Journal of Informetrics 11 (2017) 63–79

Contents lists available at ScienceDirect

Journal of Informetrics

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

Do patent citations indicate knowledge linkage? The evidence from text similarities between patents and their citations

Lixin Chen

Department of Management and Economics, North China University of Water Conservancy and Electric Power, Zhengzhou, Henan, China

ARTICLE INFO

Article history:
Received 20 October 2015
Received in revised form 10 April 2016
Accepted 10 April 2016
Available online 21 November 2016

Keywords:
Patent citation
Knowledge linkage
Text similarity
Citing-cited patent pairs
Non-citing-cited patent pairs
Examiner citation
Applicant citation

ABSTRACT

Whether patent citations indicate knowledge linkage is still a controversial issue, which is very important for the widespread use of the patent citation analysis method. We hypothesize that there exists technological knowledge linkage between patents and their citations, and that the linkage can be detected through measuring text similarities between them. To test the hypothesis, we selected citing-cited patent pairs as the observation group and selected patent pairs without citing-cited relationship as the control group. Using the VSM with WF-IDF weighting method, we calculated text similarity values of the two groups. Through comparing text similarity values between the two groups, we validate that in the vast majority of cases text similarity values of citing-cited pairs are much higher than those of non-citing-cited pairs. The study in nano-technology field shows that the above results are the same, although patents in the same technological area are more relevant than in different technological areas. Furthermore, by comparing text similarities between applicant and examiner citing-cited pairs, the results show that in more cases examiner citations indicate knowledge linkage a bit better than applicant citations. Preferably, examiner citations can be regarded as not only the supplement of applicant citations but also the more important technological background and the prior art closely related to the patents. Compared to applicant citations, examiner citations are a good indicator of knowledge linkage rather than an incomplete and noisy indicator. In short, the results suggest that most certainly patent citations can indicate knowledge linkage, and more likely examiner citations can indicate knowledge linkage a bit better than applicant citations, especially for the component of patent claims. Therefore, we accept the hypothesis that patent citations can indicate knowledge linkage, which is the basic assumption of the patent citation analysis method.

1. Introduction

Patent citations, similar to paper citations, are the references of patents. Listing references is an obligation of the patent applicant, who should comply with the legal requirement to disclose the prior art derived from previous patents, and who should supply a complete description of the state of the art in the field of the invention as the technological background (Criscuolo & Verspagen, 2008; Jaffe, Trajtenberg, & Fogarty, 2000). Paper citations can be extensively applied to investigate the linkage of scientific knowledge for tracking science development (Garfield, Sher, & Torpie, 1964), revealing knowledge flow and diffusion (Liu & Rousseau, 2010), mapping science structure (Leydesdorff & Rafols, 2009; Park & Leydesdorff, 2009)

E-mail address: lynnchenlixin@163.com

http://dx.doi.org/10.1016/j.joi.2016.04.018
1751-1577/

and so on. Can patent citations be used to reveal the linkage of technological knowledge? This issue, however, has been in dispute ever since the patent citation analysis method began to be used.

Many researchers believe that patent citations, which are the same as paper citations, can reflect the linkage of technological knowledge. Narin and Olivastro (1988) suggest that patent citations can indicate linkages between companies, between technological areas, and between technology and science. Narin and his cooperators (Carpenter, Cooper, & Narin, 1980; Narin & Olivastro, 1992; Narin & Olivastro, 1998; Narin, Hamilton, & Olivastro, 1995; Narin, Hamilton, & Olivastro, 1997) have used the scientific papers within patent references to measure the linkage between science and technology. In recent decades, patent citations have been extensively used as a proxy for measuring technological knowledge linkage (Callaert, Grouwels, & Looy, 2012; Criscuoloa & Verspagen, 2008; Hu, Chen, Huang, & Roco, 2007; Lo, 2010; Meyer, 2001; Ribeiro, Ruiz, Bernardes, & Albuquerque, 2010; Schmoch, 1997; Tijsen, 2005; Verbeek et al., 2002).

However, some researchers (Jaffe et al., 2000) hold the opposite opinions and suggest that patent citations indicate the technological knowledge linkage incompletely. Patent citations look like paper citations, but they are different in many respects (Meyer, 2000; Michel & Bettels, 2001). According to US patent laws, not only the applicant is required to provide references when filing a patent application, but also the patent examiner need to add references for judging patentability through searching the related prior art during the patent examining period.[1] The survey by Jaffe et al. (2000) shows that inventors respond that one half patent citations indicate no knowledge spillover, since addition of citations by the patent examiner is unknown to the inventor and has no effect on the invention. They suggest that the inventor citations, instead of the total citations, should be taken as an indicator of knowledge flow. Li and Meng (2010) insist that simply applying patent citation data to indicate knowledge linkage is both conceptually and technologically illogical and unreasonable. Li, Chambers, Ding, Zhang, and Meng (2014) further point out that only scientific papers self-cited by the inventor should be used for measuring linkage between science and technology, and the examiner citation and non-self-citation by inventor papers should be excluded due to excessive "noise".

Obviously, it is a complex and controversial issue whether patent citations can indicate knowledge linkage or not. The disputations mainly focus on the differences of the two types of patent citations that are respectively added by applicants and examiners (Alcacer & Gittelman, 2006; Alcacer, Gittelman, & Sampat, 2009). Jaffe et al. (2000) regard examiner citations as an incomplete and noisy indicator for measuring knowledge flow and diffusion. However, basing their study on European Patent Office search reports, Criscuoloa and Verspagen (2008) find that there are mainly two kinds of examiner citations: documents of particular relevance that restrict patent application claims, accounting for 36%; and references related to technological background, accounting for 62%. Patent attorneys anticipate citations most likely to be added by examiners, so that examiner and applicant citations may come to resemble each other closely (Alcacer & Gittelman, 2006). That means examiner citations are relevant to applicant citations, which are used for disclosing prior art and describing technological background. Li et al. (2014) suggest that examiner citations can indicate knowledge linkage logically because the citing behavior of examiner is regulated by patent laws.

Despite many limitations, criticisms and disputations, by using the patent citation analysis method, many researchers have revealed technological knowledge flow, diffusion and transfer (Chen & Hicks, 2004; Hu & Jaffe, 2003; Nelson, 2009; Park & Suh, 2013), have traced the technological trajectories and the technological frontiers (Epicoco, 2013; Érdi et al., 2013; Martinelli, 2012; Mina, Ramlogan, Tampubolon, & Metcalfe, 2007), have mapped the technological knowledge domain to display its structure and its relation (Lai & Wu, 2005; Wang, Zhang, & Xu, 2011; Weng & Daim, 2012; Yeh, Sung, Yang, Tsai, & Chen, 2013), have explored the technology and patent classification (Shih & Liu, 2010), and even have evaluated patent, technology and innovation ability at the levels of organization, region and country (Albert, Avery, Narin, & McAllister, 1991; Carpenter & Narin, 1983; Lanjouw & Schankeman, 2004; Verspagen, 2000; Wartburg, Teichert, & Rost, 2005; Yeh et al., 2013; Yoon & Park, 2004).

Similar to the paper citation analysis method, the patent citation analysis method must be provided with a basic condition or assumption: patent citations can reflect knowledge linkage. Undoubtedly, the assumption is an essential and crucial issue for the widely use of the patent citation analysis method. Only based on this condition, can technological knowledge flow and diffusion be disclosed, can the technological frontiers be traced, can the technological knowledge structure be mapped, and can technology innovations be evaluated by using the method. Therefore, we pose a hypothesis that patent citations can indicate knowledge linkage: patents are similar to or are relevant to their citations in technological knowledge, and that the linkage can be detected through measuring text similarities between them. Through the empirical research, we try to gather supporting evidence for the hypothesis.

2. Data

The United States Patent and Trademark Office (USPTO) provides all kinds of patent documents freely. We constructed the origin database through downloading the full text documents of USPTO patents that were granted during the 1976–2013 period, with over 300G storage capacity in all. In our research, only utility patents, which are the most common of all patents, were taken as the study object. There are more than four million full text documents of utility patent in the origin database.

[1] Applicant citations are submitted on the filing date (the priority date), which is the submission date of a patent application, whereas examiner citations are added during the patent examining period, after the filing date but before the granting date (the date of authorizing the patent).

Patent citations include patent and non-patent literatures.[2] Because of some limitations,[3] we just select citing-cited USPTO utility patent pairs as the dataset for the research. The vast majority of USPTO patents are utility patents and most of their citations are to other USPTO patents; it means that the dataset of the citing-cited USPTO utility patent pairs is a representative sample for measuring the text similarities between patents and their citations.

The utility patents granted in December 2013 and January 2009 were taken as the samples. There are more than 25 thousand utility patents granted in December 2013, and 23,505 patents among them meet the above-mentioned condition that a patent has at least one USPTO utility patent as a reference. These 23,505 patents contain 501,585 citations, which are composed of the utility patents granted by USPTO in the period of 1976–2013. The cited patents granted before 1976 were removed for lack of full text documents. We took these 501,585 citing-cited patent pairs, which are named citing-cited pairs, as the observation group.

For comparative analysis, we randomly selected 1/30 of the utility patents granted in December 2013 and 1/10 of the utility patents granted in January 2009 as dataset of the control group. There are more than 11 thousand utility patents granted in January 2009. We obtained nearly 100 million patent pairs; one patent of each pair was granted in 2013, and the other was granted in 2009. A 4–5 year period is chosen as the time gap of two patents, since a patent is mostly cited 4–5 years after granted according to our statistical results. We removed such a kind of patent pair at least one patent of which cited the other, with the result that the two patents matched in the same pair are not in a citing-cited relationship. Finally, we obtained 990,616 patent pairs without citing-cited relationship, which are named non-citing-cited pairs, as the control group.

3. Methodology

In the field of Patent Bibliometrics, to explore and reveal the technological knowledge linkage, in addition to the controversial method of patent citation analysis, another common and important method is patent content analysis. The patent content analysis method uses some analysis techniques, such as co-word analysis (Engelsman & van Raan, 1994), semantic analysis (Park, Yoon, & Kim, 2012), inventions functional trees analysis (Cascini & Zini, 2008) and patent text mining (Feldman & Sanger, 2007), to reveal the technological knowledge linkage (Fattori, Pedrazzi, & Turra, 2003; Feng & Leng, 2012; Oostdijk, Verberne, & Koster, 2010; Tseng, Lin, & Lin, 2007; Yoon & Park, 2004).

Our hypothesis is that patent citations can indicate knowledge linkage, which means that patents and their citations should be similar or relevant in text content. Whether they are similar or not can be detected through measuring text similarities between them. It is possible that any two patents are similar to each other, but in general the text similarity between two patents without citing-cited relationship should be weaker than those with citing-cited relationship. If the text similarity value between a citing-cited pair is generally higher than that between a non-citing-cited pair (that means there exists knowledge linkage or relevance within a citing-cited pair), then the hypothesis can be accepted; otherwise, the hypothesis can be rejected.

There are many methods to measure text similarity, among which the vector space model (VSM) (Salton, Wong, & Yang, 1975) is an important and mature one. Magerman, van Looy, and Song (2010) validate the use of text mining techniques based on the VSM to detect text similarities between patent documents and scientific publications. Ahlgren and Colliander (2009) point out that the text mining techniques based on the VSM have higher accuracy when compared with several other document similarity measuring methods.

