Journal of Informetrics 5 (2011) 400–412

Contents lists available at ScienceDirect

Journal of Informetrics

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

Identifying missing relevant patent citation links by using bibliographic coupling in LED illuminating technology

Dar-Zen Chen[a,*], Mu-Hsuan Huang[b], Hui-Chen Hsieh[c], Chang-Pin Lin[d]

[a] Dept. of Mechanical Engineering and Institute of Industrial Engineering, National Taiwan University, No. 1, Sec. 4, Roosevelt Road, Taipei, Taiwan
[b] Dept. of Library and Information Science, National Taiwan University, Taipei, Taiwan
[c] Institute of Industrial Engineering, National Taiwan University, Taipei, Taiwan
[d] Dept. of Mechanical and Mechatronic Engineering, National Taiwan Ocean University, Keelung, Taiwan

ARTICLE INFO

Article history:
Received 24 September 2010
Received in revised form 18 February 2011
Accepted 22 February 2011

Keywords:
Patent citation network
Hidden relevant patent
Citation time lag
Bibliographical coupling
Pareto principle

ABSTRACT

This study uses bibliographic coupling to identify missing relevant patent links, in order to construct a comprehensive citation network. Missing citation links can be added by taking the missing relevant patent links into account. The Pareto principle is used to determine the threshold of bibliographic coupling strength, in order to identify the missing relevant patent links. Comparisons between the original patent citation network and the comprehensive patent citation network with the missing relevant patent links are illustrated at both the patent and assignee levels. Light emitting diode (LED) illuminating technology is chosen as the case study. The relationships between the patents and the assignees are obviously enhanced after adding the missing relevant patent links. The results show that the growth rates on both the total number and the average number of links have apparently improved at the patent level. At the assignee level, the number of linked assignees and the average number of links between two assignees are increased. The differences between the two citation networks are further examined by means of the Freeman vertex betweenness centrality and Johnson's hierarchical clustering. The patents with more new links to other patents have distinct results in terms of the Freeman vertex betweenness centrality. The enhancement of links among patents also results in different clustering.

1. Introduction

Patent citations have been used extensively in measuring the impact of a patent. The more times that a patent is cited, the more impact it has on other patents (Jaffe, Trajtenberg, & Fogarty, 2000; Trajtenberg, 1990). The impact of a patent is also referred as its quality. From this point of view, Lanjouw, Schankerman, and Street (2004) develop an index for assessing patent quality by calculating the numbers of claims, the cited numbers, the citing numbers, and the patent family size. Atallah and Rodriguez (2006) consider not only the number of times that a patent is directly cited, but also the number of patents indirectly citing that patent. Patent citations also show the relationship among patents. Citation paths are very useful in understanding knowledge flow, industrial trends, and technology developments. Hu and Jaffe (2003) identify the knowledge flow from the U.S. and Japan to Korea and Taiwan using patent citations as an indicator. Wartburg, Teichert, and Rost (2005) analyze the patent citations by means of multi-stage measurements of the inventive progress. Multi-stage measurement considers not only direct citations, but also indirect citations and bibliographic coupling. The study presents promising evidence for multi-stage patent citation analysis in illustrating technological changes.

* Corresponding author. Tel.: +886 2 2366 2723; fax: +886 2 2369 2178.
E-mail address: dzchen@ntu.edu.tw (D.-Z. Chen).

1751-1577/$ – see front matter.
doi:10.1016/j.joi.2011.02.005

The patent citation network (PCN) can be used to construct and reveal the relationship among patents. Verspagen (2007) maps the technology trajectories of fuel cells with the use of the PCN. The analysis of citation paths suggests that technological trajectories in fuel cell research are indeed selective and cumulative. Li, Chen, Huang, and Roco (2007) present a network view of patent citation relations that provides a better global understanding of the knowledge-diffusion process.

Citation analysis is used to identify the relationships between citing and cited documents. Meyer (2000) studies the similarities and differences between patent citations and paper citations. They are both widely perceived as measurements of the impact of technology and have much in common, in that the findings in one field can be used as inspiration for research in the other. Patent citations and paper citations share some common properties. Case and Higgins (2000) conclude a reason that researchers cite documents, namely, the references cited in a patent are also a review of prior studies. However, not all of the relevant prior studies are cited as references. As a result, some relevant information may be missing or simply unused. Wilson (1995) studies the causes of unused relevant information among scholarly papers, and then proposes three reasons for unused relevant information, namely: (1) failure to find, (2) information overload, and (3) non-use policy. The citing motivation of patent inventors is similar to that of paper authors. The above reasons for unused relevant information are also highly likely to occur among patents. Unlike papers, a patent for an invention is the grant of a property right to the technique or design innovation of an inventor or assignee. It brings out the seriously rivalrous relationship among patents. An assignee may deliberately choose not to cite the relevant patents of competitors. Moreover, citable materials such as relevant prior publications or patents may be unused due to the failure to find them.

The patent analyses mentioned above, such as patent quality, knowledge flow and spillovers, industrial trends, and technology trajectory, all rely on the discussion of citations among patents. However, there may be some missing relevant patents, and the analyses of patent citations will be inaccurate due to incomplete information on the relationship among patents. Adding the missing relevant patent links (MRPLs) in the citation network would provide a more comprehensive view of the relationship among patents. However, there are few research studies that deal with identifying and utilizing MRPLs for patent citation analysis. This research aims to identify MRPLs through citations, and to make up the missing links for a comprehensive patent citation relationship. The MRPLs can be revealed by the extensive relationships of citations. Bibliographic coupling (BC, proposed by Kessler, 1963) and co-citation (CC, proposed by Small, 1973) are methods currently used for retrieving relevant documents. BC is constructed by the citing relationship, while CC is constructed by the cited relationship. Cleverdon (1967), Harter (1971), Swanson (1971), Small (1973), Braam, Moed, and van Raan (1991), and Chen, Sung, and Kuan (2010) employ BC or CC to discover the relevant literatures that were not found during ordinary studies. Small and Griffith (1974), Garfield (1994), Persson (1994), Morris, Yen, Wu, and Asnake (2003), and Jarneving (2007) use BC or CC clustering to explore the research fronts. Comparisons between these two methods have been performed in several research works (Morris et al., 2003; van den Besselaar & Heimeriks, 2006). BC is immediately available upon publication of the later-issued patent from a BC pair; however, it takes time to retrieve the CC between a pair of patents. Compared with CC, BC provides more current and immediate information about patents. Therefore, BC is chosen for identifying MRPLs in this research.

