIT, we can change the structure of clusters. This change led to the appearance of what we call clustering phenomenon, which is a classification of the possible mutations of clusters resulting from the inclusion of unconnected elements. We believe that the study of the characteristics of those mutations is important both for understanding the structure of the patent network and for practical usage of our method, as decision makers may be interested in just focusing on either clusters with high relatedness or clusters providing new information.

To identify the phenomenon, we tracked core patents before and after applying UCIT, and we observed that the clusters where they belonged to categorized them into some types of mutations. Then, we measured the semantic coherence and ratio of new information for each phenomenon. Finally, the characteristics of the clustering phenomena were discussed.

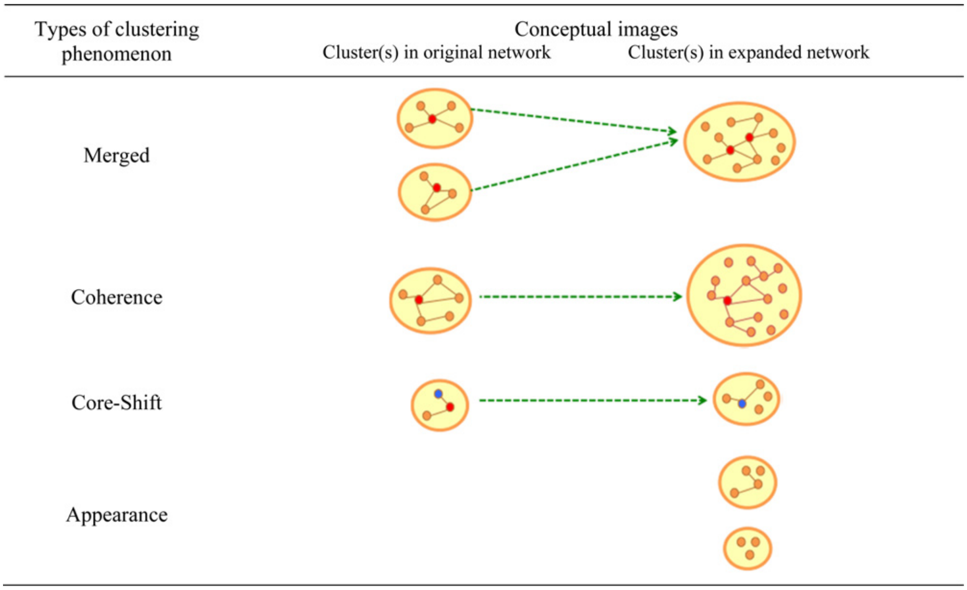

Types of clustering phenomenon
Conceptual images
Cluster(s) in original network
Cluster(s) in expanded network
Merged
Coherence
Core-Shift
Appearance

Fig. 5. Four types of clustering phenomenon. (For interpretation of the references to colour in the text, the reader is referred to the web version of this article.)

5.1. *Observation of the clustering phenomena*

We tracked the manner in which clusters in the original network changed compared to those in the expanded network and observed four types of clustering phenomenon. We labeled them as *Merged*, *Coherence*, *Core-Shift*, and *Appearance*. Conceptual images of the four types of clustering phenomenon are depicted in Fig. 5. The small orange and red dots represent patents and core patents, respectively. The lines indicate links between patents, and the big circles refer to clusters.