In our research, the text mining techniques based on cosine similarity for the VSM are used to detect similarities between patents and their citations. The method is very efficient, especially for sparse vectors (such as vectors of patent title and abstract), as only the non-zero dimensions need to be considered. Assume that there are two text documents a and b, which can be expressed by the term frequency vectors: $V(a) = \{tf_{t1,a}, tf_{t2,a}, tf_{t3,a}, \ldots, tf_{ti,a}, \ldots, tf_{tn,a}\}$ and $V(b) = \{tf_{t1,b}, tf_{t2,b}, tf_{t3,b}, \ldots, tf_{ti,b}, \ldots, tf_{tn,b}\}$, in which $tf_{ti,a}$ indicates the term frequency of term i in document a. The text similarity value can be calculated by the cosine function:

$$sim(a, b) = \cos\left(V(a), (V(b)) = \frac{V(a) \cdot V(b)}{|V(a)||V(b)|}\right. \quad (1)$$

The similarity value represents the ratio of the dot product and the module product of two term frequency vectors.

The above VSM method does not consider the impacts of document frequency and term frequency. In our research, term frequency (tf) is weighted by using the sub-linear scaling method (Manning, Raghavan, & Schütze, 2008):

$$wf_{t,a} = 1 + \log tf_{t,a} (tf_{t,a} > 0), \text{else } wf_{t,a} = 0 (tf_{t,a} = 0). \quad (2)$$

[2] According to our statistical results based on 11.7 million citations from 270 thousand patents granted in 2012, USPTO patents as references account for 61%, foreign patents as references make up 18%, and non-patent references account for 21% of the total citations. The proportions of the three kinds of references are nearly the same as those of Cotropia et al. (2013), whose study shows that the proportions are 64%, 14% and 22% respectively, whereas Michel and Bettels (2001) find that USPTO patents as references account for 90%.

[3] For non-patent references, such as journal articles, books, conference proceedings, technical standards and reports, it is very difficult to obtain full text documents. For foreign patents as references, such as China patents and Japan patents, are non-English literatures; it is extremely difficult to measure text similarities owing to language differences.

Table 1
Similarity values of titles and abstracts from citing-cited and non-citing-cited pairs.

value	title-title				abstract-abstract			
	cit pairs	(%)	non-cit pairs	(%)	cit pairs	(%)	non-cit pairs	(%)
1	4285	0.85	17	0.00	2381	0.47	1	0.00
(0.9, 1)	1155	0.23	6	0.00	830	0.17	3	0.00
(0.8, 0.9]	2446	0.49	14	0.00	497	0.10	4	0.00
(0.7, 0.8]	4187	0.83	41	0.00	563	0.11	2	0.00
(0.6, 0.7]	7572	1.51	73	0.01	1119	0.22	7	0.00
(0.5, 0.6]	12,300	2.45	184	0.02	1954	0.39	12	0.00
(0.4, 0.5]	20,069	4.00	544	0.05	6633	1.32	33	0.00
(0.3, 0.4]	31,472	6.27	1509	0.15	24,137	4.81	151	0.02
(0.2, 0.3]	46,445	9.26	5381	0.54	72,768	14.51	1436	0.14
(0.1, 0.2]	54,367	10.84	22,068	2.23	159,369	31.77	23,958	2.42
(0, 0.1]	82,430	16.43	176,874	17.85	206,740	41.22	716,366	72.32
0	234,857	46.82	783,905	79.13	24,594	4.90	248,643	25.10

Document frequency is weighted by using the inverse document frequency (*idf*) method (Robertson, 2004; Sparck-Jones, 1972):

$$idf_t = \log \frac{N}{df_t} \tag{3}$$

in which N is the number of total documents, and df_t is the number of documents in which term t appears. *Wf-idf* instead of *tf-idf* (Salton & Buckley, 1988; Wu, Luk, Wong, & Kwok, 2008) is used as the weighting method:

$$wf - idf_{t,a} = wf_{t,a} \times idf_t \tag{4}$$

The *wf-idf* weighting method decreases the weight of common words appearing in most documents and increases the weight of terms with low document frequency, and it contributes to getting relatively better effects when measuring text similarity. In addition, for improving the accuracy of measuring text similarity, the stop words given by Lextek International are used to delete some of the most common and short function words, and the Porter stemming algorithm (Porter, 1980) is applied to remove the commoner morphological and inflexional endings from words when processing patent texts.

Using cosine similarity for the VSM with *wf-idf* term weighting method, we can calculate text similarity values of citing-cited and non-citing-cited pairs. Furthermore, by comparing text similarity values between the two groups, we can examine whether text similarity values of citing-cited pairs are generally higher than those of non-citing-cited pairs or not, for accepting or rejecting the hypothesis.

4. Analyses and results

4.1. Local comparative analyses of text similarities of the four components between citing-cited and non-citing-cited pairs

A patent document is mainly composed of the four components: title, abstract, description and claims. Using self-made software, we calculated the text similarity values between title–title (one title of which is from the citing patent and the other is from the cited patent), between abstract–abstract, between description–description and between claims–claims derived from the 501,585 citing-cited pairs. In the same way, we calculated the text similarity values of the four components from the 990,616 non-citing-cited pairs.

Table 1 shows the distribution of text similarity values of titles. The value varies from 0 to 1.[4] The greater a value is, the more similar the two titles are. If a value is equal to one, then the two vocabularies of patent title terms are the same. There are 4285 citing-cited pairs and 17 non-citing-cited pairs whose title–title text similarity values are equal to 1. It means that the former patent pairs contain more entirely similar titles than the latter pairs. If a text similarity value is equal to zero, then the two vocabularies of patent title terms are completely different. For the former pairs 47% of the values are zero, whereas for the latter pairs 79% of the values are zero. That means the former pairs contain less completely different titles

[4] The cosine value ranges from 0 to 1, whereas the Pearson correlation coefficient runs from −1 to +1. A value of cosine > 0.2 is in the neighborhood of $r = 0$ in the article of Egghe and Leydesdorff (2009). However, our case is different; there are 64,918 terms, which means the dimensions of the vector are very large, but the most elements are zero owing to term sparseness (many terms do not appear in a patent document), especially for title and abstract. So in our case, the average values $\bar{x}$, $\bar{y}$ are approximate to zero, and the Pearson's correlation coefficient $Correl(X, Y) = \frac{\sum (x-\bar{x})(y-\bar{y})}{\sqrt{\sum (x-\bar{x})^2 \sum (y-\bar{y})^2}}$ is near to the value of cosine $Cos(X, Y) = \frac{\sum xy}{\sqrt{\sum x^2 \sum y^2}}$.

Table 2
Similarity values of descriptions and claims from citing-cited and non-citing-cited pairs.

value	description-description				claims-claims			
	cit pairs	(%)	non-cit pairs	(%)	cit pairs	(%)	non-cit pairs	(%)
1	30	0.01	0	0.00	2	0.00	0	0.00
(0.9, 1)	4958	0.99	3	0.00	507	0.10	0	0.00
(0.8, 0.9]	1259	0.25	1	0.00	786	0.16	0	0.00
(0.7, 0.8]	1586	0.32	0	0.00	1051	0.21	4	0.00
(0.6, 0.7]	3088	0.62	9	0.00	1545	0.31	12	0.00
(0.5, 0.6]	8222	1.64	25	0.00	3177	0.63	23	0.00
(0.4, 0.5]	40,277	8.03	287	0.03	9693	1.93	91	0.01
(0.3, 0.4]	153,429	30.59	4892	0.49	37,539	7.48	607	0.06
(0.2, 0.3]	213,476	42.56	70,541	7.12	116,925	23.31	4507	0.45
(0.1, 0.2]	71,510	14.26	534,639	53.97	211,021	42.07	80,717	8.15
(0, 0.1]	3304	0.66	380,219	38.38	119,339	23.79	904,655	91.32
0	446	0.09	0	0.00	0	0.00	0	0.00

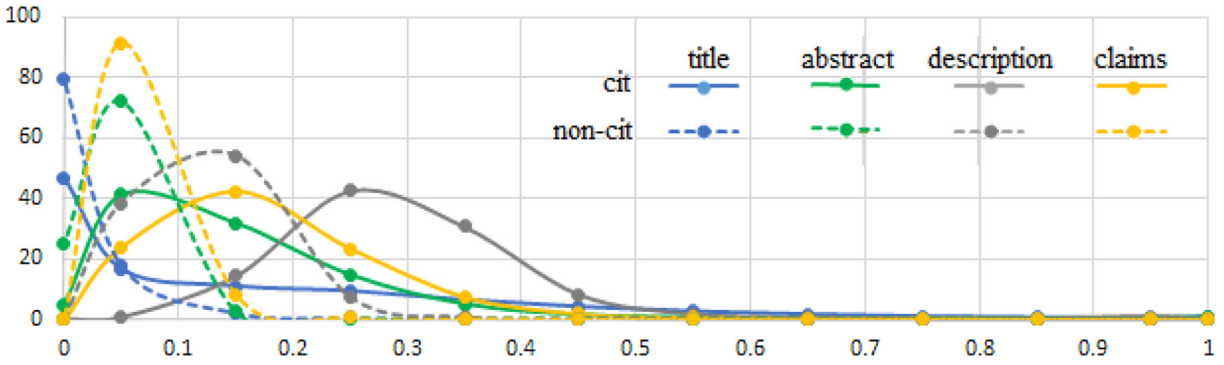

Fig. 1. Distributions of text similarity values of the four components.

than the latter pairs. For the former pairs 37% of the values are higher than 0.1, whereas for the latter pairs the percentage is 3%. For the former pairs 26% of the values are higher than 0.2, whereas for the latter pairs the proportion is not more than eight thousandths. It indicates that the title–title text similarity values of citing-cited pairs are most likely higher than those of non-citing-cited pairs.

Table 1 also shows the distribution of text similarity values of abstracts. For citing-cited pairs only 5% of the values are zero, whereas for non-citing-cited pairs 25% of the values are zero. For the former pairs 54% of the values are higher than 0.1, whereas for the latter pairs the proportion is 3%. For the former pairs 22% of the values are higher than 0.2, whereas for the latter pairs the proportion is two thousandths. It indicates that the abstract–abstract text similarity values of citing-cited pairs are most likely higher than those of non-citing-cited pairs.

Table 2 shows the distribution of text similarity values of descriptions. For citing-cited pairs 85% of the values are higher than 0.2, whereas for non-citing-cited pairs the proportion is 8%. For the former pairs 42% of the values are higher than 0.3, whereas for the latter pairs the proportion is five thousandths. It indicates that the description–description text similarity values of citing-cited pairs are most likely higher than those of non-citing-cited pairs.

Table 2 also shows the distribution of text similarity values of claims. For citing-cited pairs 76% of the values are higher than 0.1, whereas for non-citing-cited pairs the proportion is 9%. For the former pairs 34% of the values are higher than 0.2, whereas for the latter pairs the proportion is five thousandths. It indicates that the claims–claims text similarity values of citing-cited pairs are most likely higher than those of non-citing-cited pairs.

From Fig. 1, it can be seen that the four curves of citing-cited pairs are located at the right side of those of non-citing-cited pairs. When text similarity value becomes higher and is over the threshold value (the values of titles, abstracts, claims and descriptions are about 0.05, 0.1, 0.1 and 0.2 respectively), the distribution density of the former pairs is higher than that of the latter pairs. It indicates that text similarity values of the four components of the former pairs are distributed more in the relatively higher value zone, whereas those of the latter pairs are distributed more in the relatively lower value zone.