The value of currency is another concern in this paper. Currency, especially for research and development (R&D) staff, is not an option, but rather a requirement. The preservation of self-status or standing is a very strong private motivation. Briefing others if and when the demand arises is, on the contrary, a social expectation (Wilson, 1993). Furthermore, social pressure reinforces the demand of ethics or law to prevent one from exposure to contempt or being encumbered with a malpractice suit (Keeton, 1984). R&D staff members are similar to players in their own competitive fields; they are unlikely to be successful unless they maintain current knowledge (Bourdieu, 1991). In the proposed method, a comprehensive patent citation network (CPCN) can be established by taking MRPLs into account. In order to identify MRPLs, it is necessary to understand the citation pattern among patents. The citation time lag (CTL) is the time it takes for patents to be cited as references. The CTL is a way to investigate the citation pattern from the aspect of time. In general, the CTL is defined as the time difference between the application date of a citing patent and the issue date of its cited patent. If the CTL of a patent pair is smaller than zero, a relevant patent that is issued might not be cited as a reference because of an irresistible reason, such as the failure to find. It is worthy to note that we found that a high portion of missing links encounter such a situation. It reveals the fact that most of the missing links were caused by the failure to find. The CPCN enhances the concurrent nature throughout the retrieval of the missing relevant patent pairs. Besides, the patent citation relationship is discussed in a network view. Not only the patent level, but also the assignee level, is studied for the PCN. The results also show that the growth rates of both the total number and the average number of links have obviously improved at the patent level; the number of linked assignees and the average number of links between two assignees are increased at the assignee level. Furthermore, at the patent level, the average shortest timeframe for inventors using the issued patents is reduced. At the assignee level, the differences between the original patent citation network (OPCN) and the CPCN will be further evaluated by the Freeman vertex betweenness centrality (Freeman, 1977) and by Johnson's hierarchical clustering with the average-link method (Johnson, 1967), which illustrates the improvement gained by adding the MRPLs.

It is noteworthy that the proposed method is a purely citation-based algorithm. Even though various integrations of citation-based and content-based/text-based algorithms have been extensively developed during the past decade (Cohn & Hofmann, 2001; Fujii, 2007; Fujii, Iwayama, & Kando, 2007; Strohman, Croft, & Jensen, 2007; Torres, McNee, Abel, Konstan, & Riedl, 2004), we firmly believe the claims of Krier and Zaccà (2002) that there can be the intentional use of non-standard terminology, vague terms, and legalistic language in patent documents. The keywords in patent documents in particular

may be written in quite a different style from the description; and cannot be regarded as reliable information for further analysis; which is the reason that the contents or texts of patent documents are not involved in the proposed algorithm.

Light emitting diode (LED) illuminating technology is chosen as the case study. This study assumes that potentially missing links are true missing links, and uses bibliographic coupling to identify the missing relevant patent links in order to construct a comprehensive citation network. It then uses the Pareto principle to determine the threshold of bibliographic coupling strength, in order to identify the missing relevant patent links. Finally, comparisons between the original patent citation network and the comprehensive patent citation network with missing relevant patent links are illustrated at both the patent and assignee levels.

2. Methodology

In order to construct a CPCN, this study aims to identify MRPLs by taking into account the citation links created with the use of bibliographic coupling. The Pareto principle is used to determine the threshold of the bibliographic coupling strength, so as to identify the missing relevant patent links. A systematic algorithm is proposed to extract MRPLs from either patent information or assignee information.

The collection of patent documents from the specific technological domains is the first preparatory work. Patents in specific technology fields are collected based on the various terms of queries and limited conditions from the patent database. Then, the patents are arranged by increasing order of issued date. After collecting the data, the proposed algorithm is composed of three steps. First, the patent citation matrix is constructed. Second, the BC strength of each pair is calculated and the MRPLs are revealed by filtering out the pairs without existing citations and having lower BC strength. Finally, the original citations are integrated with the MRPLs. The detailed process of the proposed algorithm is explained in the following sections.

2.1. Comprehensive patent citation network

Stage 1: Constructing the original patent citation matrix

The OPCN is constructed by the citing and cited relationship among patents. The patents are represented by vertices and the citations by arcs. An arc from vertex i to vertex j denotes that the patent j is in the reference list of patent i. The patent i is defined as the citing patent, and the patent j is defined as the cited patent. A patent can be both a citing and cited patent if it has references and also appears as a reference in other patents' reference list. The vertex-adjacency matrix $\boldsymbol{A}$ for an OPCN at the patent level can be defined as:

$$a_{ij} = \begin{cases} 1 & \text{if the patent } j \text{ is in the reference list of the patent } i \\ 0 & \text{otherwise} \end{cases} \quad (1)$$

where $\boldsymbol{A}$ is an asymmetric $m \times m$ matrix, i.e., $a_{ij} \neq a_{ji}$, $m = |P|$, and P is the set of patents.

Stage 2: Identifying the MRPLs

Two or more documents are said to be bibliographically coupled if they have cited the same references. The strength of the BC is defined as the number of common references. In general, the more references they both cite, the more common technical background they are both based on for development (Kessler, 1963). That is to say, the higher the BC strength between the two patents, the higher the relevance of them (Huang, Chiang, & Chen, 2003). In this paper, the BC strength of each patent pair is calculated, and then the threshold of the BC strength is used to determine whether or not a patent pair without existing citations has enough relevance. The vertex-adjacency matrix $\boldsymbol{B}$ for the BC network of MRPLs at the patent level is defined as:

$$b_{ij} = \begin{cases} 1 & \text{if there are } r \text{ } BC \text{ pairs between patents } i \text{ and } j \text{ for } r \geq \alpha \\ 0 & \text{otherwise} \end{cases} \quad (2)$$

where matrix $\boldsymbol{B}$ is a symmetric $m \times m$ matrix, i.e., $b_{ij} = b_{ji}$ .The value of b_{ij} is one if the BC strength of patents i and j is larger than a specific threshold α, otherwise, it is zero. In this paper, the Pareto principle is used to determine a reasonable α value.

Stage 3: Constructing the comprehensive patent citation matrix

While the patent pairs without existing citations, but that have greater BC strength, are revealed, we integrate the original citations with MRPLs into the construction of the CPCN. The mathematical vertex-adjacency matrix $\boldsymbol{C}$ for the CPCN at the patent level can be defined as:

$$c_{ij} = a_{ij} + b_{ij} \quad (3)$$

where matrix $\boldsymbol{C}$ is an asymmetric $m \times m$ matrix, i.e., $c_{ij} \neq c_{ij}$. The same equations can be applied for analyzing at assignee level using aggregated coupling input values.

2.2. Measurements

2.2.1. Citation time lag

The citation behaviors are highly related in terms of time. For a better understanding of the citation behaviors from the point of view of time, the citation time lag (CTL) is calculated. The CTL is the time length for issued patents being available as references for other inventors. From a mathematical point of view, CTL_{ij} is the time lag between the application date of the citing patent i (APD_i) and the issue date of its cited patent j (ISD_j), which is defined as:

$$CTL_{ij} = APD_i - ISD_j. \tag{4}$$

Normally, patent i cites patent j with the issued date earlier than the application date of patent i, i.e., $ISD_j < APD_i$; the value of CTL_{ij} is greater than zero. A patent can also cite other patents during its examining period. Therefore, the application date of patent i can be earlier than the issued date of patent j, i.e., $ISD_j \geq APD_i$; the value of CTL_{ij} is below zero.

The minimum CTL (MCTL) of a patent i represents the shortest time for patents being cited as references, which is the minimum time lag between the application date of patent i and the issue dates of all of its n_i cited patents, and can be written as:

$$MCTL_i = Min\{CTL_{ij} | 1 \leq j \leq n_i\}. \tag{5}$$

For each year, the available prior data pool is different. The median of the MCTL for patents being cited as references in year y is referred to as the Median(MCTL)$_y$. Finally, the average value of Median(MCTL)$_y$ can be obtained, which means how many times on average an issued patent takes to be cited as a reference.