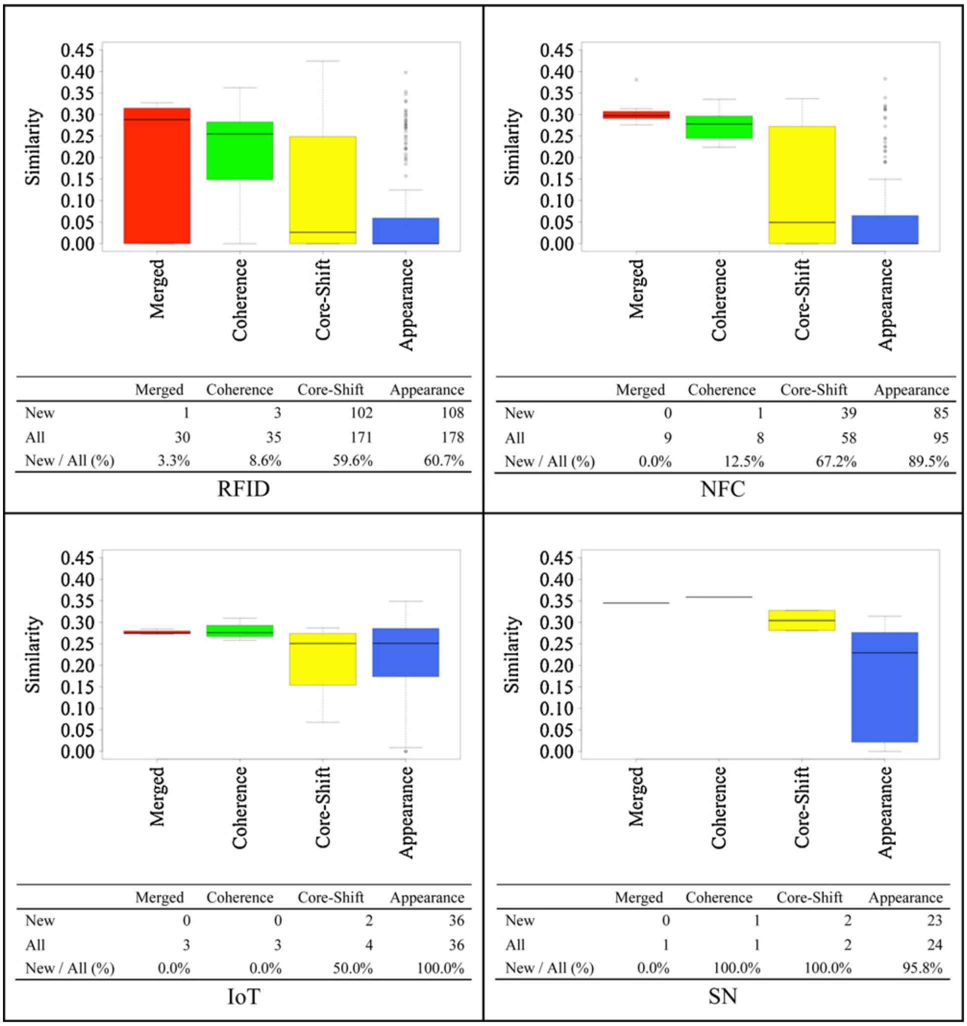
Merged represents a situation in which core patents of two or more clusters in the original network merged into one cluster in the expanded network. *Coherence* refers to the phenomenon in which the expansion strategy did not affect the cluster in a way that other core patents were merged together. *Core-Shift* refers to a situation in which any patent, except core patents, from clusters in the original network shifted to core patents (shown in blue in Fig. 5) in clusters in the expanded network. Finally, *Appearance* refers to the phenomenon in which new clusters that were not in the original network appeared in the expanded network. Clusters in the expanded network may contain connected and unconnected patents. We analyzed how UCIT affected the network structure of patents with the above framework.

5.2. *Characteristics of the four types of clustering phenomenon*

Fig. 6 provides boxplots and tables for each clustering phenomenon. Boxplots were used to show the semantic coherence of clusters in expanded networks, and tables were used to show the percentages of clusters having new information. Based on the boxplots in Fig. 6, *Merged* and *Coherence* clusters showed better similarity scores than did *Core-Shift* and *Appearance*. Thus, *Merged* and *Coherence* can be regarded as more semantically coherent clusters. By contrast, the tables in Fig. 6 indicate that *Core-Shift* and *Appearance* possess much higher percentages of clusters with new information. A special case arose in SN, wherein the *Coherence* cluster reached 100% among clusters having new information. This occurred because only a single *Coherence* cluster appeared in SN. In general, the performance of UCIT proved the value of adding unconnected components to the original network in two regards:

- Unconnected components added semantically coherent information when they were aggregated into *Merged* or *Coherence* clusters.
- Unconnected components added information when they were aggregated into *Core-Shift* or *Appearance* clusters.