In short, the results of the local comparative analyses show that the text similarity values of the four components of citing-cited pairs are most likely higher than those of non-citing-cited pairs. However, for citing-cited pairs, 63% of the values of titles and 46% of the values of abstracts are lower than 0.1, as low as the values of non-citing-cited pairs (97% of those values are lower than 0.1). It means that a large number of titles and abstracts from citing-cited pairs are not similar to each other. Nevertheless, it cannot be concluded that a half citing-cited pairs are not relevant in text content. Both abstracts and titles are too short to express patent documents fully and to measure text similarity accurately. Therefore, it is necessary to compare

Table 3
Correlation coefficients of text similarity values of the four components.

	title	abstract	description	claims	subtotal
title	–	0.255	0.211	0.235	0.701
abstract	0.255	–	0.451	0.577	1.282
description	0.211	0.451	–	0.662	1.324
claims	0.235	0.577	0.662	–	1.473

Note: For any two patents that hardly have a citing-cited relationship, so we use the similarity values of the non-citing-cited pairs (990,616 in all) instead of the citing-cited pairs to calculate the correlation coefficient.

Table 4
Comprehensive similarity values of citing-cited and non-citing-cited pairs.

value	cit pairs	(%)	non-cit pairs	(%)
1	0	0.00	0	0.00
(0.9, 1)	1387	0.28	1	0.00
(0.8, 0.9]	1583	0.32	2	0.00
(0.7, 0.8]	1277	0.25	0	0.00
(0.6, 0.7]	1408	0.28	2	0.00
(0.5, 0.6]	2832	0.56	6	0.00
(0.4, 0.5]	9181	1.83	28	0.00
(0.3, 0.4]	39,860	7.95	198	0.02
(0.2, 0.3]	130,309	25.98	2388	0.24
(0.1, 0.2]	244,146	48.67	79,129	7.99
(0, 0.1]	69,602	13.88	908,862	91.75
0	0	0.00	0	0.00

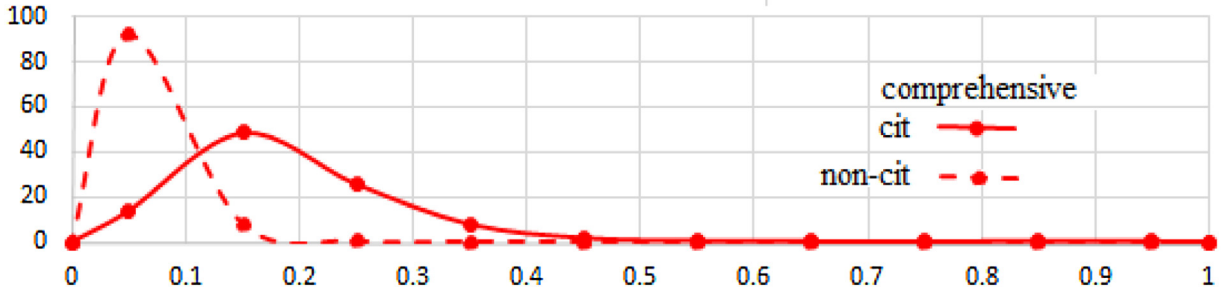

Fig. 2. Distributions of the comprehensive similarity values of citing-cited and non-citing-cited pairs.

the differences of text similarities between citing-cited and non-citing-cited pairs at the global level by integrating the four components.

4.2. Global comparative analysis of text similarities between citing-cited and non-citing-cited pairs

The four components of a patent document are different in text length and function, but are related to each other in content. The patent title is an invention theme, the abstract is a brief summary, the description is an illustration to specify and disclose an invention, and the claim is a lawsuit of the extent of the protection conferred by a patent. When calculating the text similarity between two patent documents, it is necessary to consider the differences and connections of the four components. We give a simple comprehensive method to measure the text similarities of two patent documents according to the Pearson product-moment correlation coefficients of similarity values of the four components (see Table 3). The comprehensive text similarity value V integrating the four components can be calculated by using the following equation:

$$V = w_t V_t + w_a V_a + w_d V_d + w_c V_c = 0.147 V_t + 0.268 V_a + 0.277 V_d + 0.308 V_c \quad (5)$$

in which, w_t, w_a, w_d and w_c are the weights of the four components, V_t, V_a, V_d and V_c are respectively the text similarity values of them (the total value of the matrix is $sum = 4.780$, that of title is $sum_t = 0.701$, and the weight of title is $w_t = sum_t/sum = 0.147$).

Using Eq. (5), the comprehensive text similarity values of citing-cited and non-citing-cited pairs were calculated (see Table 4). For the former pairs 14% are lower than 0.1 and 86% of the values are higher than 0.1, whereas for the latter pairs 92% of the values are lower than 0.1 and 8% of the values are higher than 0.1. For the former pairs 37% of the values are higher than 0.2, whereas the proportion of the latter pairs is three thousandths. It indicates that most of the comprehensive text similarity values of citing-cited pairs are higher than those of non-citing-cited pairs.

From Fig. 2, it can be seen that when the text similarity value is over the threshold value (about 0.1), the distributed density of citing-cited pairs is greater than that of non-citing-cited pairs. For the former pairs most values are higher than

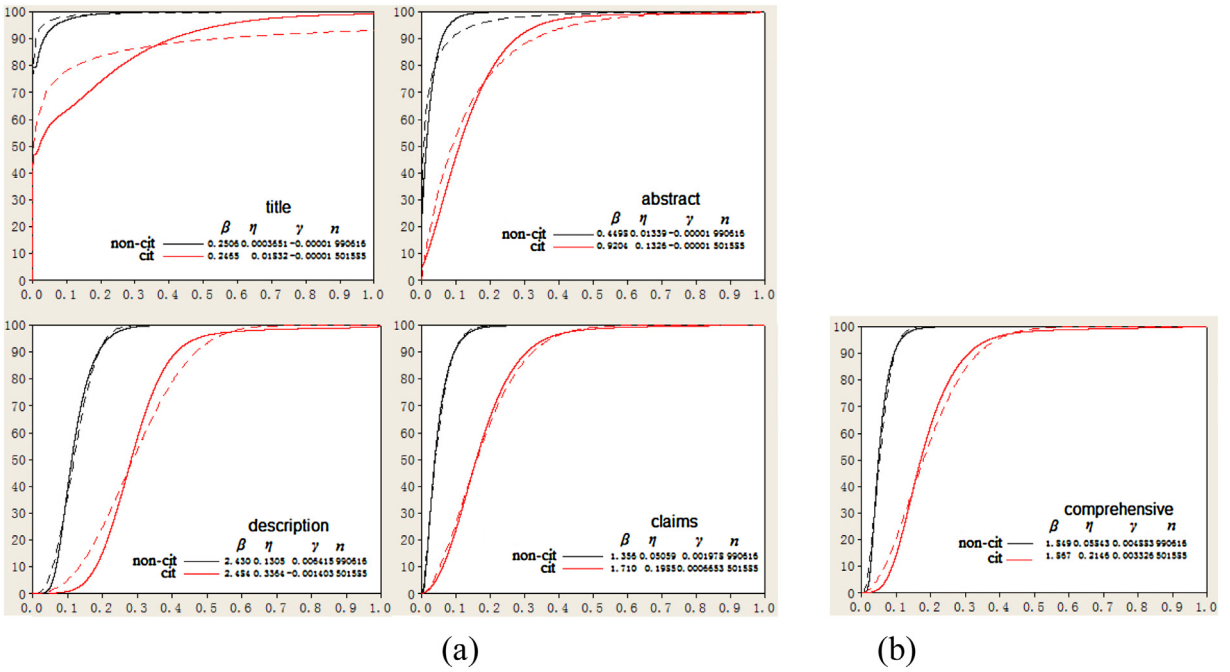

Fig. 3. Cumulative distribution functions of text similarity values of citing-cited and non-citing-cited pairs.

0.1 and are distributed in the relatively higher value zone, whereas for the latter pairs most values are not more than 0.1 and are concentrated in the relatively lower value zone. The results suggest that most citations can indicate knowledge linkage well, but still a small part of citations represents knowledge linkage incompletely (for citing-cited pairs 14% of the values are lower than 0.1).

From Figs. 1–2, it can be seen that the distribution curves of text similarity values are like Weibull. The 3-parameter Weibull probability density function is given by:

$$f(t) = \frac{\beta}{\eta}\left(\frac{t-\gamma}{\eta}\right)^{\beta-1} e^{-\left(\frac{t-\gamma}{\eta}\right)^{\beta}} \tag{6}$$

in which, β is shape parameter, η is scale parameter and γ is location parameter. By using the software Minitab, Weibull cumulative distribution functions of the text similarity values were fitted (see Fig. 3). It can be seen that the five curves of non-citing-cited pairs are located at the left side and are steeper than those of citing-cited pairs. That means there is a much higher probability that the similarity values of citing-cited pairs are higher than those of non-citing-cited pairs.

4.3. Differences between applicant and examiner citations in text similarities

4.3.1. Comparative analysis of text similarities between applicant and examiner citing-cited pairs

The above results show that a small part of citations represents knowledge linkage incompletely. Some researchers (Jaffe et al., 2000) believe that examiner citations carry an amount of "noise" and cannot indicate knowledge linkage, but some others (Li et al., 2014) suggest that logically they can because the citing behavior of examiner is regulated by patent laws. Whether examiner citations can indicate knowledge linkage or not is the focus of the disputation. Therefore, it is necessary to investigate whose text similarity values are higher between applicant and examiner citing-cited pairs.

In the observation group, 23,505 patents contain 501,585 citations, which consist of 407,207 applicant citations and 94,378 examiner citations. Examiner citations account for 19% (51% on the average patent), whereas applicant citations account for 81% (49% on the average patent) of the all citations.[5] That means the majority of patent citations are listed by applicants, and the minority are added by examiners. The distributions of the comprehensive similarity values of applicant

[5] However, using data on the citing patents granted between 2001 and 2003, Alcacer et al. (2009) find that examiner citations account for 41% (63% on the average patent). Using 1% sample of the patents issued in 2007, Cotropia et al. (2013) show that 34% of citations to USPTO patents are from examiners. This reflects that the proportion of citations added by examiners has been decreasing dramatically, and accordingly the proportion of citations added by applicants has been increasing observably. In recent decades, information retrieval systems and web search engines have been rapidly developed and extensively applied, so that applicants can obtain information they need comprehensively and conveniently, and they can submit more and more references. Besides, Sampat (2010) and Cotropia (2009) point out that submitting too many references may be beneficial to the applicant since the examiner may hardly evaluate them efficiently and effectively; if granted, the patent would earn a presumption of validity as against the cited art in litigation.