2.2.2. Growth rate of the total or average links

An index for evaluating the performance of the CPCN is defined on the growth rate of the total links in the network, which represents the MRPLs as a percentage of the existing citations in the OPCN. It can be calculated from the incidence matrices $\boldsymbol{A}$, $\boldsymbol{B}$, and $\boldsymbol{C}$, where $n(\boldsymbol{A})$, $n(\boldsymbol{B})$, and $n(\boldsymbol{C})$ are the numbers of non-zero elements in matrices $\boldsymbol{A}$, $\boldsymbol{B}$, and $\boldsymbol{C}$, respectively. Because the growth rate of the total links of the network is L, G_L is defined by:

$$G_L = \frac{n(\boldsymbol{C}) - n(\boldsymbol{A})}{n(\boldsymbol{A})} = \frac{n(\boldsymbol{B})}{n(\boldsymbol{A})} \tag{6}$$

Another index is the growth rate of the average links in the network, which represents the average MRPLs for each pair of patents under consideration, and thus can be calculated as follows. Because the growth rate of the average links of the network is L, $G_{avg(L)}$ is defined by:

$$G_{avg(L)} = \frac{(n(\boldsymbol{C})/rank(\boldsymbol{C})) - (n(\boldsymbol{A})/rank(\boldsymbol{A}))}{n(\boldsymbol{A})/rank(\boldsymbol{A})} \tag{7}$$

2.2.3. Freeman vertex betweenness centrality

The Freeman vertex betweenness centrality is used to identify the role of brokerage in a network (Freeman, 1977). Vertices that occur on the shortest paths between other vertices have higher betweenness, and are usually considered to be a better broker in relating objects, such as patents and assignees. The Freeman vertex betweenness centrality of vertex j, $C_B(j)$, is calculated by:

$$C_B(j) = \sum_{i \neq j \neq k \in U} \frac{p_{ik}(j)}{p_{ik}} \tag{8}$$

where p_{ik} is the number of shortest paths from assignee i to assignee k, and $p_{ik}(j)$ is the number of shortest paths from assignee i to assignee k that pass through assignee j.

3. Results and discussion

In order to demonstrate the feasibility of the research methodology, the light emitting diode (LED) illuminating technology is chosen for the case study in this paper. United States Patent Classification (USPC) categories are used to represent different patent technology fields. The LED illuminating technology is represented by the current USPC 362(ILLUMINATION)/800(LIGHT EMITTING DIODE), and the date range of the retrieved data is from 1976/01/01 to 2008/12/31. There are 1810 patents collected in this LED illuminating technology category.

3.1. Original patent citation network

3.1.1. Original patent citation network at patent level

The retrieved citing and cited relationship among the patents is as follows. There are a total of 1810 issued patents, which include 1375 (76%) patents with citing, 1049 (58%) patents that have been cited, and 253 (14%) non-citing and non-cited patents. The total citing number of the citing patents is 7414 within these issued patents, while the average citing number

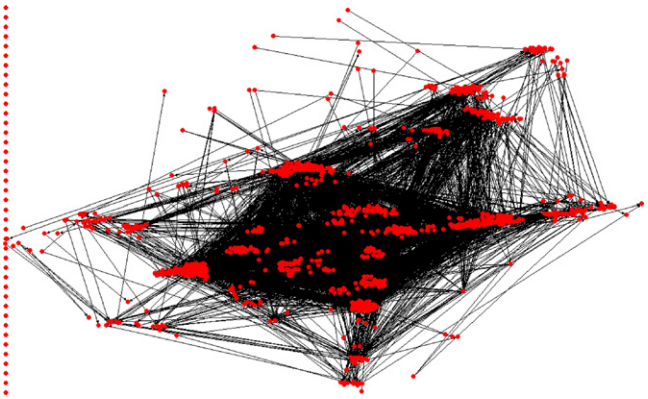

Fig. 1. OPCN at patent level in LED illuminating technology.

is 5.4 and the average cited number is 7.1. There are 1557 (86%) patents connected by their citing and cited relationship, and the average links between each pair of patents is approximately 4.8.

The time differences between the issue date (*ISD*) of the cited patent and the application date (*APD*) of the citing patent are also calculated. This is for the purpose of understanding the citation behavior from the aspect of timeframe. There are 6410 (86%) citations with cited patents issued before the application date of the citing patents ($APD_{\text{citing}} > ISD_{\text{cited}}$). The other 1004 (14%) citations with cited patents issued after the application date of the citing patents ($APD_{\text{citing}} \leq ISD_{\text{cited}}$) are added by examiners or by inventors into the information disclosure statement (IDS) during the examination process. The average value of Median$(MCTL)_y$ is 85.8 weeks in the LED domain. This means that, on average, it takes an issued patent 85.8 weeks to be cited as a reference, which is quite reasonable under real-world conditions. After a valuable patent is applied or issued, firms pick up the relevant ideas at once and then begin a series of R&D tasks. This valuable patent will not appear in the reference lists of subsequent patents until the successful R&D is patented.

In this paper, the networks are analyzed through the use of the social network analysis software UCINET, which was developed by Borgatti, Everett, and Freeman (2002). UCINET provides the following visualization that can be embellished by the functions of NetDraw. The OPCN at the patent level is shown in Fig. 1. Most of the patents are densely connected with each other. Because of the complex and confusing links of the OPCN at the patent level, the relationships in the network at the assignee level will be discussed in Section 3.1.2 for a clearer view.

3.1.2. Original patent citation network at assignee level

In the above OPCN, there are 1117 assignees among these 1810 patents, and 34 patents are owned by multi-assignees. This paper focuses on studying the assignees who own more patents than others, based on the number of assignees and the distribution of the numbers of their patents, in this case, the top 22, as shown in Table 1. These assignees are referred to as the major assignees in the current USPC 362/800.

All cited and citing percentages presented are based only on the citations within the current USPC 362/800. As shown in Table 1, most of the major assignees have over 50% of their patents cited by other patents. There are four assignees without self-citations (the total of others citing and cited), namely, Toyoda Gosei, Eastman Kodak, Philips Electronics, and U.S. Philips. All of the major assignees have over 55% of their patents citing other patents within the current USPC 362/800. This demonstrates a close interaction between the major assignees and the other assignees.

There are 82 pairs of assignees in this case study, and each pair of linked assignees has 3.83 citations between them on average. There could also be missing relationships among these major assignees. The strongest relationship (46 citations) is between Stanley Electric and 911 Emergency Products. On average, each major assignee is linked to 2.18 assignees through a citation relationship. Stanley Electric has the maximum 15 linked assignees, while Osram Sylvania, Matsushita Electric, Minnesota Mining & Manufacturing, Streamlight, Rohm, Avago Technologies, Armament Systems & Procedures, and Eastman Kodak have no linked assignees.

In order to explore the strong links between the major assignees, the network is simplified as shown in Fig. 2.There are 20 pairs of the top strongly linked major assignees, and they consist of 14 major assignees. As can be seen in the network of Fig. 2, Color Kinetics and 911 Emergency Products have many links to other major assignees in the top-20 strong relationships. They are both linked to Stanley Electric and U.S. Philips. However, there is no link between Color Kinetics and 911 Emergency Products, which suggests a possible missing relationship between the two.

3.2. Identifying missing relevant patent links

In order to identify the MRPLs, it is necessary to understand the patent citation behaviors. During the development period, the inventor searches for relevant prior studies according to his/her research topic. The relevant patents issued before the application date of this invention are very likely to be cited as references. However, some relevant patents may be missing

Table 1
Top 22 assignees in LED illuminating technology.