A trade-off was observed in the sense that, even though adding unconnected components to *Core-Shift* or *Appearance* clusters created clusters with new information, they were generally of low semantic coherence. The emergence of new clusters might be able to comprehend relevant patent clusters that were missed in the original dataset. However, their semantic coherence is usually low, and some of them may need to be discarded because they are expected to have little relevance to the scope of the analysis. Nevertheless, we need to note that there is a large dispersion in the semantic coherence of *Core-Shift* and *Appearance* clusters. Some clusters in these two categories also reached high levels of semantic coherence.



	Merged	Coherence	Core-Shift	Appearance
New	1	3	102	108
All	30	35	171	178
New / All (%)	3.3%	8.6%	59.6%	60.7%

RFID

	Merged	Coherence	Core-Shift	Appearance
New	0	1	39	85
All	9	8	58	95
New / All (%)	0.0%	12.5%	67.2%	89.5%

NFC

	Merged	Coherence	Core-Shift	Appearance
New	0	0	2	36
All	3	3	4	36
New / All (%)	0.0%	0.0%	50.0%	100.0%

IoT

	Merged	Coherence	Core-Shift	Appearance
New	0	1	2	23
All	1	1	2	24
New / All (%)	0.0%	100.0%	100.0%	95.8%

SN

Fig. 6. Characteristic of the four types of clustering phenomena.

Thus, we cannot simply regard these categories as noise as they contributed to both new information and semantic coherence. Hence, the various functions of unconnected components can be described as follows:

(A) Increase the semantic coherence of clusters with old information
(B) Decrease the semantic coherence of clusters with old information
(C) Increase the semantic coherence of clusters with new information
(D) Decrease the semantic coherence of clusters with new information

In a previous study, Al-Shboul and Myaeng (2014) examined topic drift, which refers to the change of meaning or to the use of a query for an unintended direction following query expansion, and identified the following categories of topics: rising topics are new topics that improve query effectiveness, and drifting topics are new topics that cause query topic drift. UCIT does not refer to query expansion because dataset (n) was not changed before or after the methodology was implemented. UCIT increased the number of patents that were included inside networks within dataset (n). However, in our research, the two categories of topics resemble some roles of unconnected components. Specifically, a rising topic is analogous to Role C and a drifting topic is analogous to Role D. Roles C and D tend to be observed in relatively large networks (i.e., RFID and NFC). This may be because, in those citation networks, the original maximum component consisting of relatively old patents and unconnected components added by UCIT tends to be categorized as clusters with new information.

UCIT can provide information about not only Roles C and D, which were previously discovered, but also Roles A and B.

6. Limitations

One of the limitations of this study was identified through the analysis of IPC subclass frequency. The results showed that reachable and unreachable components tended to share highly frequent IPC subclasses; thus, patents that remained unconnected after the UCIT was applied might still have a relatedness to the maximum component in terms of IPC. However, citation networks can only deal with reachable components and this is also a deficit of the UCIT. In particular, the reachability of IoT, which represents a new and broad concept, was small. For new technologies, hybrid network construction methods could be better in terms of increasing coverage, but the inclusion of noise by using such approaches is yet to be investigated. In favor of purely citation-based networks, there is the consideration that they are not static since even unreachable components may obtain citing references in the future and become reachable. Investigation of the dynamics of change from unreachable to reachable and of the change in clustering phenomena by conducting time-series analysis is necessary in a future study.

