Scientometrics (2012) 90:515–529
DOI 10.1007/s11192-011-0522-7

Identifying patent infringement using SAO based semantic technological similarities

Hyunseok Park · Janghyeok Yoon · Kwangsoo Kim

Received: 11 May 2011 / Published online: 23 September 2011

Abstract Companies should investigate possible patent infringement and cope with potential risks because patent litigation may have a tremendous financial impact. An important factor to identify the possibility of patent infringement is the technological similarity among patents, so this paper considered technological similarity as a criterion for judging the possibility of infringement. Technological similarities can be measured by transforming patent documents into abstracted forms which contain specific technological key-findings and structural relationships among technological components in the invention. Although keyword-based technological similarity has been widely adopted for patent analysis related research, it is inadequate for identifying patent infringement because a keyword vector cannot reflect specific technological key-findings and structural relationships among technological components. As a remedy, this paper exploited a subject–action–object (SAO) based semantic technological similarity. An SAO structure explicitly describes the structural relationships among technological components in the patent, and the set of SAO structures is considered to be a detailed picture of the inventor's expertise, which is the specific key-findings in the patent. Therefore, an SAO based semantic technological similarity can identify patent infringement. Semantic similarity between SAO structures is automatically measured using SAO based semantic similarity measurement method using WordNet, and the technological relationships among patents were mapped onto a 2-dimensional space using multidimensional scaling (MDS). Furthermore, a clustering algorithm is used to automatically suggest possible patent infringement cases, allowing large sets of patents to be handled with minimal effort by human experts.

H. Park
Department of Technology and Innovation Management, Pohang University of Science and Technology, San 31, Hyoja-dong, Nam-gu, Pohang, Kyungbuk 790-784, Republic of Korea
e-mail: howgood@postech.ac.kr

J. Yoon · K. Kim (✉)
Department of Industrial and Management Engineering, Pohang University of Science and Technology, San 31, Hyoja-dong, Nam-gu, Pohang, Kyungbuk 790-784, Republic of Korea
e-mail: kskim@postech.ac.kr

J. Yoon
e-mail: janghyoon@postech.ac.kr

The proposed method will be verified by detecting real patent infringement in prostate cancer treatment technology, and we expect this method to relieve human experts' work in identifying patent infringement.

Keywords Patent mining · Patent litigation · Subject–action–object · SAO · Natural language processing · NLP · Multidimensional scaling · Patent analysis · Patent risk

JEL Classification C63 · C82

Introduction

Companies should investigate possible patent infringement and cope with possible risks because patent litigation may have a tremendous financial impact. Kodak and Polaroid's litigation case can be an example. Kodak incorporated Polaroid technology in a product without permission and was sued for infringement on Polaroid's instant camera patent. This legal dispute took about 14 years to reach a settlement, and finally Kodak had to pay almost $1 billion to Polaroid in 1991 (Hall and Ziedonis 2001). Like this, if a company gets embroiled in legal disputes for patent infringement, significant losses in time and costs can occur (Lai and Che 2009), and the damage is often so huge that it may threaten the survival of the company. To make matters worse, patent trolls have emerged who acquire ownership of a patent to engage in litigation rather than to actually use it to produce a product, rapidly increasing patent infringement litigation (Soo et al. 2006). Therefore, patent infringement must be identified beforehand to reduce the risk of the damages from litigation. To this end, the fundamental but safest way to identify patent infringement is analysis by human experts (Durham 2004). However, patent experts cannot detect all possible infringement cases (Wallerstein et al. 1993; Majewski and Williamson 2004) because the number of patent applications has been increasing (Arundel 2001) and the increasing complexity of technologies and their convergence as industries require more technologies for individual products (Carree et al. 2000). Thus, an automated method is necessary to support experts.

In fact, various factors such as patentability, technological similarity, and scope of claims should be considered to identify the possibility of patent infringement. However, this research considered only technological similarity (Lanjouw and Schankerman 2001); we recognized technologically closely related patents as possible patent infringement. The technological similarity among patents was measured by transforming patent documents into abstracted forms which contain specific technological key-findings and structural relationships among technological components in the invention (Moehrle et al. 2005). Although keyword-based technological similarity has been widely adopted for patent analysis related research (Yoon 2008; Kim et al. 2008; Tsourikov et al. 2000; Yoon and Park 2004), it cannot identify patent infringement. This is because keyword vectors are composed of the frequency of predefined keywords only, so they cannot reflect the specific technological key-findings and structural relationships among components (Yoon et al. 2011).