Table 5
Comprehensive similarity values of examiner and applicant citing-cited pairs.

value	exa-cit	(%)	app-cit	(%)
1	0	0.00	0	0.00
(0.9, 1)	755	0.80	632	0.16
(0.8, 0.9]	573	0.61	1010	0.25
(0.7, 0.8]	392	0.42	885	0.22
(0.6, 0.7]	382	0.40	1026	0.25
(0.5, 0.6]	782	0.83	2050	0.50
(0.4, 0.5]	2727	2.89	6454	1.58
(0.3, 0.4]	11,560	12.25	28,300	6.95
(0.2, 0.3]	32,130	34.04	98,179	24.11
(0.1, 0.2]	39,725	42.09	204,421	50.20
(0, 0.1]	5352	5.67	64,250	15.78
0	0	0.00	0	0.00

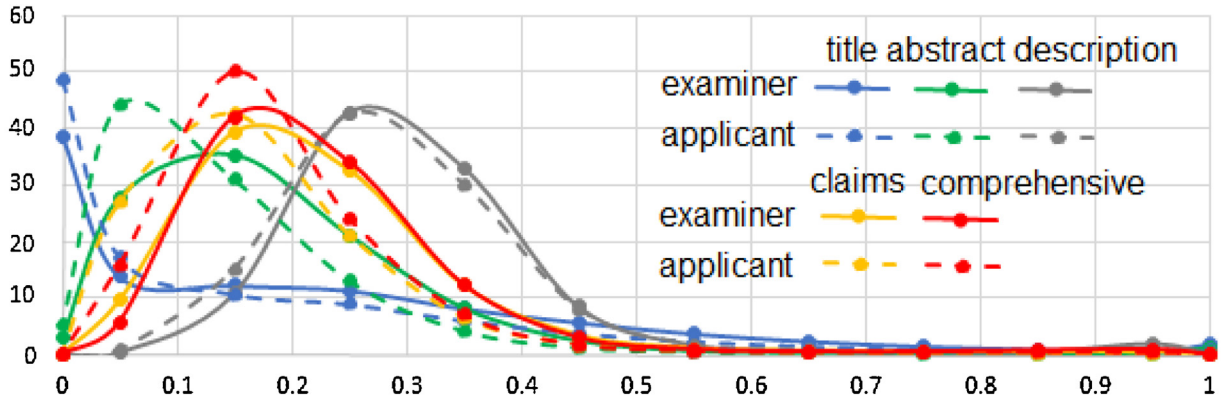

Fig. 4. Distributions of text similarity values of examiner and applicant citing-cited pairs.

and examiner citing-cited pairs are listed in Table 5 (the distributions of the similarity values of the four components are listed in Supplemental Tables 1–2).

From Fig. 4, it can be seen that the five curves of examiner citing-cited pairs are located at the right side of those of applicant citing-cited pairs. It indicates that text similarity values of the former pairs are distributed more in the relatively higher value zone, whereas those of the latter pairs are distributed more in the relatively lower value zone. From Fig. 5, it can be seen that the five curves of the latter pairs are located at the left side and are steeper than those of the former pairs. That means there is a higher probability that the similarity values of the former pairs are higher than those of the latter pairs. It can be inferred that examiner citations may indicate knowledge linkage a bit better than applicant citations.

4.3.2. Comparative analysis of text similarities between applicant and examiner citing-cited pairs within each patent

The local and the global analyses of text similarity values between examiner and applicant citing-cited pairs are macroscopically performed at the level of the whole patents and their citations, and are not microcosmically performed within each patent to investigate which one of examiner and applicant citations indicate knowledge linkage better. In the observed group, 23,505 patents contain 501,585 citations, including 4002 patents with all citations listed by applicants and 7248 patents with all citations added by examiners. There are 12,255 patents with 368,574 citations added by both applicants and examiners; among them, 310,549 citations are added by applicants, accounting for 84%, and 58,025 citations are added by examiners, accounting for 16%. We took the 12,255 patents and their 368,574 citations as the study object to compare text similarities between applicant and examiner citing-cited pairs within each patent.

Assume that a patent contains m applicant citations and n examiner citations. We can calculate the average text similarity value of n examiner citing-cited pairs and the average value of m applicant citing-cited pairs, and can calculate the difference of the two average values. Table 6 shows that the number of D_+ is greater than that of D_-. Especially for patent claims, D_+ is 1.6 times as much as D_-; the overall balance is 22%: 61% of the cases are favorable to that the claims of the citing patent are more relevant to the claims of cited patents added by examiners than listed by applicants, whereas 39% of the cases are unfavorable. That means there is a higher probability (at global level 56% of the cases are favorable and 44% unfavorable) that the average similarity value of examiner citing-cited pairs is higher than that of applicant citing-cited pairs within a patent. The results suggest that examiner citations may indicate knowledge linkage a bit better than applicant citations, especially for the component of claims, and that compared to applicant citations, more than half of the cases support that examiner citations are a good indicator of knowledge linkage rather than an incomplete and noisy indicator.

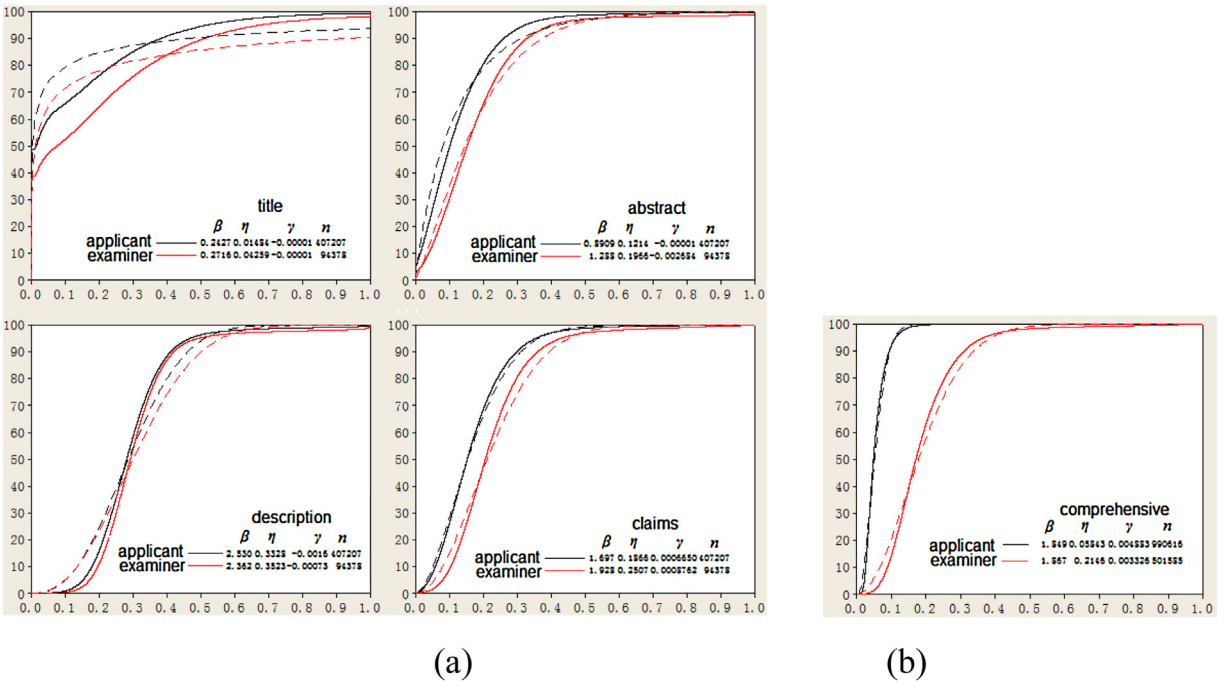

Fig. 5. Cumulative distribution functions of examiner and applicant citing-cited pairs.

Table 6
Comparisons of average similarity values between examiner and applicant citing-cited pairs within each patent.

	title		abstract		description		claims		comprehensive	
	Num	%	Num	%	Num	%	Num	%	Num	%
D_0	408	3.33	6	0.05	0	0.00	0	0.00	0	0.00
D_+	5984	48.83	6798	55.47	6738	54.98	7474	60.99	6908	56.37
D_-	5863	47.84	5451	44.48	5517	45.02	4781	39.01	5347	43.63
DD	121	0.99	1347	10.99	1221	9.96	2693	21.97	1561	12.74

Note: D_0 represents that the difference is equal to zero, D_+ represents that the difference is greater than zero, D_- represents that the difference is less than zero, and *DD* represents D_+ minus D_-.

5. Case study in the field of nano-technology

We retrieved nano-technology patents (USPC Class 977)[6] in the origin database and obtained thousands of patents. We took 1123 utility patents granted in 2013 and 425 utility patents granted in 2008 as the sample. The 1123 patents contain 15,050 citations that are composed of the utility patents granted by USPTO in the period of 1976–2013. These 15,050 citing-cited patent pairs are regarded as the observation group. A total of 477,080 non-citing-cited patent pairs (one patent of a pair is from 2013 and the other is from 2008), are regarded as the control group. By comparing text similarity values between the two groups, we continue to verify whether patent citations indicate knowledge linkage or not in the same technological area.

By using the same methods, the text similarity values of the four components of the two groups were calculated (see Supplemental Tables 3–4). For citing-cited pairs, 34% of the values of titles and 48% of the values of abstracts are greater than 0.1, whereas for non-citing-cited pairs the percentages are respectively 11% and 14%. For the former pairs, 87% of the values of descriptions are greater than 0.2 and 45% of the values are greater than 0.3, whereas for the latter pairs the proportions are respectively 44% and 5%. For the former pairs, all the values of claims are greater than 0.1 and 68% of the values are greater than 0.2, whereas for the latter pairs the proportions are respectively 28% and 4%. By using Eq. (5), the comprehensive text similarity values (seen in Table 7) were calculated. For citing-cited pairs 16% of the values are lower than 0.1 (84% of the values are greater than 0.1), whereas for non-citing-cited pairs, 63% of the values are lower than 0.1 (37% of the values are greater than 0.1). For the former pairs 32% of the values are greater than 0.2, whereas for the latter pairs the proportion is

[6] Nano-technology (US patent class 977) seems like a single unified field of technology but basically is a "container" consisting of several subfields, and therefore the analysis might give slightly different results for the subfields: The text similarity values of citing-cited pairs might be greater and the differences between citing-cited and non-citing-cited pairs in text similarity values might be narrower.

Table 7
Comprehensive similarity values in the field of nano-technology.

value	cit pairs	(%)	non-cit pairs	(%)
1	0	0.00	0	0.00
(0.9, 1)	43	0.29	1	0.00
(0.8, 0.9]	36	0.24	2	0.00
(0.7, 0.8]	43	0.29	1	0.00
(0.6, 0.7]	33	0.22	1	0.00
(0.5, 0.6]	69	0.46	5	0.00
(0.4, 0.5]	163	1.08	40	0.01
(0.3, 0.4]	949	6.31	1261	0.26
(0.2, 0.3]	3509	23.32	16,238	3.40
(0.1, 0.2]	7790	51.76	159,033	33.33
(0, 0.1]	2415	16.05	300,498	62.99
0	0	0.00	0	0.00

Table 8
Relative scores of mean, median and mean rank of text similarity values of all kinds of pairs.

		non-cit	cit	app	exm	n-cit-nano	cit-nano	app-nano	exm-nano
TI	mean	0.09	1	0.91	1.38	0.24	0.85	0.78	1.25
	mean rank	0.67	1	0.98	1.09	0.74	0.92	0.90	1.02
AB	mean	0.17	1	0.93	1.31	0.36	0.90	0.83	1.34
	median	0.13	1	0.92	1.38	0.28	0.85	0.79	1.30
	mean rank	0.46	1	0.97	1.12	0.64	0.94	0.92	1.08
DE	mean	0.41	1	0.99	1.05	0.65	1.01	0.99	1.15
	median	0.40	1	0.99	1.03	0.67	1.02	1.01	1.10
	mean rank	0.33	1	0.99	1.03	0.68	1.02	1.00	1.09
CL	mean	0.27	1	0.94	1.26	0.45	0.89	0.82	1.27
	median	0.24	1	0.94	1.27	0.42	0.86	0.81	1.24
	mean rank	0.40	1	0.97	1.12	0.62	0.93	0.90	1.10
CO	mean	0.29	1	0.96	1.19	0.50	0.94	0.89	1.23
	median	0.29	1	0.96	1.19	0.50	0.93	0.90	1.20
	mean rank	0.34	1	0.98	1.09	0.64	0.97	0.95	1.09

Note: The mean, median and mean rank of text similarity values of citing-cited pairs are regarded as the base scores. The relative scores are the corresponding values divided by the base scores. For title–title pairs, relative scores of median values are not listed because some median values are zero.