Coverage was a key indicator in this research. We looked for clusters that increased the coverage for the sake of a more complete citation network analysis. However, reaching a 100% coverage, i.e., the case wherein all patents in the dataset are present in the network, may not necessarily be good. Some patents with little relatedness can be introduced in the network. This should not be an issue when the patent retrieval strategy ensures the relatedness of all patents in the dataset. If this is not the case, some measures have to be undertaken to remove or avoid the inclusion of noisy patents in the network. Here, we present two possible options for noise removal. One possibility is using the technique used by Shibata, Kajikawa, & Sakata, 2011b, who plotted a direct citation network using a large layout algorithm that concentrates highly connected clusters in the center of the network so that peripheral clusters expected to be little related to those in the center can be spotted right away and then removed after a brief inspection. The other possibility, derived from our own method, is by setting a threshold in the similarity boxplots presented in Fig. 6. For instance, clusters whose inner coherence is less than average similarity can be detected as noisy clusters and thus should be removed. However, the threshold value has not been studied and needs to be explored in a future research study.

7. Conclusion

In this study, we proposed a network expansion method called UCIT. It contributes to the inclusion of unconnected components to the maximum component of a patent network. Case studies of IoT-related technologies were conducted to test the effectiveness of UCIT. In general, clusters in expanded networks obtained by UCIT were better than those in the original network in terms of semantic coherence and brought new information that was missed in the original network. We observed that UCIT increased the number of nodes, especially for relatively small networks (i.e., IoT and SN). The best increment in coverage occurs by including an additional auxiliary dataset of all cited patents listed in the original set of patents.

We also classified four clustering phenomena caused by adding unconnected patents with UCIT and analyzed their characteristics. Specifically, UCIT contributed to an increase in the semantic coherence in *Merged* and *Coherence* clusters and provided new information when aggregated as *Core-Shift* and *Appearance* types of clusters. In practice, inventors or patent officers may benefit from the more complete picture of technology networks brought by UCIT, and by the analysis of clustering phenomena, when evaluating the advancement or uniqueness of the application by investigating the contents of clusters that are related to their invention; policy makers could extract insights about emerging technologies by focusing on clusters with new information (e.g., *Appearance* clusters) since their newness indicates that they are still sparse of citations.

Even though the pure citation-based approach limits our method only to patents with citation information, UCIT has proved to be useful for increasing the coverage when the size of patent citation networks is a concern.

Authors' contributions

Yasutomo Takano: Conceived and designed the analysis, collected the data, contributed data or analysis tools, performed the analysis and wrote the paper.

Cristian Mejia: Collected the data, contributed data or analysis tools, performed the analysis and wrote the paper.

Yuya Kajikawa: Conceived and designed the analysis, performed the analysis and wrote the paper.

Acknowledgments

Part of this research was financially supported by the Research Institute of Science and Technology for Society (RISTEX) of Japan Science and Technology Agency (JST). One of the authors (YT) was also supported by a Grant-in-Aid for Scientific Research from the Japan Society for the Promotion of Science (JSPS) Fellowship Program. The other author (CM) was supported by a Monbukagakusho (MEXT) scholarship.

References

Al-Shboul, B., & Myaeng, S. H. (2014). Analyzing topic drift in query expansion for information retrieval from a large-scale patent dataBase. *2014 International conference on big data and smart computing, BIGCOMP 2014*, 177–182. http://dx.doi.org/10.1109/BIGCOMP.2014.6741432

Altuntas, S., Dereli, T., & Kusiak, A. (2015). Forecasting technology success based on patent data. *Technological Forecasting and Social Change*, *96*, 202–214. http://dx.doi.org/10.1016/j.techfore.2015.03.011. Elsevier Inc.

Atzori, L., Iera, A., & Morabito, G. (2010). The Internet of Things: a survey. *Computer Networks*, *54*(15), 2787–2805. http://dx.doi.org/10.1016/j.comnet.2010.05.010. Elsevier B.V.

Bache, R. (2011). Patent retrieval—a question of access. *World Patent Information*, *33*(4), 345–351.

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

Clauset, A., Newman, M. E. J., & Moore, C. (2004). Finding community structure in very large networks. *Physical Review E*, *70*(6 Pt 2), 1–6. http://dx.doi.org/10.1103/PhysRevE.70.066111

Daim, T. U., Rueda, G., Martin, H., & Gerdsri, P. (2006). Forecasting emerging technologies: use of bibliometrics and patent analysis. *Technological Forecasting and Social Change*, *73*, 981–1012. http://dx.doi.org/10.1016/j.techfore.2006.04.004

De Solla Price, D. J. (1965). Networks of scientific papers. *Science*, *149*(3683), 510.