Thus, this paper suggests a way to identify possible patent infringement using SAO based semantic technological similarity. An SAO based approach transforms patent documents into SAO structures, each of which is composed of a subject, verb and object, the canonical form of expressing a meaning (Franzosi 1994). Unlike keyword vectors, SAO structures are directly extracted and formulated from the patent document, the document is fully taken into account (Cascini and Zini 2008). Moreover, an SAO structure explicitly describes the structural relationships among components in the patent (Yoon and Kim 2011a), and the set of SAO structures is considered to be a detailed picture of the inventor's expertise which is specific key-findings in the patent (Moehrle et al. 2005). Thus, an SAO based semantic technological similarity can identify patent infringement. The semantic similarity between SAO structures is automatically measured using SAO based semantic similarity measurement method using WordNet; WordNet is a hierarchical thesaurus of English (Miller 1995). The technological relationships among patents were mapped onto 2-dimensional space using multidimensional scaling (MDS) (Kruskal 1964) for intuitive understanding. Since a clustering algorithm automatically suggests possible infringement cases, large sets of patents can be handled with minimal effort by human experts.

Our procedure consists of (1) extracting SAO structures from a collected patent set using natural language processing (NLP) (Manning et al. 1999), (2) measuring semantic technological similarities among patents using WordNet, (3) generating a 2-dimensional semantic patent map using MDS, (4) analyzing automatically generated clusters to identify possible patent infringement. The proposed method will be verified by detecting real cases of patent infringement in prostate cancer treatment technology, and we expect this method to reduce the demands on human experts.

The organization of this paper is as follows. First, an overview of the theoretical background is presented. Second, the overall process of generating semantic patent maps and unveiling possible patent infringement are described. Third, an example case is used to exhibit the process of analysis and to assure the utility of the method. Finally, conclusions are presented.

Theoretical background

Patent infringement

According to United States patent law 35 U.S.C. § 271, patent Infringement is defined as any type of use, manufacturing, or sale of a patented invention without permission from the patent holder. Recently, the increase in importance of intellectual property and emergence of a new kind of business model such as non-practicing entity (NPE) have incurred too many patent infringement disputes. If companies are embroiled in a patent infringement case, they may pay the tremendous size of damage awards or settlement costs. Reversely, companies also have a chance to receive a huge size of damage awards or settlement costs. Therefore, companies should detect patent infringement beforehand to avoid damage awards, and then again should discover unauthorized use of patented invention to receive proper compensation and protect own intellectual properties. Furthermore, companies can prepare robust coping strategies for patent infringement by detecting before it becomes a legal matter (for details, see Crampes and Langinier 2002). The process of detecting patent infringement consists of (1) identification of possible infringement patents which have high technological similarity, (2) patent claim analysis for identifying and determining infringement.

In this paper, we focused on the preceding process for detecting patent infringement; identification of possible infringement patents by using technological similarity.

SAO structure

The technological similarities among patents are measured by transforming each patent into an abstracted form which contains the specific technological key-findings and structural relationships among technological components. Although many studies have used keyword-based technological similarity for various purposes, a keyword-based approach, as mentioned above, is inadequate here.

As a remedy, this paper adopted SAO based semantic technological similarity. Since S and O denote the components and A denotes the effect or relationship between components in the invention, an SAO structure explicitly describes the structural relationships among components in the patent (Cascini and Zini 2008), and since S denotes "means" and A–O denotes "end" in the invention, an SAO structure fully reflects specific key-findings in the patent (Moehrle et al. 2005). A given sentence "Coolants lower temperature of the engine" is an example of an SAO structure. This SAO structure is composed of S (Coolants), A (lower) and O (temperature of the engine). "Coolants" and "temperature of the engine" denote the components and "lower" clearly defines a relationship between the components. Like this, an SAO structure explicitly describes the structural relationships among technological components and fully reflects the specific key-findings in the patent, so an SAO based technological similarity is suitable for identifying patent infringement.

Recent research related to patent analysis for R&D management such as management of human resources (Moehrle et al. 2005), evaluation of patent risk (Bergmann et al. 2007, 2008), product forecasting (Gerken et al. 2010) and identification of technological trends (Yoon and Kim 2011b) has started to exploit an SAO based approach instead of a keyword-based approach.