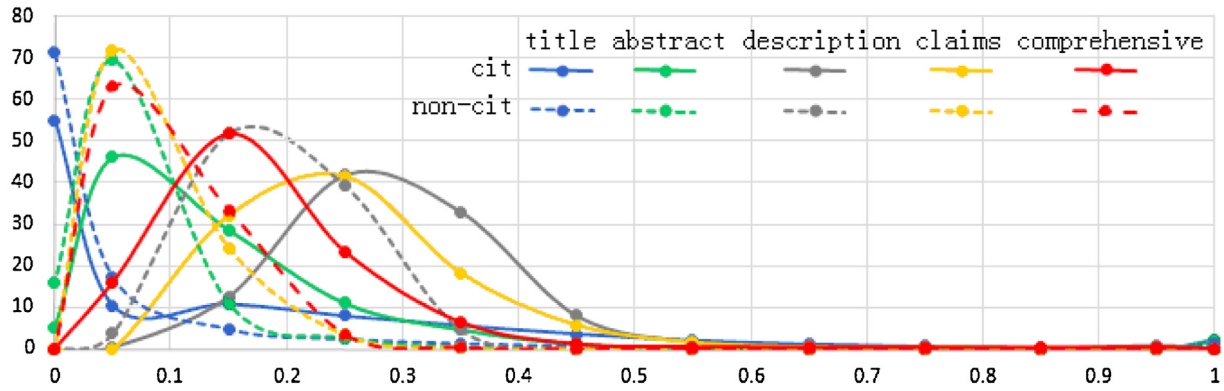

Fig. 6. Distributions of text similarity values in the field of nano-technology.

4%. It indicates that there is a much higher probability that text similarity values of citing-cited pairs are higher than those of non-citing-cited pairs in the field of nano-technology.

Figs. 6 and 7 show that the curves of citing-cited pairs are located at the right side of those of non-citing-cited pairs, and Fig. 7 shows that the five curves of non-citing-cited pairs are steeper than those of citing-cited pairs. The results show that, in the field of nano-technology, the text similarity values of citing-cited pairs are distributed more in the relatively higher value zone, whereas those of non-citing-cited pairs are distributed more in the relatively lower value zone.

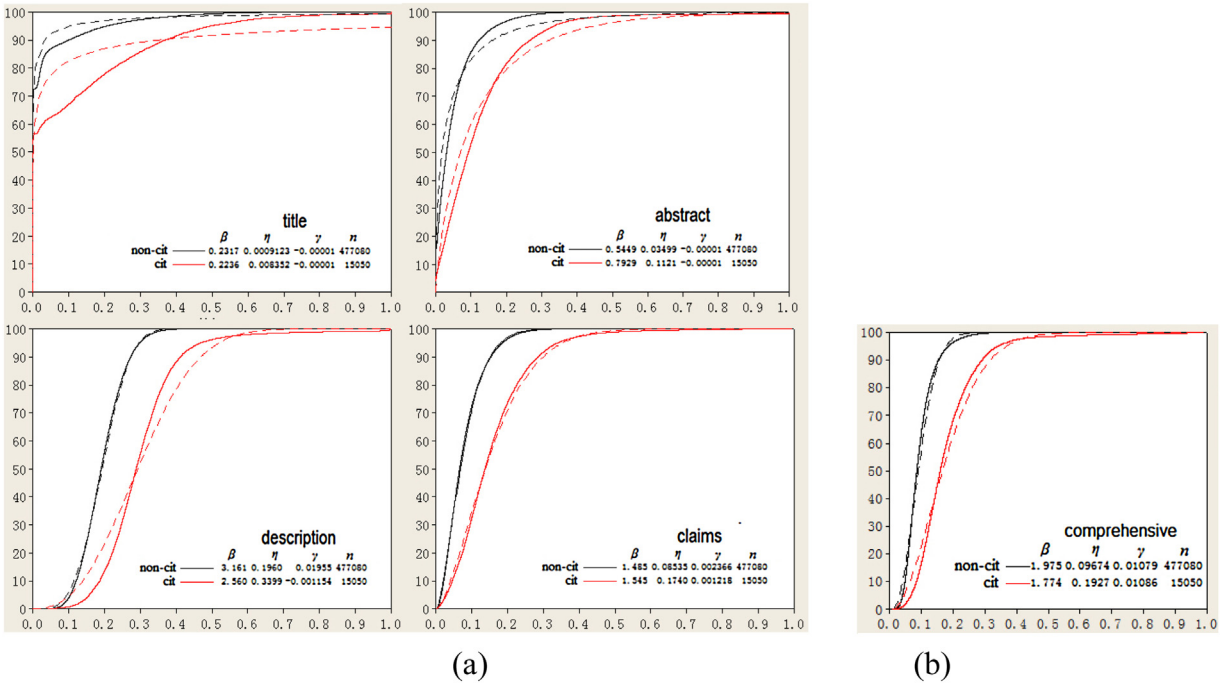

Fig. 7. Cumulative distribution functions in the field of nano-technology.

6. Comprehensive analysis and interpretation

6.1. Comparative analysis of text similarity values among all kinds of patent pairs

In the article, there are eight kinds of patent pairs: non-citing-cited pairs, citing-cited pairs, applicant citing-cited pairs, examiner citing-cited pairs and the other four pairs in the field of nano-technology. For comparing the text similarity values among all kinds of pairs, the mean, median and mean rank of these text similarity values were calculated (see Supplemental Table 5[7]) and the relative scores are shown in Table 8. Mean rank is often used to test whether there is a statistically significant difference in scores for many variables measured at the ordinal level.

Fig. 8 shows that the relative scores of the two kinds of citing-cited pairs (lines 2 and 6) are much higher than those of non-citing-cited pairs (lines 1 and 5), and the relative scores of the two kinds of examiner citing-cited pairs (lines 4 and 8) are a bit higher than those of applicant citing-cited pairs (lines 3 and 6). It can be seen that the two lines of non-citing-cited pairs (lines 1 and 5) are similar, the two lines of citing-cited pairs (line 2 and 6) are similar, the two lines of applicant citing-cited pairs (lines 3 and 7) are similar, and the two lines of examiner citing-cited pairs (lines 4 and 8) are similar, too. The gaps among the kinds of citing-cited pairs (lines 2–4, 6–8) are narrower, whereas the gaps from the kinds of citing-cited pairs (lines 2–4, 6–8) to the kinds of non-citing-cited pairs (lines 1 and 5) are wider. The two lines of applicant and examiner citing-cited pairs (lines 3 and 4) are complementary; the higher line 3 is, the lower line 4 is. The two lines of applicant and examiner citing-cited pairs in the field of nano-technology (lines 7 and 8) are complementary, too. The relative scores of the two kinds of examiner citing-cited pairs (lines 4 and 8) are the highest, whereas those of non-citing-cited pairs (lines 1 and 5) are the lowest. The relative scores of the kind of non-citing-cited pairs in the same technological area (nano-technology, line 5) are higher than those in different technological areas (line 1), but are much lower than scores of the kinds of citing-cited pairs (lines 2–4 and 6–8).

The results show that for the six kinds of citing-cited pairs, including applicant and examiner citing-cited pairs, whether from the same technological area or not, their scores are not much different. Whereas for the two kinds of non-citing-cited pairs, there are obvious differences: the scores of the kind of pairs from the same technological area (nano-technology) are

[7] Besides, from Supplemental Table 5, it can be seen that the mean and median values become greater when the text is longer in the order of title, abstract, claims and description, especially for non-citing-cited pairs. Many researchers (Metzler et al., 2007; Oliva et al., 2011) suggest that using short texts to measure text similarities performs poorly and has more deviations because of term sparseness and lack of context, especially without considering semantics of text. For citing-cited pairs, their standard deviations of text similarity values of the four components and the comprehensive are 0.1994, 0.1273, 0.1195, 0.1091 and 0.1092 respectively. The deviation of values of claims is the minimum; it means that the text similarity values of claims from citing-cited pairs are more stable than the others. From Fig. 8, it can be seen that patent description is an insensitive indicator to measure text similarity, in that its gaps from the lines of citing-cited pairs (lines 2–4 and 6–8) to the lines of non-citing-cited pairs (lines 1 and 5) are narrower than the others.

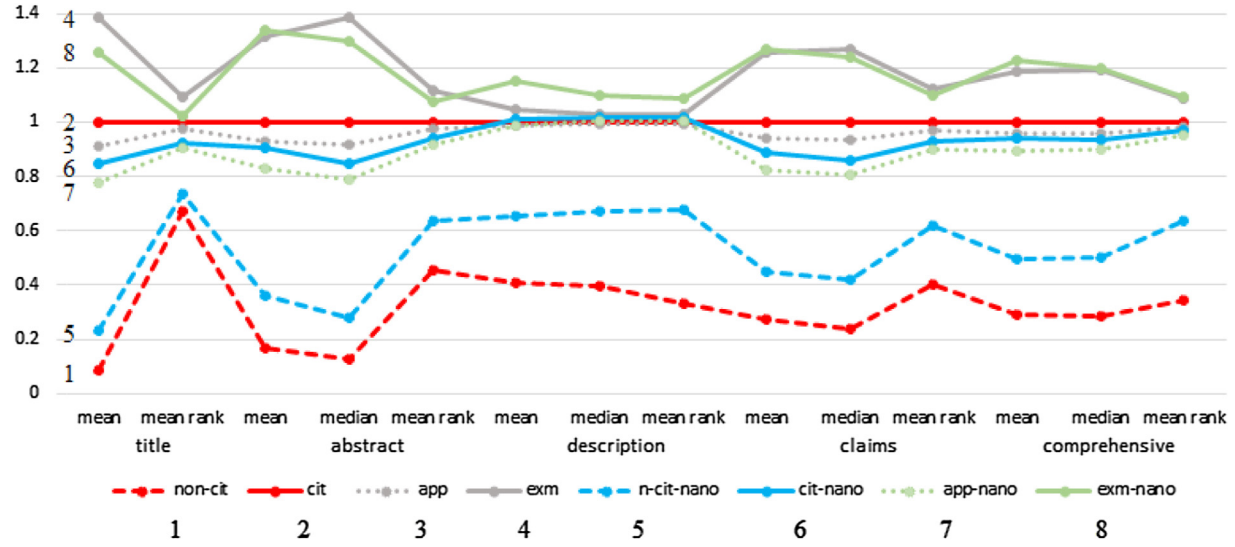

Fig. 8. Relative scores of mean, median and mean rank of text similarity values.

much higher than those from different technological areas. It stands to reason that patents from the same technological area are more relevant than from different technological areas, but there is still a much higher probability that the text similarity values of citing-cited pairs are higher than those of non-citing-cited pairs in the field of nano-technology.

6.2. *Effect sizes of difference tests of text similarity values*

For further analyzing differences of text similarity values of different kinds of patent pairs, by using the software Minitab, the differences of text similarity values between citing-cited and non-citing-cited pairs, and between examiner and applicant citing-cited pairs were tested. The Mann-Whitney (Mann & Whitney, 1947) tests indicate that it is statistically significant ($p = 0.0000$, adjusted for ties) that the text similarity values of citing-cited pairs are higher than those of non-citing-cited pairs, and that the text similarity values of examiner citing-cited pairs are higher than those of applicant citing-cited pairs, even in the field of nano-technology.