Rank	Assignee	Patents	Cited %		Citing %	
			Self	Other	Self	Other
1	Koito	30	9.2	90.8	32.7	67.3
2	911 Emergency Products	24	72.4	27.6	10.6	89.4
3	Stanley Electric	21	1.6	98.4	8.9	91.1
4	Hewlett-Packard Development	18	1.8	98.2	4.7	95.3
4	Toyoda Gosei	18	0.0	100.0	0.0	100.0
6	Osram Sylvania	16	54.5	45.5	17.6	82.4
7	Color Kinetics	14	75.3	24.7	11.0	89.0
7	GELcore	14	3.6	96.4	2.5	97.5
7	Matsushita Electric	14	9.5	90.5	2.9	97.1
10	Armament Systems & Procedures	13	73.1	26.9	19.4	80.6
10	Dialight	13	6.6	93.4	9.8	90.2
12	Cao Group	12	34.7	65.3	11.3	88.7
12	Koninklijke Philips Electronics	12	3.1	96.9	3.3	96.7
12	Rohm	12	4.8	95.2	5.6	94.4
15	Sharp	10	2.5	97.5	2.9	97.1
16	Avago Technologies	9	50.0	50.0	4.0	96.0
16	Eastman Kodak	9	0.0	100.0	0.0	100.0
16	General Electric	9	5.0	95.0	8.3	91.7
16	Minnesota Mining & Manufacturing	9	85.7	14.3	18.8	81.3
16	Philips Electronics	9	0.0	100.0	0.0	100.0
16	Streamlight	9	22.2	77.8	4.3	95.7
16	U.S. Philips	9	0.0	100.0	0.0	100.0

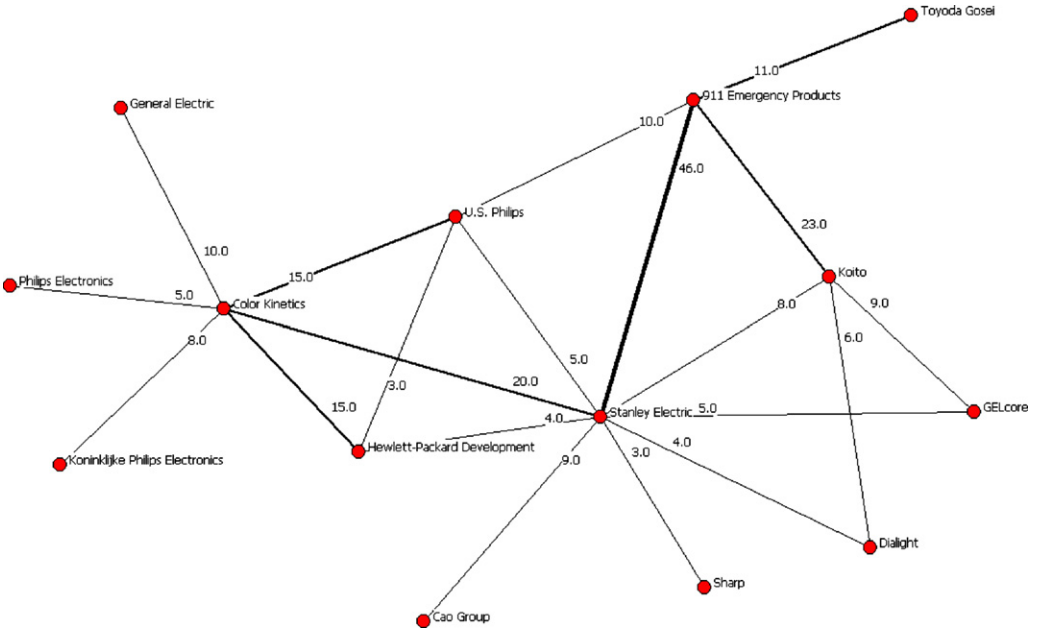

Fig. 2. Top 20 strong linked major assignees in OPCN in LED illuminating technology.

due to failure to find, policy of non-use, or information overload. Patent pairs with a strong BC, but no existing citation relation, are then considered as MRPLs in this study.

3.2.1. Threshold of BC strength

In previous research studies, BC and CC were the two popular methods for retrieving relevant documents or for exploring research fronts. As we mentioned earlier, because BC provides more current and complete information about documents than CC, it is chosen for identifying the MRPLs in this research. There are 45,225 distinct BC pairs with 73,542 units of BC. Over 94% of the BC pairs have no existing citation, and 6% of them have. Among the 7414 citations, 37% of them are identified by the existing BC, while 63% are not; therefore, the BC pairs account for 37% of the existing citations.

There can be different numbers of BC between each patent pair, which is referred to as the BC strength. The BC strength represents the correlation between two patents. In identifying the MRPLs, the BC pairs with low BC strength should be

Table 2
Number of patent pairs with different BC strengths in LED illuminating technology.

BC strength (x)	BC pairs with existing citation (y)		BC pairs without existing citation (z)	
1	1443	53.1%	33,135	78.0%
2	610	22.4%	5866	13.8%
≥3	665	24.5%	3506	8.2%
Mean	$\Sigma xy/\Sigma y = 3.41$		$\Sigma xz/\Sigma z = 1.51$	

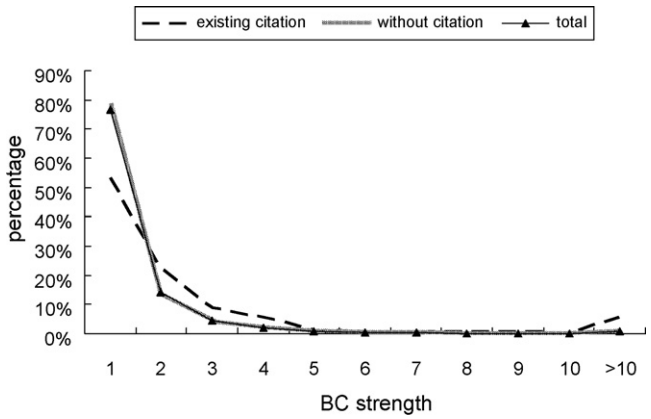

Fig. 3. BC strength distribution in LED illuminating technology.

excluded. The numbers of patent pairs with variant BC strength are shown in Table 2. Over 78% of the BC pairs without existing citations have a BC strength of 1. Therefore, the mean BC strength of the pairs without existing citations is quite low (1.51).

The strength of BC exhibits a Pareto distribution, as shown in Fig. 3. In Pareto distributions, a high-frequency or high-amplitude population is followed by a low-frequency or low-amplitude population that gradually tails off asymptotically. The events at the far end of the tail have a very low probability of occurrence (Persky, 1992).

For a BC relationship, participation is referred to as the relevance based on the BC strength. In order to determine a sufficiently large set $(100 - k)\%$ of participants, a reasonable evaluation of the relevance must be first determined. Table 3 shows the BC strengths for different k. The BC distribution is a discrete distribution, so k is determined from the percentage of the BC pairs without existing citations for different BC strengths. For the BC pairs with existing citations, the mean BC strength is 3.41, as shown in Table 2. This accounts for the missing relationship among patents. When the number of BC pairs reaches 3.41, there may be a missing relationship between them.

The mean BC strength for the $(100 - k)\%$ participants should be more than 3.41, so that the relevance is high enough to be a sufficiently large set of participants. Note that the mean values of the tail probability of Fig. 3, when w is greater than or equal to 2 and 3, are listed in Table 3. Therefore, the threshold value α in Eq. (2) is set to 3, and the major relationships consist of 8.2% (3506) BC pairs without existing citations. As shown in Fig. 4, the BC network of the MRPLs at the patent level is a dense network that represents many missing relationships among patents. It is worthy to note that the CTL of the MRPLs is 48.3 weeks, which is 37.5 weeks shorter than the CTL of the OPCN.