Semantic technological similarity measurement using WordNet

Technological similarity of patents can be identified by measuring similarity of SAO structures in patents. To this end, semantic preprocessing that combines different words with almost similar or identical meanings into representative word should be performed beforehand, and most research which employed semantic similarity measurement method for patent analysis adopted expert-based analysis to define synonym set. However, expert-based analysis requires too much human effort and cost. Furthermore, if the patent set is very large, expert-based manual categorization of synonyms is difficult or impossible to conduct. Although Bergmann's research (2007, 2008) is known as the first attempt to identify technologically similar patents by using SAO based semantic similarity measurement method for evaluating patent risk, Bergmann also exploited semantic similarity measurement method using expert-based analysis to define synonyms. To overcome the limitations, we propose the SAO based semantic technological similarity measurement using WordNet.

WordNet is a hierarchically structured lexical database for the English language that contains sets of synonyms, called synsets, of nouns, verbs, adjectives and adverbs (Miller 1995). A simple example of the structure of WordNet is as follows (Fig. 1).

Every synset contains a group of synonymous words or collocations (sequences of words that go together to form a specific meaning such as "resounding success"), and words in a synset are connected as a network of semantic relations, such as hypernym, hyponym, meronym, holonym, and coordinate term. The database of WordNet is constructed of independent modules which enable data modification and extraction and has enormous and well-verified data, and the latest version, 3.0, contains 155,287 words organized in 117,659 Synsets for a total of 206,941 word-sense pairs. Thus, it is considered to be the most suitable ontology to offer semantic information of tokens in a sentence (Yoon and Kim 2011a). The expert's effort to define synonym set can be reduced or minimized by adopting WordNet in the semantic similarity measurement method.

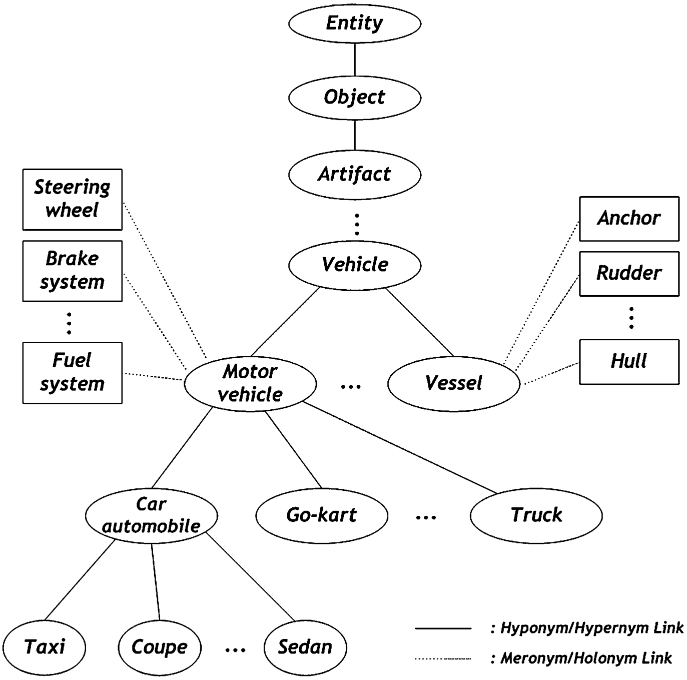


Fig. 1 An example of concept hierarchy in WordNet (Richardson and Smeaton 1995)

In this paper, WordNet is used to identify the similarity among tokens in the extracted SAO structures; the detailed method for measurement of semantic technological similarity will be presented in 3.2.

Multidimensional scaling

The measured similarity data among patents was visualized as a 2-dimensional patent map using MDS. MDS is a visualization technique for proximity data, that is, data in the form of $N \times N$ dissimilarity matrices. When the dissimilarities are distances between high-dimensional objects, MDS acts as a dimension reduction technique (Bergmann et al. 2007) and has been successfully adopted in various areas such as psychophysics, sensor analysis, marketing, and the issues of group decision making (Buja et al. 2008; Carroll and Green 1997; Chen 2009; Huang et al. 2006).

One of the most important issues in using MDS is to lessen the distance errors that occur when mapping objects in high dimensions into reduced dimensions to maintain the quality of MDS. This quality can be determined and modified by stress value. The range of stress value is from 0 (no stress) to 1 (maximum stress); MDS results are generally accepted when the stress value is less than 0.2 (Kruskal 1964).

Methodology

The overall process for identifying possible patent infringement is as follows (Fig. 2):

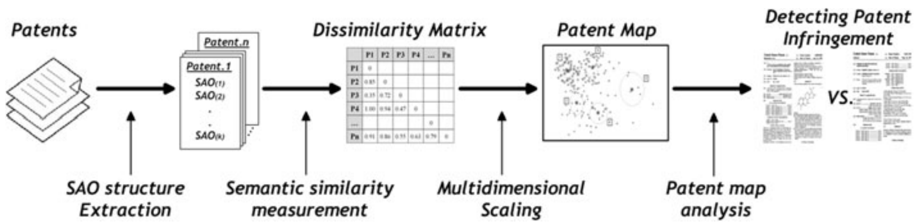


Fig. 2 Overall process for identifying possible patent infringement

SAO structure extraction from patents

The first step is to transform collected patents; patent documents into SAO structures which are a comparable and structured form. The patents related to the target technology are collected using *International Patent Classifications* (*IPC*) and related keyword retrieval from a patent database, such as United States Patent and Trademark Office (USPTO).