However, hypothesis testing (significance testing) has serious limitations and flaws (Cohen, 1962, 1990). The results of significance test are easily and commonly misinterpreted, because the test is biased by the sample size (a large sample size can lead to some comparisons being significant) (Cohen, 1994). Hypothesis testing produces arbitrarily dichotomous yes-no answers (accepting or rejecting the null hypothesis), while discarding important information about magnitude favorable or unfavorable to the null hypothesis (Ellis, 2010).

In order to determine the practical importance of difference tests, we need to measure effect size (Huberty, 2002; Kelley & Preacher, 2012), which complements hypothesis testing and is helpful for making decision (it is the first criterion for evaluating a statistical claim). Cureton (1956) introduced the rank-biserial correlation as an effect size for the Mann-Whitney U test. It is a simple measuring method related to the common language effect size (McGraw & Wong, 1992; Vargha & Delaney, 2000), and has been improved and generalized by many scholars (Cliff, 1993; Kerby, 2014; Newcombe, 2006; Wendt, 1972).

Kerby (2014) gives a simple difference formula, the rank-biserial correlation $r = f - u$, in which f is the proportion of pairs favorable to the hypothesis and u is that not favorable. The correlation r is directional, with positive values indicating that the results support the hypothesis, and does not require any assumptions about the shape or spread of the two distributions. It can be calculated from U (U_1 or U_2) or W (W_1 or W_2), and from the sizes (N_1 and N_2) of the two groups, with the Mann-Whitney U test:

$$r = (U_1 - U_2)/(N_1 \times N_2) \tag{7}$$

$$U_1 + U_2 = N_1 \times N_2 \tag{8}$$

$$U_1 = W_1 - N_1(N_1 + 1)/2 \tag{9}$$

in which, U is statistic and W is the rank sum of a group. Referencing to the guidelines of Pearson's r given by Cohen (1992), we can regard the effect size of a Mann-Whitney U test as tiny, small, medium, large and huge when rank-biserial correlation r is from [0,0.1), [0.1, 0.3), [0.3, 0.5), [0.5, 0.7), [0.7, 1] respectively.

Table 9 shows the effect sizes of differences of text similarity values. The effect sizes of difference $V_C - V_N$ are huge, except that the effect size of patent titles is medium. Among the four components, patent title is the shortest text; using short texts to measure text similarities performs poorly (Metzler, Dumais, & Meek, 2007; Oliva, Serrano, del Castillo, & Iglesias, 2011).

Table 9
Effect sizes of differences of text similarity values.

ii	N_1	N_2	R_{TI}	R_{AB}	R_{DE}	R_{CL}	R_{CO}
$V_C - V_N$	501,585	990,616	0.40	0.74	0.91	0.83	0.91
$V_E - V_N$	94,378	990,616	0.50	0.85	0.93	0.93	0.96
$V_A - V_N$	407,207	990,616	0.37	0.72	0.91	0.81	0.90
$V_E - V_A$	94,378	407,207	0.15	0.26	0.08	0.32	0.26
$V_{CN} - V_{NN}$	15,050	477,080	0.22	0.45	0.66	0.52	0.78
$V_{EN} - V_{NN}$	2145	477,080	0.34	0.64	0.78	0.77	0.82
$V_{AN} - V_{NN}$	12,905	477,080	0.20	0.42	0.64	0.48	0.61
$V_{EN} - V_{AN}$	2145	12,905	0.15	0.28	0.20	0.39	0.34

Note: The table shows the effect sizes of differences of text similarity values of the four components and the comprehensive, between citing-cited and non-citing-cited pairs, between examiner citing-cited and non-citing-cited pairs, between applicant citing-cited and non-citing-cited pairs, between examiner and applicant citing-cited pairs, and those in the field of nano-technology. For the first test of $V_C - V_N$, the hypothesis is the difference $V_C - V_N$ is greater than zero. For patent titles, of the total of 501,585 × 990,616 = 496,878,126,360 comparison pairs, U_1 = 346,976,826,084 (70%) are favorable to the hypothesis and U_2 = 149,901,300,276 (30%) are unfavorable; the favorable evidence outweighs the unfavorable 70%–30%, so the overall balance is 0.70 minus 0.30, yielding a rank-biserial correlation $r = f - u = 0.70–0.30 = 0.40$.

For the comprehensive values, the overall balance is 0.91, which means 95.5% are favorable to the hypothesis that the text similarity values of citing-cited pairs are higher than those of non-citing-cited pairs, and only 4.5% are unfavorable.

But the effect sizes of difference $V_E - V_A$ are not large. For the comprehensive values, the overall balance is 0.26 (63% are favorable and 37% unfavorable) and the effect size is small. For the values of patent descriptions, the overall balance is only 0.08 (54% are favorable and 46% unfavorable) and the effect size is tiny; hence, it is hard to accept the hypothesis that the text similarity values of examiner citing-cited pairs are greater than those of applicant citing-cited pairs only based on patent descriptions. For patent titles, abstracts and claims, the effect sizes are small, small and medium respectively. In the field of nano-technology, the effect sizes of difference $V_{CN} - V_{NN}$ are much lower than those of $V_C - V_N$, whereas the effect sizes of $V_{EN} - V_{AN}$ are a bit bigger than those of $V_E - V_A$. The effect sizes of $V_E - V_N$ ($V_{EN} - V_{NN}$) are a bit bigger than those of $V_C - V_N$ ($V_{CN} - V_{NN}$), whereas the effect sizes of $V_A - V_N$ ($V_{AN} - V_{NN}$) are a bit smaller than $V_C - V_N$ ($V_{EN} - V_{NN}$).

In brief, for the comprehensive values, 95.5% (89% in the field of nano-technology) of the cases are in accord with the hypothesis that $V_C - V_N > 0$ ($V_{CN} - V_{NN} > 0$), and two-thirds (63% and 67%) of the cases are in accord with the hypotheses that $V_E - V_A > 0$ and $V_{EN} - V_{AN} > 0$. For the differences of $V_E - V_A$ and $V_{EN} - V_{AN}$, the effect sizes from patent claims are bigger than from the other three components; it shows that there is more (66% and 69.5%) of the cases to support that patent claims are much more relevant to claims of cited patents added by examiners than listed by applicants. It can be suggested that in practice most certainly (95% are favorable) patent citations can indicate knowledge linkage, and examiner citations more likely (63% are favorable) indicate knowledge linkage a bit better than applicant citations, especially for the component of claims (66% are favorable).

6.3. Interpretation of the results based on patent examination processes

Patent citations include two types: applicant citations and examiner citations. They involve two kinds of knowledge linkage: knowledge diffusion and knowledge relevance. Knowledge diffusion means adaptations and applications of knowledge, while knowledge relevance is approximated by knowledge similarity measure based on word, term and text, etc. For knowledge relevance, the results show that most certainly both applicant citations and examiner citations can indicate knowledge relevance. As for knowledge diffusion, some researchers (Jaffe et al., 2000) suggest that only applicant citations can indicate knowledge diffusion, because inventors are unaware of examiner citations and do not utilize them for creating inventions.

However, it is just one side of the issue, not the whole. On the other side, applicants may disclose prior art incompletely and imperfectly, and even deliberately withhold some closely related prior art that invalidates their claims (Cotropia, Lemley, & Sampat, 2013; Lampe, 2012). In contrast, the examiners must reference the related prior patents or publications for judging the application's patentability according to patent laws. Of course, not every invention is novel to qualify for a patent. The main requirement for obtaining a patent on an invention is that it must be novel: it must differ from all previous inventions or existing knowledge. The nature of patent examination is rejection of claims.[8] Basing their study on the patent data and information from USPTO, Cotropia et al. (2013) point out that about 36% of examiner citations were used to reject the claims of patent applications and 76% of the granted patents had at least one rejection of claims based on prior art. Applicants

[8] Patent claims, which state the novelties in an invention and declare the exclusive ownership rights, are the core of a patent. The examiners' duty is to review and determine whether an invention should be granted a patent by comparing its claims with the prior art to judge whether it complies with the novelty requirement. According to USPTO, rejection of claims (706) is explained as: The refusal to grant claims because the subject matter as claimed is considered unpatentable is called a "rejection." (1) If the invention is not considered patentable, or not considered patentable as claimed, the claims, or those considered unpatentable will be rejected. (2) In rejecting claims for want of novelty or for obviousness, the examiner must cite the best references at his or her command. When a reference is complex or shows or describes inventions other than that claimed by the applicant, the particular part relied on must be designated as nearly as practicable. The pertinence of each reference, if not apparent, must be clearly explained and each rejected claim specified. For more details about rejection of claims, see http://www.uspto.gov/web/offices/pac/mpep/s706.html

may prefer to take patents that support rather than obviously undermine or impinge on their claims as references (Hegde & Sampat, 2009; Lampe, 2012). They may unintentionally ignore or intentionally withhold the prior art that negates or harms the novelty of patent applications, and they may overwhelmingly submit low quality or irrelevant prior art across all areas because they cannot retrieve the relevant art as professionally as examiners can and they are not required by law to search the prior art or outsource search to professionals (Cotropia et al., 2013; Hegde & Sampat, 2009; Lampe, 2012). Examiners rarely use applicant citations in their rejections to limit or narrow patent claims, relying almost exclusively on examiner citations (Cotropia et al., 2013). They must professionally search the most relevant prior art and must cite the best references for judging the novelty of patent applications. In addition, even if examiner citations are unknown to the inventor and are not utilized for developing the invention, it just means that examiner citations do not directly indicate knowledge flow and diffusion. Nonetheless, examiner citations may indirectly indicate knowledge flow and diffusion through affecting applicants and forcing them to amend their claims. During patent examination process, the examiner and the applicant exchange information. The examiner replies to the applicant in the office actions, in which examiner citations are listed when rejection is based on prior art. The applicant can abandon, narrow or amend some of the patent claims for avoiding prior art, or can give a reasonable explanation to the examiner for the novelty of the invention; otherwise, the applicant will not obtain a patent (but he/she can appeal). That means examiner citations may indicate knowledge diffusion indirectly, because they are mainly used to judge patentability and compel the applicant to amend the patent claims for avoiding the prior art. This is the main cause of the results of the article: Examiner citations rather than applicant citations are more likely to relate to the citing patent closely; compared to the other three components, claims of cited patents added by examiners rather than listed by applicants are more liable to associate with claims of the citing patent in text similarity.

By analyzing patent examination process, we interpret the cause of the fact that examiner citations are more likely to indicate knowledge linkage a bit better than applicant citations, especially for the component of patent claims. Preferably, examiner citations can be regarded as not only the supplement of applicant citations but also the indispensable technological background and the prior art related to the patents. We suggest that logically examiner citations can indicate knowledge linkage, in that examiner citations are mainly used for patent rejections that influence applicants and oblige them to abandon, limit or amend some of their claims for obtaining patents.

7. Conclusions and discussions

By comparing text similarity values between the two groups, it can be validated that, in the vast majority of cases, text similarity values of citing-cited pairs are much higher than those of non-citing-cited pairs. The study in the field of nanotechnology shows that, in the majority of cases, the results are the same although patents in the same technological area are more relevant than in different technological areas. Furthermore, by comparing text similarities between applicant and examiner citing-cited pairs, the results show that, in more cases, text similarity values of examiner citing-cited pairs are a bit higher than those of applicant citing-cited pairs. Preferably, examiner citations can be regarded as not only the supplement of applicant citations but also the more important technological background and the prior art closely related to the patents. Compared to applicant citations, logically examiner citations are a good indicator of knowledge linkage rather than an incomplete and noisy indicator. In short, the results suggest that almost certainly patent citations can indicate knowledge linkage, and more likely examiner citations can indicate knowledge linkage a bit better than applicant citations, especially for the component of patent claims. Therefore, we accept the hypothesis that patent citations can indicate knowledge linkage, although a small part of citations represents knowledge linkage incompletely.