Fajardo-Ortiz, D., Ortega-Sánchez-de-Tagle, J., & Castaño, V. M. (2015). Hegemonic structure of basic, clinical and patented knowledge on ebola research: a US Army reductionist initiative. *Journal of Translational Medicine*, *13*(1), 124. http://dx.doi.org/10.1186/s12967-015-0496-y

Gretarsson, B., O Donovan, J., Asuncion, A., & Newman, D. (2011). TopicNets: visual analysis of large text corpora with topic modeling. *ACM Transactions on Intelligent Systems and Technology V*, (2), 1–26. http://dx.doi.org/10.1126/science.1178206

Hall, B. H., Jaffe, A., & Trajtenberg, M. (2005). Market value and patent citations. *RAND Journal of Economics*, *36*, 16–38. http://dx.doi.org/10.2307/1593752

Hamedani, M., & Kim, S.-W. (2014). On computing similarity in academic literature data: methods and evaluation. *Web-Age information management: Waim 2014 international workshops*, *vol. 8597*, 403–412. http://dx.doi.org/10.1007/978-3-319-11538-2_37

Harhoff, D., Narin, F., Scherer, F. M., & Vopel, K. (1999). Citation frequency and the value of patented inventions. *Review of Economics and Statistics*, *81*(3), 511–515. http://dx.doi.org/10.1162/003465399558265

Hu, G., & Liu, W. (2015). Nano/micro-electro mechanical systems: a patent view. *Journal of Nanoparticle Research*, *17*(12), 465. http://dx.doi.org/10.1007/s11051-015-3273-1. Springer Netherlands

Hu, Z., Fang, S., & Liang, T. (2014). Empirical study of constructing a knowledge organization system of patent documents using topic modeling. *Scientometrics*, *100*(19), 787–799. http://dx.doi.org/10.1007/s11192-014-1328-1

Karki, A. M., & Krishnan, K. S. (1998). Patent Citation Analysis: A Policy Analysis Tool, World Patent Information 19(4), 269–272 http://dx.doi.org/10.1016/S0172-2190(97)00033-1.

Kessler, M. (1963). Bibliographic Coupling Between Scientific Papers. *American Documentation 14*(1): 10–25.

Klavans, R., Boyack, K. W. (2015). Which Type of Citation Analysis Generates the Most Accurate Taxonomy of Scientific and Technical Knowledge? E-Print, 1–26. arXiv:151105078v2.

Leydesdorff, L. (2008). Patent classifications as indicators of intellectual organization. *Journal of the American Society for Information Science & Technology*, *59*(10), 1582–1597. http://dx.doi.org/10.1002/asi.20814

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. http://dx.doi.org/10.1007/s11192-012-0923-2

Lopez, P., & Romary, L. (2010). Experiments with citation mining and key-term extraction for prior art search. *CLEF 2010—Conference on multilingual and multimodal information access evaluation*.

Mahdabi, P., & Crestani, F. (2014). The effect of citation analysis on query expansion for patent retrieval. *Information Retrieval*, *17*(5-6), 412–429. http://dx.doi.org/10.1007/s10791-013-9232-5

Marco, A. C. (2007). The dynamics of patent citations. *Economics Letters*, *94*, 290–296. http://dx.doi.org/10.1016/j.econlet.2006.08.014

Marra, M., Emrouznejad, A., Ho, W., & Edwards, J. S. (2015). The value of indirect ties in citation networks: SNA analysis with OWA operator weights. *Information Sciences*, *314*, 135–151. http://dx.doi.org/10.1016/j.ins.2015.02.017. Elsevier Inc.