A patent document is mainly composed of several textual sections, such as title, abstract, description of invention and claims. This research extracted SAO structures from the abstract and claims in the patent since the abstract is a summary of the invention and claims is the utmost important part in patent which describes the exclusiveness of the invention and the boundaries of patent protection (Cascini et al. 2004). SAO structures can be extracted using NLP parser such as Stanford parser (2011) and Minipar (Lin 2003) and using commercial linguistic analyzers such as Knowledgist™ (Tsourikov et al. 2000) which can analyze grammatical structures of textual information.

Measurement of semantic technological similarity among patents

In this step, semantic technological similarities among all collected patents are measured using extracted SAO structures as shown in Fig. 3.

The technological similarity between two patents can be calculated by using similarity coefficient such as Sorensen coefficient, Jaccard coefficient, cosine coefficient, and inclusion coefficient (Moehrle 2010). This paper simply adopted Sorensen coefficient and

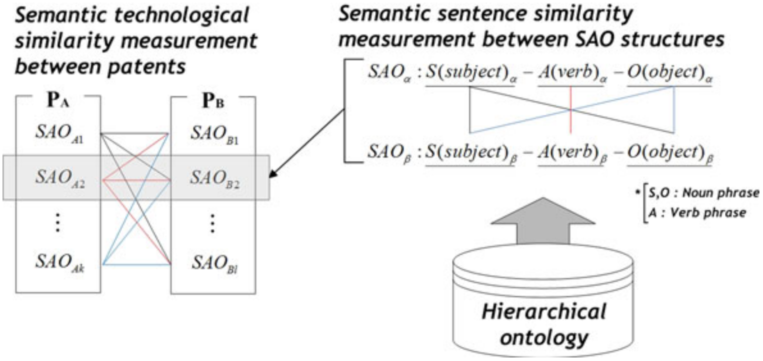

Semantic technological similarity measurement between patents
P_A
P_B
SAO_A1
SAO_A2
SAO_Ak
SAO_B1
SAO_B2
SAO_Bl
Semantic sentence similarity measurement between SAO structures
SAO_α : S(subject)_α – A(verb)_α – O(object)_α
SAO_β : S(subject)_β – A(verb)_β – O(object)_β
S,O : Noun phrase
A : Verb phrase
Hierarchical ontology

Fig. 3 Concepts for measurement of semantic technological similarity between patents

thus technological similarity is determined by how many semantically identical SAO structures are shared in the patents. The technological similarity between two patents is defined as

$$\mathrm{SIM}(X,Y) = \frac{2 \times N(X,Y)}{N(X) + N(Y)}$$

Here, $N(X)$ and $N(Y)$ were the number of SAO structures in patent X and Y, and $N(X,Y)$ is the number of semantically identical SAO structures.

The number of semantically identical SAO structures in two patents is determined by measuring the sentence similarity between extracted SAO structures, which is defined as follows (Dao and Simpson 2002):

$$\mathrm{MatAvr}(A,B) = \frac{2 \times \mathrm{Match}(A,B)}{|A| + |B|}$$

Here, Match(A,B) are the matching word tokens between SAO structures A and B. These $|A|$ and $|B|$ are the number of tokens each SAO structure has. If MatAvr is larger than the threshold value(p), which can be set by experts, two SAO structures are considered to be semantically identical.

When determining identicalness between SAO structures, the similarity between each token should be measured. WordNet was used with the following formula (Wu and Palmer 1994):

$$\mathrm{sim}_{\text{Wu and Palmer}}(c_1, c_2) = \frac{2 \times d(c_3)}{d(c_1) + d(c_2)}$$

where c_3 is the maximally specific superclass of c_1 and c_2, $d(c_3)$ is the depth from the root of the taxonomy, and $d(c_1)$ and $d(c_2)$ are the depths of c_1 and c_2 on the path through c_3. $\mathrm{sim}_{\text{Wu and Palmer}}$ ranges from 0 (completely different) to 1 (identical). For example, in Fig. 1, if c_1 and c_2 are 'Taxi' and 'Truck', c_3 is 'Motor vehicle' (for more details of semantic distance in WordNet, see Budanitsky and Hirst 2001).