3.2.2. Characteristic of the missing links

We examine a characteristic of those missing links in terms of the CTL index. The CTL is the time length for issued patents being available as references for other inventors. The CTL will lose its meaning if a patent i cites a later issued patent j ($APD_i \leq ISD_j$), because there is no information available about the cited patent when the citing patent files its application. In other words, a relevant patent issued might not be cited as a reference because of an irresistible reason: failure to find. In our experiment, we found that there are 2297 missing links with a CTL smaller than zero, which are the majority (up to 65.52%) among all missing links. It reveals the fact that most of the missing links were caused by failure to find, which is

Table 3
BC strength in determining relevant pairs in LED illuminating technology.

BC strength for determining $(100 - k)\%$ sufficiently large set of participants ($\geq w$)	$k\%$	$(100 - k)\%$	Mean = $\Sigma xz/\Sigma z$
≥2	78.0	22.0	3.32
≥3	91.8	8.2	5.53

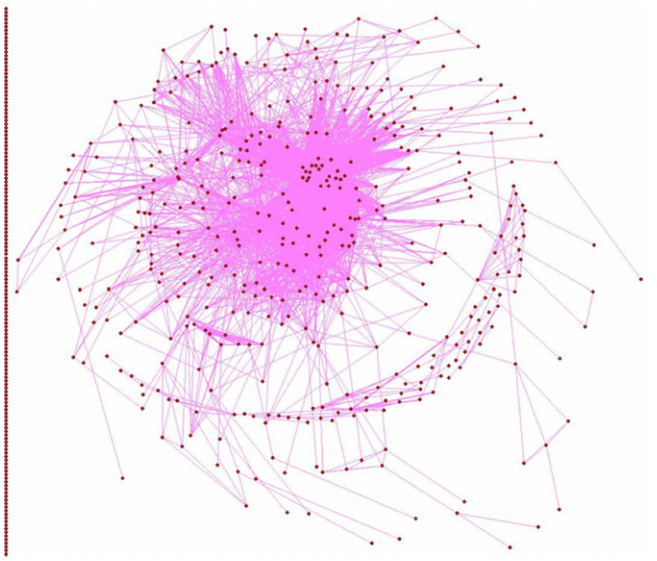

Fig. 4. BC network of HRPLs at patent level in LED illuminating technology.

one of the important sources for current information. Keeping up with currency is necessary for keeping a competitive edge (Laskin, 1994).

We further removed the examiner-added citations and calculated the BC strength only with the inventor-added citations in order to retrieve the suggested missing links. The fraction of the suggested missing links that were added by examiners was then calculated, yielding a value of 2.5%. Nevertheless, we do not feel surprised at such a low fraction due to the majority (about 80%) of the suggested missing links caused by the failure to find. Generally, if two patents do not only have a very similar topic but also have very close temporal relationship, inventors of these two patents may not have the chance to cite each other. Meanwhile, examiners might also not be able to create a citation relationship between these two patents. In sum, the total number of the suggested missing links that were also added by examiners is scarce since this study is focused on extracting the current information in order to keep a competitive edge.

3.3. Comprehensive patent citation network

3.3.1. Comprehensive patent citation network at patent level

A CPCN can be established by taking into account the MRPLs that were discovered in the previous sections. The 1557 patents are now connected by citing, cited, and an additional BC relationship, as shown in Fig. 5. The network of the CPCN at the patent level is quite complex, as can be seen in Fig. 5. The average number of links in the CPCN for LED is 7. There are 7414 citations in the original OPCN. After constructing the BC relationship among the patents and identifying the strong BC pairs, it turns out that 3506 MRPLs referred to the add-in links. By considering the MRPLs and making up the missing links between the patents, the number of links in the CPCN is raised to 10,920, including citations in the OPCN and the MRPLs identified by predefined strong BC pairs. The number of total links grows 47% from the OPCN to the CPCN after taking the MRPLs into account. The growth rate of average links of the connected patents from the OPCN to the CPCN is 46%, which was 4.8 in the OPCN and increased to 7 in the CPCN. In this case study, the growth rates on both the total number and the average number of links have reached nearly 50%, revealing a tremendous improvement in establishing relationships among patents, by finding out the possible missing links through BC. The CTL for the CPCN is obviously different from the one for the corresponding OPCN. After adding the MRPLs, the mean of the Median$(\mathrm{MCTL})_y$ is reduced by 3.3 weeks, from 85.8 weeks to 82.5 weeks.

3.3.2. Comprehensive patent citation network at assignee level

The CPCN for the major assignees is constructed based on the procedure presented in Section 2.1 except to use aggregated coupling as input values. The network is denser than that of the OPCN. Table 4 shows the number of linked assignees and the rank for each assignee in the OPCN vs. the CPCN. On average, an assignee is linked to 2.18 assignees in the OPCN and is linked to 11.86 assignees in the CPCN. Nonetheless, not all of the assignees link to more assignees in the CPCN than in the OPCN.

The total number of assignee pairs in the CPCN is 151, while there are only 82 pairs in the original OPCN. The average number of links between two assignees increases from 3.83 in the OPCN to 51.01 in the CPCN. The relationships among patents have obviously enhanced after the MRPLs are included. After adding the MRPLs, the links can be classified into three types: new links, enhanced links, and unchanged links. The relationships are shown in Fig. 6, which illustrates the links of the

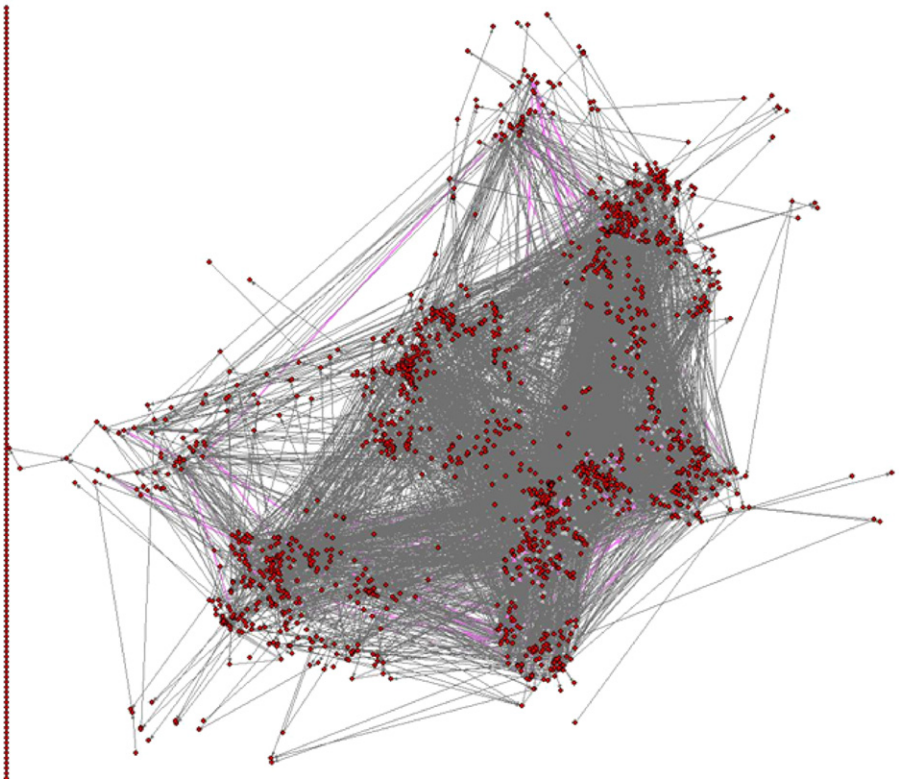

Fig. 5. CPCN at patent level in LED illuminating technology.

top-20 strong linked major assignees in the CPCN. The new links of the top-20 strong linked major assignees after adding the MRPLs are 36 pairs. Two pairs of assignees not connected in the OPCN are now strongly linked in the CPCN: 911 Emergency Products vs. Color Kinetics, and 911 Emergency Products vs. Cao Group. The enhanced links are between Color Kinetics and two major assignees, including Hewlett-Packard Development and Koninklijke Philips Electronics. The other enhanced links

Table 4
Number of linked assignees for each assignee in OPCN vs. CPCN in LED illuminating technology.