In recent several decades, patent citations have been widely used as a very important proxy for exploring linkage between technology and science, for measuring technological knowledge diffusion and relevance, for revealing technological development tracking and even for evaluating the importance, impact or quality of patents. No doubt, the patent citation analysis method must be provided with the assumption that patent citations can indicate knowledge linkage. The assumption is a foundational and basic issue of the patent citation analysis method. By comparing text similarities between citing-cited and non-citing-cited pairs, the study provides good empirical evidence supporting the assumption of knowledge relevance: patents are generally more relevant to their citations (cited patents) than other patents in text similarities, even in the same technological area. By analyzing patent examination process and by interpreting the cause of the fact that in more cases text similarity values of examiner citing-cited pairs are a bit higher than those of applicant citing-cited pairs, the study provides indirect (logical) evidence supporting the assumption of knowledge diffusion. Applicant citations can indicate knowledge diffusion directly, because the applicants are required by law for disclosing prior art as the pre-existing background technological knowledge known by the inventors and utilized to develop the new inventions. Examiner citations can indicate knowledge diffusion indirectly, because the examiners are required for searching the closely related prior art as the references mainly used for patent rejections, which force the applicants to abandon, or narrow, or amend patent claims for obtaining patents. In short, we suggested that patent citations, including applicant citations and examiner citations, can indicate knowledge linkage, containing knowledge relevance and logically involving knowledge diffusion. We preliminarily and tentatively validate the basic assumption of the patent citation analysis method: patent citations can indicate knowledge linkage.

In the study, we do not consider semantics when measuring text similarity, so the knowledge linage of two texts with the same meaning but which are different in words cannot be inspected by the text similarity measure only based on the VSM. In the patent text mining process, some researchers (Shih & Liu, 2010; Taduri, Lau, Law, Yu, & Kesan, 2011) construct knowledge

ontology or use semantic analysis based on WordNet,[9] in order to increase the accuracy of text similarity calculating. In further researches, we will consider semantics, especially for the short-text of patent title and abstract. Through integrating citation analysis with content analysis (Ding et al., 2014; Zhang, Ding, & Milojević, 2013), we will reveal how a patent citation links to the citing patent and how patent citations indicate knowledge linkage specifically, especially for examiner citations across different areas of technology or over time, in order to gather new detailed evidence for the hypothesis.

Acknowledgements

I thank Professor Liming Liang (Henan Normal University) and Professor Zeyuan Liu (Dalian University of Technology) for guidance, thank my colleagues, Professor Qi Wang, Qianjin Hong, Guangjun Li, Jiyu Liu and Tao Chen for their help, and thank two referees for painstaking comments. This study is supported by the National Natural Science Foundation of China (Grant No. 71373252).

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at http://dx.doi.org/10.1016/j.joi.2016.04.018.

References

Ahlgren, P., & Colliander, C. (2009). Document-document similarity approaches and science mapping: Experimental comparison of five approaches. *Journal of Informetrics*, *3*(1), 49–63. http://dx.doi.org/10.1016/j.joi.2008.11.003
Albert, M., Avery, D., Narin, F., & McAllister, P. (1991). Direct validation of citation counts as indicators of industrially important patents. *Research Policy*, *20*(3), 251–259. http://dx.doi.org/10.1016/0048-7333(91)90055-u
Alcacer, J., & Gittelman, M. (2006). Patent citations as a measure of knowledge flows: The influence of examiner citations. *The Review of Economics and Statistics*, *88*, 774–779. http://dx.doi.org/10.1162/rest.88.4.774
Alcacer, J., Gittelman, M., & Sampat, B. (2009). Applicant and examiner citations in US patents: An overview and analysis. *Research Policy*, *38*, 415–427. http://dx.doi.org/10.2139/ssrn.1273016
Callaert, J., Grouwels, J., & Looy, B. V. (2012). Delineating the scientific footprint in technology: Identifying scientific publications within non-patent references. *Scientometrics*, *91*(2), 383–398. http://dx.doi.org/10.1007/s11192-011-0573-9
Carpenter, M. P., & Narin, F. (1983). Validation study: Patent citations as indicators of science and foreign dependence. *World Patent Information*, *5*(3), 180–185. http://dx.doi.org/10.1016/0172-2190(83)90139-4
Carpenter, M. P., Cooper, M., & Narin, F. (1980). Linkage between basic research literature and patents. *Research Management*, *23*(2), 30–35.
Cascini, G., & Zini, M. (2008). Measuring patent similarity by comparing inventions functional trees. In G. Cascini (Ed.), *International federation for information processing: Computer-aided innovation* (pp. 31–42). http://dx.doi.org/10.1007/978-0-387-09697-1_3
Chen, C. M., & Hicks, D. (2004). Tracing knowledge diffusion. *Scientometrics*, *59*(2), 199–211. http://dx.doi.org/10.1023/B:SCIE.0000018528.59913.48
Cliff, N. (1993). Dominance statistics: Ordinal analyses to answer ordinal questions. *Psychological Bulletin*, *114*(3), 494. http://dx.doi.org/10.1037/0033-2909.114.3.494
Cohen, J. (1962). The statistical power of abnormal-social psychological research. *Journal of Abnormal Psychology*, *65*(3), 145–153. http://dx.doi.org/10.1037/h0045186
Cohen, J. (1990). What I have learned (so far). *American Psychologist*, *45*(12), 1304. http://dx.doi.org/10.1037/0003-066x.45.12.1304
Cohen, J. (1992). A power primer. *Psychological Bulletin*, *112*(1), 155–159. http://dx.doi.org/10.1037/0033-2909.112.1.155
Cohen, J. (1994). The earth is round ($p < .05$). *American Psychologist*, *49*, 997–1003. http://dx.doi.org/10.1037/0003-066X.49.12.997
Cotropia, C. A., Lemley, M. A., & Sampat, B. N. (2013). Do applicant patent citations matter? *Research Policy*, *42*(4), 844–854. http://dx.doi.org/10.2139/ssrn.1656568
Cotropia, C. A. (2009). Modernizing the inequitable conduct doctrine in patent law. *Berkeley Technology Law Journal*, *24*, 723–783.
Criscuoloa, P., & Verspagen, B. (2008). Does it matter where patent citations come from? Inventor vs. examiner citations in European patents. *Research Policy*, *37*(9), 1892–1908. http://dx.doi.org/10.1016/j.respol.2008.07.011
Cureton, E. E. (1956). Rank-biserial correlation. *Psychometrika*, *21*(3), 287–290. http://dx.doi.org/10.1007/bf02289138
Ding, Y., Zhang, G., Chambers, T., Song, M., Wang, X., & Zhai, C. (2014). Content-based citation analysis: The next generation of citation analysis. *Journal of the Association for Information Science and Technology*, *65*(9), 1820–1833. http://dx.doi.org/10.1002/asi.23256
Egghe, L., & Leydesdorff, L. (2009). The relation between Pearson's correlation coefficient r and Salton's cosine measure. *Journal of the American Society for Information Science and Technology*, *60*(5), 1027–1036. http://dx.doi.org/10.1002/asi.21009
Ellis, P. D. (2010). *The essential guide to effect sizes: Statistical power, meta-Analysis, and the interpretation of research results*. Cambridge: Cambridge University Press.
Engelsman, E. C., & van Raan, T. (1994). A patent-based cartography of technology. *Research Policy*, *23*(1), 1–26. http://dx.doi.org/10.1016/0048-7333(94)90024-8
Epicoco, M. (2013). Knowledge patterns and sources of leadership: Mapping the semiconductor miniaturization trajectory. *Research Policy*, *42*(1), 180–195. http://dx.doi.org/10.1016/j.respol.2012.06.009
Érdi, P., Makovi, K., Somogyvári, Z., Strandburg, K., Tobochnik, J., Volf, P., et al., (2013). Prediction of emerging technologies based on analysis of the US patent citation network, Scientometrics, 95(1), 225–242. doi: 10.1007/s11192-012-0796-4.
Fattori, M., Pedrazzi, G., & Turra, R. (2003). Text mining applied to patent mapping: A practical business case. *World Patent Information*, *25*(4), 335–342. http://dx.doi.org/10.1016/S0172-2190(03)00113-3
Feldman, R., & Sanger, J. (2007). *The text mining handbook: Advanced approaches in analyzing unstructured data*. pp. 1–18. London: Cambridge University Press.
Feng, X., & Leng, F. H. (2012). Patent text mining and informetric-based patent technology morphological analysis: An empirical study. *Technology Analysis & Strategic Management*, *24*(5), 467–479. http://dx.doi.org/10.1080/09537325.2012.674669
Garfield, E., Sher, I. H., & Torpie, R. J. (1964). *The use of citation data in writing the history of science*. Philadelphia: Institute for Scientific Information.

[9] WordNet is a large lexical database of English, and was created in the Cognitive Science Laboratory of Princeton University. For more details about WordNet, see http://wordnet.princeton.edu/wordnet/.