Michel, J., & Bettels, B. (2001). Patent citation analysis: a close look at the basic input data from patent research reports. *Scientometrics*, *51*(1), 185–201. http://dx.doi.org/10.1023/A:1010577030871

Nakamura, H., Suzuki, S., & Kajikawa, Y. (2015). The effect of patent family information in patent citation network analysis: a comparative case study in the drivetrain domain. *Scientometrics*, *104*(2), 437–452. http://dx.doi.org/10.1007/s11192-015-1626-2

Narin, F. (1994). Patent bibliometrics. *Scientometrics*, *30*(1), 147–155. http://dx.doi.org/10.1007/BF02017219

Newman, M. E. J. (2006). Modularity and community structure in networks. *Proceedings of the National Academy of Sciences of the United States of America*, *103*(23), 8577–8582. http://dx.doi.org/10.1073/pnas.0601602103

Ogawa, T., & Kajikawa, Y. (2014). Assessing the industrial opportunity of academic research with patent relatedness: a case study on polymer electrolyte fuel cells. *Technological Forecasting and Social Change*, *90*, 469–475. http://dx.doi.org/10.1016/j.techfore.2014.04.002. Elsevier Inc.

Shibata, N., Kajikawa, Y., & Sakata, I. (2010). Extracting the commercialization gap between science and technology—case study of a solar cell. *Technological Forecasting and Social Change*, *77*(7), 1147–1155. http://dx.doi.org/10.1016/j.techfore.2010.03.008. Elsevier Inc.

Shibata, N., Kajikawa, Y., & Sakata, I. (2011a). Detecting potential technological fronts by comparing scientific papers and patents. *Foresight*, *13*(5), 51–60.

Shibata, N., Kajikawa, Y., & Sakata, I. (2011b). Measuring relatedness between communities in a citation network. *Journal of the American Society for Information Science and Technology*, *62*(7), 1360–1369. http://dx.doi.org/10.1002/asi.21477

Shibata, N., Kajikawa, Y., Takeda, Y., & Matsushima, K. (2009). Comparative study on methods of detecting research fronts using different types of citation. *Journal of the American Society for Information Science and Technology*, *60*(3), 571–580.

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. http://dx.doi.org/10.1002/asi.4630240406

Thomson Innovation. (2016). *Family in Thomson Innovation*.. Retrieved 15.02.16. http://www.thomsoninnovation.com/tip-innovation/support/help/patent_fields.htm#family_members

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. http://dx.doi.org/10.1142/S0219525907000945

Waltman, L., & Van Eck, N. J. (2012). A new methodology for constructing a publication-level classification system of science. *Journal of the American Society for Information Science and Technology*, *63*(12), 2378–2392. http://dx.doi.org/10.1002/asi.22748

Wang, B., Liu, S., Ding, K., Liu, Z., & Xu, J. (2014). Identifying technological topics and institution-topic distribution probability for patent competitive intelligence analysis: a case study in LTE technology. *Scientometrics*, (101), 685–704. http://dx.doi.org/10.1007/s11192-014-1342-3

Whitmore, A., Agarwal, A., & Da Xu, Li. (2014). The Internet of Things—a survey of topics and trends. *Information Systems Frontiers*, (March), 1–14. http://dx.doi.org/10.1007/s10796-014-9489-2

Wilson, P. (1995). Unused relevant information in research and development. *Journal of the American Society for Information Science*, *46*(1), 45–51. http://dx.doi.org/10.1002/(SICI)1097-4571(199501)46:1<45::AID-ASI5>3.0.CO;2-X. Wiley Subscription Services, Inc., A Wiley Company

WIPO, World Intellectual Property Organization. (2015). *Guide to the international patent classification*. Retrieved 15.02.16. http://www.wipo.int/export/sites/www/classifications/ipc/en/guide/guide_ipc.pdf

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

Zhang, Y., Zhang, G., Chen, H., Porter, A. L., Zhu, D., & Lu, J. (2016). Topic analysis and forecasting for science, technology and innovation: methodology with a case study focusing on big data research. *Technological Forecasting and Social Change*, http://dx.doi.org/10.1016/j.techfore.2016.01.015. Elsevier B.V.