Assignee	Rank		Number of linked assignees		
	CPCN	OPCN	OPCN	CPCN	Add-in
Color Kinetics	1	2	6	21	15
911 Emergency Products	1	5	4	21	17
Hewlett-Packard Development	3	5	4	18	14
GELcore	4	7	3	16	13
Cao Group	4	11	1	16	15
Matsushita Electric	4	15	0	16	16
Stanley Electric	7	1	9	15	6
Koninklijke Philips Electronics	8	11	1	14	13
Toyoda Gosei	9	8	2	13	11
Dialight	9	8	2	13	11
Koito	11	2	6	11	5
General Electric	11	8	2	11	9
Sharp	13	11	1	10	9
Philips Electronics	13	11	1	10	9
Osram Sylvania	13	15	0	10	10
Avago Technologies	16	15	0	9	9
Minnesota Mining & Manufacturing	17	15	0	8	8
Rohm	18	15	0	7	7
Streamlight	18	15	0	7	7
U.S. Philips	20	2	6	6	0
Eastman Kodak	20	15	0	6	6
Armament Systems & Procedures	22	15	0	3	3
Mean			2.18	11.86	

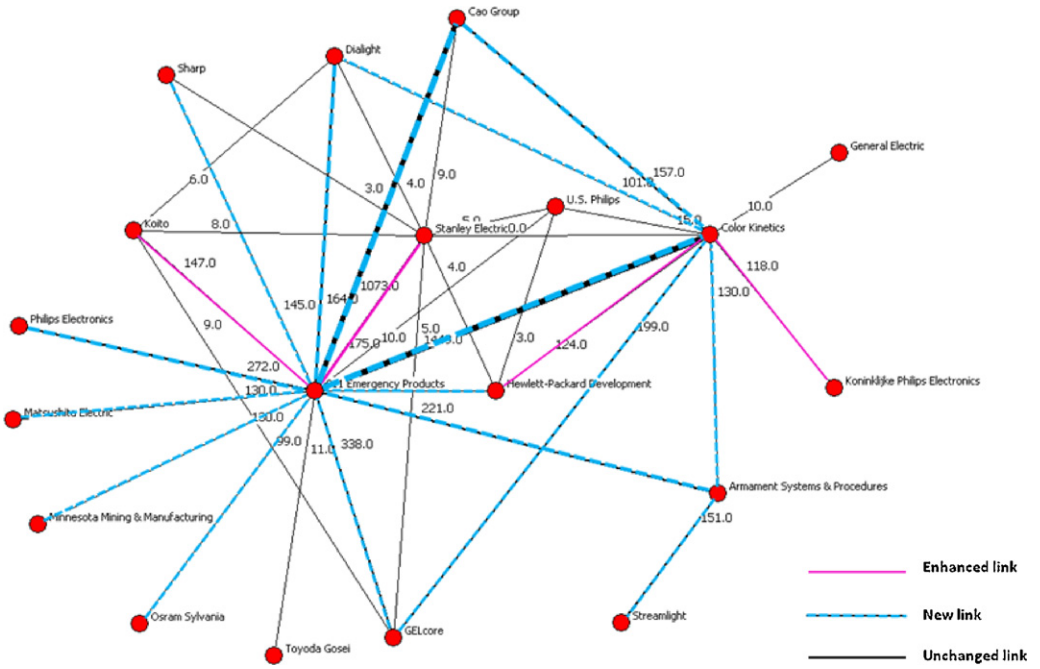

Fig. 6. Links of the top 20 strong linked major assignees in CPCN in LED illuminating technology.

are between 911 Emergency Products and two major assignees, including Koito and Stanley Electronics. There are 16 links that remain unchanged when considering the MRPLs.

In a network at the assignee level, the Freeman vertex betweenness centrality can be used to evaluate the ability of being the technology transmitter among other assignees. As shown in Table 5, the ranks on betweenness centrality for 6 assignees move up in the list, while 10 assignees stay the same and 6 assignees move down after adding the MRPLs to the network. 911 Emergency Products and Cao Group are originally in the third and fourth quarters in the OPCN, respectively, and are moved to the first quarter in the CPCN. These assignees of increasing rank are linked to more assignees after the MRPLs are taken into consideration. They have a greater possibility to be the intermediary of the shortest path for other assignees.

Table 5
The Freeman vertex betweenness centrality of OPCN vs. CPCN in LED illuminating technology.

Assignee	Rank			Quarter		$C_B(j)$	
	OPCN	CPCN	Variation	OPCN	CPCN	OPCN	CPCN
Hewlett-Packard Development	1	3	−2	1	1	15.06	3.016
Stanley Electric	2	6	−4	1	2	11.276	1.739
Color Kinetics	3	1	2	1	1	10.735	8.19
Koito	4	16	−12	1	4	10.44	0.665
U.S. Philips	5	5	0	1	1	9.191	2.192
General Electric	6	12	−6	2	3	4.225	0.884
Avago Technologies	7	15	−8	2	3	3.686	0.773
Philips Electronics	8	9	−1	2	2	3.524	1.352
GELcore	9	8	1	2	2	3.505	1.461
Dialight	10	17	−7	2	4	3.33	0.537
Minnesota Mining & Manufacturing	11	19	−8	3	4	3	0.172
Matsushita Electric	12	7	5	3	2	1.835	1.688
Toyoda Gosei	13	11	2	3	3	1.458	0.948
Koninklijke Philips Electronics	14	10	4	3	2	1.454	1.21
911 Emergency Products	15	1	14	3	1	1.329	8.19
Sharp	16	18	−2	4	4	0.373	0.388
Cao Group	17	4	13	4	1	0.159	2.802
Osram Sylvania	18	13	5	4	3	0.103	0.878
Rohm	19	20	−1	4	4	0.079	0.169
Streamlight	20	14	6	4	3	0	0.794
Eastman Kodak	20	21	−1	4	4	0	0.048
Armament Systems & Procedures	20	22	−2	4	4	0	0