Hegde, D., & Sampat, B. (2009). Examiner citations, applicant citations, and the private value of patents. *Economics Letters*, *105*(3), 287–289. http://dx.doi.org/10.1016/j.econlet.2009.08.019
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. http://dx.doi.org/10.3386/w8528
Hu, D., Chen, H., Huang, Z., & Roco, M. C. (2007). Longitudinal study on patent citations to academic research articles in nanotechnology (1997–2004). *Journal of Nanoparticle Research*, *9*(4), 529–542. http://dx.doi.org/10.1007/s11051-007-9215-9
Huberty, C. J. (2002). A history of effect size indices. *Educational and Psychological Measurement*, *62*(2), 227–240. http://dx.doi.org/10.1177/0013164402062002002
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. http://dx.doi.org/10.1257/aer.90.2.215
Kelley, K., & Preacher, K. J. (2012). On effect size. *Psychological Methods*, *17*(2), 137–152. http://dx.doi.org/10.1037/a0028086
Kerby, D. S. (2014). The simple difference formula: An approach to teaching nonparametric correlation. *Innovative Teaching*, *3*, 1–9. http://dx.doi.org/10.2466/11.IT.3.1
Lai, K. K., & Wu, S. J. (2005). Using the patent co-citation approach to establish a new patent classification system. *Information Processing & Management*, *41*(2), 313–330. http://dx.doi.org/10.1016/j.ipm.2003.11.004
Lampe, R. (2012). Strategic citation. *Review of Economics and Statistics*, *94*(1), 320–333. http://dx.doi.org/10.2139/ssrn.984123
Lanjouw, J. O., & Schankeman, M. (2004). Patent quality and research productivity: Measuring innovation with multiple indicators. *Economic Journal*, *114*(495), 441–465. http://dx.doi.org/10.1111/j.1468-0297.2004.00216.x
Leydesdorff, L., & Rafols, I. (2009). A global map of science based on the ISI subject categories. *Journal of the American Society for Information Science and Technology*, *60*(2), 348–362. http://dx.doi.org/10.1002/asi.20967
Li, R., & Meng, L. (2010). On the framing of patent citations and academic paper citations in reflecting knowledge linkage: A discussion of the discrepancy of their divergent value-orientations. *Chinese Journal of Library and Information Science*, *3*, 37–45.
Li, R., Chambers, T., Ding, Y., Zhang, G., & Meng, L. (2014). Patent citation analysis: Calculating science linkage based on citing motivation. *Journal of the Association for Information Science and Technology*, *65*(5), 1007–1017. http://dx.doi.org/10.1002/asi.23054
Liu, Y., & Rousseau, R. (2010). Knowledge diffusion through publications and citations: A case study using ESI-fields as unit of diffusion. *Journal of the American Society for Information Science and Technology*, *61*(2), 340–351. http://dx.doi.org/10.1002/asi.21248
Lo, S. S. (2010). Scientific linkage of science research and technology development: A case of genetic engineering research. *Scientometrics*, *82*(1), 109–120. http://dx.doi.org/10.1007/s11192-009-0036-8
Magerman, T., van Looy, B., & Song, X. (2010). Exploring the feasibility and accuracy of latent semantic analysis based text mining techniques to detect similarity between patent documents and scientific publications. *Scientometrics*, *82*(2), 289–306. http://dx.doi.org/10.1007/s11192-009-0046-6
Mann, H. B., & Whitney, D. R. (1947). On a test of whether one of two random variables is stochastically larger than the other. *Annals of Mathematical Statistics*, *18*(1), 50–60. http://dx.doi.org/10.2307/2236101
Manning, C. D., Raghavan, P., & Schütze, H. (2008). *Introduction to information retrieval*. pp. 115–116. Cambridge University Press.
Martinelli, A. (2012). An emerging paradigm or just another trajectory? Understanding the nature of technological changes using engineering heuristics in the telecommunications switching industry. *Research Policy*, *41*(2), 414–429. http://dx.doi.org/10.1016/j.respol.2011.10.012
McGraw, K. O., & Wong, S. P. (1992). A common language effect size statistic. *Psychological Bulletin*, *111*(2), 361–365. http://dx.doi.org/10.1037/0033-2909.111.2.361
Metzler, D., Dumais, S., & Meek, C. (2007). Similarity measures for short segments of text. *Advances in Information Retrieval, Lecture Notes in Computer Science*, *4425*, 16–27. http://dx.doi.org/10.1007/978-3-540-71496-5_5
Meyer, M. (2000). What is special about patent citations? Differences between scientific and patent citations. *Scientometrics*, *49*, 93–123. http://dx.doi.org/10.1023/A: 1005613325648
Meyer, M. (2001). Patent citation analysis in a novel field of technology: An exploration of nano-science and nano-technology. *Scientometrics*, *51*, 163–183. http://dx.doi.org/10.1023/A: 1010572914033
Michel, J., & Bettels, B. (2001). Patent citation analysis. A closer look at the basic input data from patent search reports. *Scientometrics*, *51*, 795–816. http://dx.doi.org/10.1023/A: 1010577030871
Mina, A., Ramlogan, R., Tampubolon, G., & Metcalfe, J. S. (2007). Mapping evolutionary trajectories: Applications to the growth and transformation of medical knowledge. *Research Policy*, *36*(5), 789–806. http://dx.doi.org/10.1016/j.respol.2006.12.007
Narin, F., & Olivastro, D. (1988). Technology indicators based on patents and patent citations. In A. F. J. Van Raan (Ed.), *Handbook of quantitative studies of science and technology* (pp. 465–507). North Holland: Elsevier Publishers. http://dx.doi.org/10.1016/B978-0-444-70537-2. 50020-9
Narin, F., & Olivastro, D. (1992). Status report: Linkage between technology and science. *Research Policy*, *21*(3), 237–249. http://dx.doi.org/10.1016/0048-7333(92)90018-y
Narin, F., & Olivastro, D. (1998). Linkage between patents and papers: An interim EPO/US comparison. *Scientometrics*, *41*(1), 51–59. http://dx.doi.org/10.1007/BF02457966
Narin, F., Hamilton, K. S., & Olivastro, D. (1995). Linkage between agency-supported research and patented industrial technology. *Research Evaluation*, *5*(3), 183–187. http://dx.doi.org/10.1093/rev/5.3.183
Narin, F., Hamilton, K. S., & Olivastro, D. (1997). The increasing linkage between U.S. technology and public science. *Research Policy*, *26*(3), 317–330. http://dx.doi.org/10.1016/s0048-7333(97)00013-9
Nelson, A. J. (2009). Measuring knowledge spillovers: What patents, licenses and publications reveal about innovation diffusion. *Research Policy*, *38*(6), 994–1005. http://dx.doi.org/10.2139/ssrn.1598495
Newcombe, R. G. (2006). Confidence intervals for an effect size measure based on the Mann-Whitney statistic. Part 1: General issues and tail-area-based methods. *Statistics in Medicine*, *25*(4), 543–557. http://dx.doi.org/10.1002/sim.2323
Oliva, J., Serrano, J. I., del Castillo, M. D., & Iglesias, A. (2011). SyMSS: A syntax-based measure for short-text semantic similarity. *Data & Knowledge Engineering*, *70*(4), 390–405. http://dx.doi.org/10.1016/j.datak.2011.01.002
Oostdijk, N., Verberne, S., & Koster, C. (2010). Constructing a broad-coverage lexicon for text mining in the patent domain. *Proceedings of the international conference of language resources and evaluation*, 2292–2298.
Park, H. W., & Leydesdorff, L. (2009). Knowledge linkage structures in communication studies using citation analysis among communication journals. *Scientometrics*, *81*(1), 157–175. http://dx.doi.org/10.1007/s11192-009-2119-y
Park, H. W., & Suh, S. H. (2013). Scientific and technological knowledge flow and technological innovation: Quantitative approach using patent citation. *Asian Journal of Technology Innovation*, *21*(1), 153–169. http://dx.doi.org/10.1080/19761597.2013.815482
Park, H., Yoon, J., & Kim, K. (2012). Identifying patent infringement using SAO based semantic technological similarities. *Scientometrics*, *90*(2), 515–529. http://dx.doi.org/10.1007/s11192-011-0522-7
Porter, M. F. (1980). An algorithm for suffix stripping. *Program*, *14*(3), 130–137. http://dx.doi.org/10.1108/00330330610681286
Ribeiro, L. C., Ruiz, R. M., Bernardes, A. T., & Albuquerque, E. M. (2010). Matrices of science and technology interactions and patterns of structured growth: Implications for development. *Scientometrics*, *83*(1), 55–75. http://dx.doi.org/10.1007/s11192-009-0020-3
Robertson, S. (2004). Understanding inverse document frequency: On theoretical arguments for IDF. *Journal of Documentation*, *60*(5), 503–520. http://dx.doi.org/10.1108/00220410410560582
Salton, G., & Buckley, C. (1988). Term-weighting approaches in automatic text retrieval. *Information Processing & Management*, *24*(5), 513–523. http://dx.doi.org/10.1016/0306-4573(88)90021-0

Salton, G., Wong, A., & Yang, C. S. (1975). A vector space model for automatic indexing. *Communications of the ACM*, *18*(11), 613–620. http://dx.doi.org/10.1145/361219.361220
Sampat, B. (2010). When do applicants search for prior art? *Journal of Law and Economics*, *53*(2), 399–416. http://dx.doi.org/10.1086/651959
Schmoch, U. (1997). Indicators and the relations between science and technology. *Scientometrics*, *38*(1), 103–116. http://dx.doi.org/10.1007/bf02461126
Shih, M. J., & Liu, D. R. (2010). Patent classification using ontology-based patent network analysis. *Proceedings of the Pacific Asia conference on information systems*, 962–972.
Sparck-Jones, K. (1972). A statistical interpretation of term specificity and its application in retrieval. *Journal of Documentation*, *28*(1), 11–21. http://dx.doi.org/10.1108/eb026526
Taduri, S., Lau, G. T., Law, K. H., Yu, H., & Kesan, J. P. (2011). Developing an ontology for the US patent system. *Proceedings of the 12th annual international digital government research conference: Digital government innovation in challenging times*, 157–166. http://dx.doi.org/10.1145/2037556.2037579
Tijsen, R. (2005). Measuring and evaluating science-technology connections and interactions. In H. Moed, W. Glanzel, & U. Schmoch (Eds.), *HandBook of quantitative science and technology research: The use of publication and patent statistics in studies of S&T systems* (pp. 695–715). Dordrecht: Kluwer Academic Publishers. http://dx.doi.org/10.1007/1-4020-2755-9_32
Tseng, Y. H., Lin, C. J., & Lin, Y. I. (2007). Text mining techniques for patent analysis. *Information Processing & Management*, *43*(5), 1216–1247. http://dx.doi.org/10.1016/j.ipm.2006.11.011
Vargha, A., & Delaney, H. D. (2000). A critique and improvement of the CL common language effect size statistics of McGraw and Wong. *Journal of Educational and Behavioral Statistics*, *25*(2), 101–132. http://dx.doi.org/10.3102/10769986025002101
Verbeek, A., Debackere, K., Luwel, M., Andries, P., Zimmermann, E., & Deleus, F. (2002). Linking science to technology: Using bibliographic references in patents to build linkage schemes. *Scientometrics*, *54*(3), 399–420. http://dx.doi.org/10.1023/A:1016034516731
Verspagen, B. (2000). The role of large multinationals in the Dutch technology infrastructure: A patent citation analysis. *Scientometrics*, *47*(2), 427–448. http://dx.doi.org/10.1023/A:1005607614347
Wang, X., Zhang, X., & Xu, S. (2011). Patent co-citation networks of Fortune 500 companies. *Scientometrics*, *88*(3), 761–770. http://dx.doi.org/10.1007/s11192-011-0414-x
Wartburg, I. V., Teichert, T., & Rost, K. (2005). Inventive progress measured by multi-stage patent citation analysis. *Research Policy*, *34*(10), 1591–1607. http://dx.doi.org/10.1016/j.respol.2005.08.001
Wendt, H. W. (1972). Dealing with a common problem in social science: A simplified rank-biserial coefficient of correlation based on the U statistic. *European Journal of Social Psychology*, *2*(4), 463–465. http://dx.doi.org/10.1002/ejsp.2420020412
Weng, C., & Daim, T. (2012). Structural differentiation and its implications—Core/periphery structure of the technological network. *Journal of the Knowledge Economy*, *3*(4), 327–342. http://dx.doi.org/10.1007/s13132-011-0048-5
Wu, H. C., Luk, R. W. P., Wong, K. F., & Kwok, K. L. (2008). Interpreting tf-idf term weights as making relevance decisions. *ACM Transactions on Information Systems*, *26*(3), 1–37. http://dx.doi.org/10.1145/1361684.1361686
Yeh, H.Y., Sung, Y.S., Yang, H.W., Tsai W.C., Chen D.Z., (2013). The bibliographic coupling approach to filter the cited and uncited patent citations: A case of electric vehicle technology. Scientometrics, 94(1), 75–93.
Yoon, B., & Park, Y. (2004). A text-mining-based patent network: Analytical tool for high-technology trend. *The Journal of High Technology Management Research*, *15*(1), 37–50. http://dx.doi.org/10.1016/j.hitech.2003.09.003
Zhang, G., Ding, Y., & Milojević, S. (2013). Citation content analysis (CCA): A framework for syntactic and semantic analysis of citation content. *Journal of the American Society for Information Science and Technology*, *64*(7), 1490–1503. http://dx.doi.org/10.1002/asi.22850