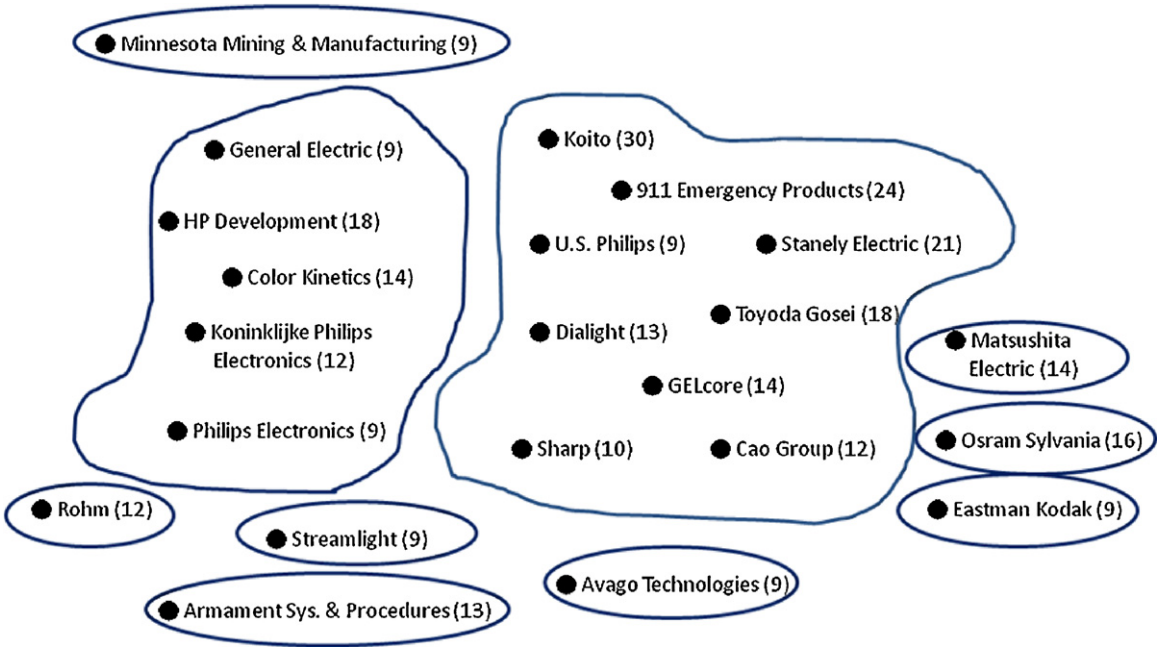

Fig. 7. The clustering result in OPCN of the major assignees in LED illuminating technology.

Therefore, the value of the Freeman vertex betweenness centrality of these assignees is higher than that of the assignees with small changes to their links. It is also found that the strong central vertices of the OPCN are the top-five assignees in the first quarter, while those of the CPCN are Color Kinetics and 911 Emergency Products.

The clustering method proposed by Johnson (1967) is based on the distances (similarities) between these clusters. In this research, the average-link method is chosen for computing the distance (similarity), which is considered to be the average distance from all members of one cluster to all members of another cluster. The cluster formations in the OPCN and the CPCN are shown in Figs. 7 and 8, respectively. They exhibit the different cluster number and the composition of each cluster in the

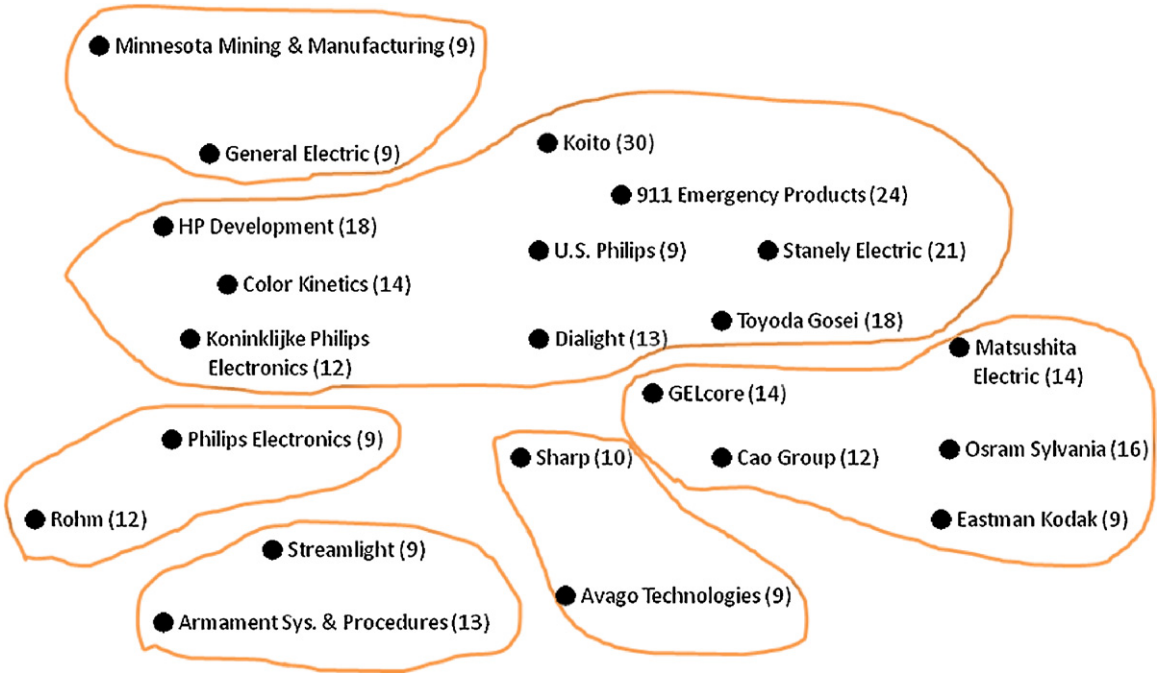

Fig. 8. The clustering result in CPCN of the major assignees in LED illuminating technology.

OPCN and the CPCN. The new links between the assignees that are not connected in the OPCN cause such variation. Due to the add-in MRPLs, there are some new links in the CPCN, which result in different clusters from the OPCN to the CPCN. In this LED illuminating technology case, the clustering result of the OPCN at the assignee level for the major assignees is 10 clusters, where eight clusters are in isolation. After adding the MRPLs, the CPCN reduces to 6 clusters, which is 4 less than the result of the OPCN. Also, the isolated clusters in the OPCN are linked to other clusters in the CPCN.

4. Conclusions

Inventors record relevant information as prior art, but not all of the relevant prior studies would be cited as references. As we observed earlier, the failure to find is the main reason that renders the inventor unable to cite prior art. This work aims to identify the missing relevant information about patents, and then further constructs the CPCN. In identifying the MRPLs, BC shows effectiveness in revealing the relevance between patents. Not all of the BC pairs are strong enough to be identified as MRPLs. The Pareto principle is applicable in identifying strong BC pairs as the Pareto distribution of the BC strength. It is considered that MRPLs are the BC pairs without existing citations, but having a greater mean strength than the BC pairs with existing citations. In the LED case, the threshold value of the strength of BC for a patent pair to be MRPLs is 3, and thus there are 3506 BC pairs of patents considered to be MRPLs. Some of the missing citation links can then be made up by adding the MRPLs, and a CPCN can be constructed as well.

At the patent level, the growth rates on both the total number and the average number of links have reached nearly 50%, showing an effective result of identifying the MRPLs. The average shortest time frame for inventors using the issued patents is reduced from 85.8 weeks in the OPCN to 82.5 weeks in the CPCN. At the assignee level, an assignee is linked to nine more assignees in the CPCN than the linked assignees of 2.18 in the OPCN. The average number of links between two assignees is increased from 3.83 in the OPCN to 51.01 in the CPCN. The relationships among patents/assignees are obviously enhanced after adding the MRPLs. There are two pairs of assignees who are not connected in the OPCN, but are now strongly linked in the CPCN, namely, 911 Emergency Products vs. Color Kinetics, and 911 Emergency Products vs. Cao Group. The differences between the OPCN and the CPCN are studied through the analysis of the Freeman vertex betweenness centrality and Johnson's hierarchical clustering. After embedding the MRPLs, the Freeman vertex betweenness centrality of each assignee is changed. The assignees linked to more assignees in the CPCN are more likely to be the intermediary of the technology transition for other assignees. The results of clustering are also changed due to the new links between patents. The number of clusters is 10 in the OPCN and becomes 6 in the CPCN with a different composition of each cluster.

The CPCN provides a more comprehensive view of the relationships among patents after taking currency information into account. Patent citations are extensively used for evaluating the impact of a patent and the relationship among patents, including the measurement of the quality of patents, the knowledge flow among countries or institutes, technology developments, and industrial trends. By using the approach of identifying MRPLs that is described in this paper, patent citation analysis can be applied more appropriately for future research.

Acknowledgements

We would like to thank the anonymous reviewers and the editor for their valuable comments and suggestions.

References

Atallah, G., & Rodriguez, G. (2006). Indirect patent citations. *Scientometrics*, *67*(3), 437–465.
Braam, R. R., Moed, H. F., & van Raan, A. F. J. (1991). Mapping of science by combined co-citation and word analysis. I. Structural aspects. *Journal of the American Society for Information Science*, *42*(4), 233–251.
Borgatti, S. P., Everett, M. G., & Freeman, L. C. (2002). *UCINET 6 for windows: Software for social network analysis*. Retrieved March 14, 2011, from http://pages.uoregon.edu/vburris/hc431/Ucinet_Guide.pdf.
Bourdieu, P. (1991). The peculiar history of scientific reason. *Sociological Forum*, *6*(1), 3–26.
Case, D. O., & Higgins, G. M. (2000). How can we investigate citation behavior? A study of reasons for citing literature in communication. *Journal of the American Society for Information Science*, *51*(7), 635–645.
Chen, D. Z., Sung, Y. S., & Kuan, C. H. (2010). Identifying core patents by citations, bibliographic coupling and co-citation. In *Proceedings of the Eleventh International Conference on Science and Technology Indicators: Book of Abstracts* (pp. 69–70).
Cleverdon, C. W. (1967). The Cranfield tests on index language devices. *Aslib Proceedings*, *19*(6), 173–194.
Cohn, D., & Hofmann, T. (2001). The missing link—A probabilistic model of document content and hypertext connectivity. *Advances in Neural Information Processing Systems*, *13*, 430–436.
Freeman, L. C. (1977). A set of measures of centrality based on betweenness. *Sociometry*, *40*(1), 35–41.
Fujii, A. (2007). Enhancing patent retrieval by citation analysis. In *Proceedings of the 30th annual ACM SIGIR conference on research and development in information retrieval* (pp. 793–794).
Fujii, A., Iwayama, M., & Kando, N. (2007). Overview of the patent retrieval task at the NTCIR-6 workshop. In *Proceedings of NTCIR-6 workshop meeting* (pp. 359–365).
Garfield, E. (1994). Research fronts. *Current Contents*, *41*, 3–7.
Harter, S. P. (1971). The Cranfield II relevance assessments: A critical evaluation. *The Library Quarterly*, *41*(3), 229–243.
Hu, A. G. Z., & Jaffe, A. B. (2003). Patent citations and international knowledge flow: The cases of Korea and Taiwan. *International Journal of Industrial Organization*, *21*(6), 849–880.
Huang, M. H., Chiang, L. Y., & Chen, D. Z. (2003). Constructing a patent citation map using bibliographic coupling: A study of Taiwan's high-tech companies. *Scientometrics*, *58*(3), 489–506.
Jaffe, A. B., Trajtenberg, M., & Fogarty, M. S. (2000). Knowledge spillovers and patent citations: Evidence from a survey of inventors. *American Economic Review*, *90*(2), 215–218.

Jarneving, B. (2007). Bibliographic coupling and its application to research-front and other core documents. *Journal of Informetrics*, *1*(4), 287–307.
Johnson, S. C. (1967). Hierarchical clustering schemes. *Psychometrika*, *32*(3), 241–254.
Keeton, W. P. (Ed.). (1984). *Prosser and Keeton on the law of torts*. St. Paul, MN: West Publishing Co.
Kessler, M. M. (1963). Bibliographic coupling between scientific papers. *American Documentation*, *14*(1), 10–25.
Krier, M., & Zaccà, F. (2002). Automatic categorisation applications at the European patent office. *World Patent Information*, *24*(3), 187–196.
Lanjouw, J. O., Schankerman, M. A., & Street, H. (2004). Patent quality and research productivity: Measuring innovation with multiple indicators. *Economic Journal*, *114*(495), 441–465.
Laskin, D. M. (1994). Dealing with information overload. *Journal of Oral and Maxillofacial Surgery*, *52*(7), 661.
Li, X., Chen, H., Huang, Z., & Roco, M. C. (2007). Patent citation network in nanotechnology (1976–2004). *Journal of Nanoparticle Research*, *9*(3), 337–352.
Meyer, M. (2000). What is special about patent citations? Differences between scientific and patent citations. *Scientometrics*, *49*(1), 93–123.
Morris, S. A., Yen, G. G., Wu, Z., & Asnake, B. (2003). Time line visualization of research fronts. *Journal of the American Society for Information Science and Technology*, *54*(5), 413–422.
Persky, J. (1992). Restrospectives: Pareto's law. *Journal of Economic Perspectives*, *6*(2), 181–192.
Persson, O. (1994). The intellectual base and research fronts of JASIS 1986–1990. *Journal of the American Society for Information Science*, *45*(1), 31–38.
Small, H. G. (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.
Small, H. G., & Griffith, B. C. (1974). The structure of scientific literatures. (I) Identifying and graphing specialties. *Science Studies*, *4*(1), 17–40.
Strohman, T., Croft, W. B., & Jensen, D. (2007). Recommending citations for academic papers. In *Proceedings of the 30th annual international ACM SIGIR conference on research and development in information retrieval* (5th ed., pp. 705–706).
Swanson, D. R. (1971). Some unexplained aspects of the Cranfield tests of indexing performance factors. *The Library Quarterly*, *41*(3), 223–228.
Torres, R., McNee, S. M., Abel, M., Konstan, J. A., & Riedl, J. (2004). Enhancing digital libraries with TechLens+. In *Proceedings of the international conference on digital libraries* (pp. 228–236).
Trajtenberg, M. (1990). A penny for your quotes: Patent citations and the value of innovations. *The Rand Journal of Economics*, *21*(1), 172–187.
van den Besselaar, P., & Heimeriks, G. (2006). Mapping research topics using word-reference co-occurrences: A method and an exploratory case study. *Scientometrics*, *68*(3), 377–393.
Verspagen, B. (2007). Mapping technological trajectories as patent citation networks: A study on the history of fuel cell research. *Advances in Complex Systems*, *10*(1), 93.
Wartburg, I. V., Teichert, T., & Rost, K. (2005). Inventive progress measured by multi-stage patent citation analysis. *Research Policy*, *34*(10), 1591–1607.
Wilson, P. (1993). The value of currency. *Journal of the American Society for Information Science*, *41*(4), 632–642.
Wilson, P. (1995). Unused relevant information in research and development. *Journal of the American Society for Information Science*, *46*(1), 45–51.