Although WordNet was adopted to measure the semantic sentence similarity due to its enormous and well-verified data, it doesn't contain all domain-specific terms including abbreviations. Thus, as a supplement, we defined some domain-specific terms by reviewing the extracted SAO structures and added them to WordNet to enforce the ontologies.

The similarity data were visualized through MDS onto 2-dimensional space. Since MDS requires a dissimilarity matrix, measured similarity data should be transformed into dissimilarity data as follows:

$$\mathrm{DSIM}(X,Y) = 1 - \mathrm{SIM}(X,Y),$$

where the range of *DSIM* is from 0 (identical) to 1 (completely different).

Generation of semantic patent map using MDS

In this paper, MDS is adopted to visualize the measured dissimilarity matrix onto 2-dimensional patent maps. Metric MDS or nonmetric MDS can be selected according to whether the similarity or dissimilarities data are quantitative (metric MDS) or qualitative (nonmetric MDS) (Wickelmaier 2003; Mead 1992). Various MDS algorithms such as PROXCAL, ALSCAL, SMACOF and PREFSCAL are available, and commercial statistics or network analysis software packages such as SPSS, UCINET, NetMiner and Krackplot provide these algorithms. Particularly, ALSCAL and PROXCAL algorithms often have been adopted for various research of patent analysis. ALSCAL allows for input data only to be in the form of dissimilarity matrix while PROXCAL allows for input data to be either in the form of similarity matrix or dissimilarity matrix. There are very little differences between the results using ALSCAL or PROXCAL (Boslaugh and Watters 2008). Thus, we simply adopted ALSCAL algorithm and NetMiner in this paper.

Identifying possible patent infringement through patent map analysis

This step identifies possible patent infringement through analysis of clusters on the generated semantic patent map. First of all, clusters that consist of more than two patents located closely on the patent map can be identified by clustering algorithms such as hierarchical and non-hierarchical clustering algorithms (Johnson 1967). Particularly, the generated clusters with relatively lower average distance among patents can be considered as important spots for identifying possible patent infringement since lower average distance signifies that patents in the cluster are highly similar. Thus, in this paper, we evaluated agglomerative hierarchical clustering algorithm which proceeds by a series of fusions of the n objects into groups (Davidson and Ravi 2005), found it to be the most suitable, and adopted this algorithm.

When analyzing patents in each cluster, some or all patents could be filed by a single assignee. These patents are a patent portfolio, a collection of related patents owned by a single individual or company to obtain monetary benefits such as a market monopoly position and non-monetary benefits such as first-mover advantages and defense against rival portfolio holders (Ernst 1998). Thus a patent portfolio should be treated as a single patent in the analysis. For example, if a cluster had four patents which are filed by three different assignees, two out of four patents filed by a single assignee are a patent portfolio and should be treated as a single patent, and thus the analysis considers three items: two patents and one patent portfolio. In another case, if four patents are filed by a single assignee, no patent infringement could occur in this cluster.

Empirical study: prostate cancer treatment technology

Outline of the patent set: prostate cancer treatment technology related patents

The proposed method was applied to prostate cancer treatment technology to verify its usefulness. Since prostate cancer treatment technology belongs to the pharmaceutical domain, where patent infringement has occurred very often, collected patents in this technology are adequate for verifying the proposed method. The patent set contains 96 patents and was constructed by (1) Patents in infringement litigation between Merck's one E.P. patent and Smithkline Beecham Corp's three U.S. patents, (2) Patent portfolio set of ASTA Medica's four US patents and Cornell Research Foundation's seven US patents, (3) 81 randomly selected patents from 264 patents that were collected by prostate cancer treatment related keyword retrieval, and patent documents extracted from the USPTO database.

SAO based semantic technological similarities

SAO structure extraction

As a first step, SAO structures were extracted from the abstract and claims of the collected patent set using Knowledgist™ 2.5, a commercial NLP program. But, some SAO structures were not extracted properly, and we complemented these insufficient parts with expert knowledge. As shown in Table 1, the extracted SAO structures contain the specific technological key-findings in the patent and explicitly describe the structural relationships among technological components in the patent.

Semantic technological similarity measurement

In this step, semantic technological similarity was measured among all sets of SAO structures. The threshold value (p) was set to 0.8 to determine whether extracted SAO structures are semantically identical. The dissimilarity matrix is a (96×96) symmetric matrix with a zero diagonal, and one part of calculated dissimilarity matrix is shown in Fig. 4.

Generating semantic patent maps

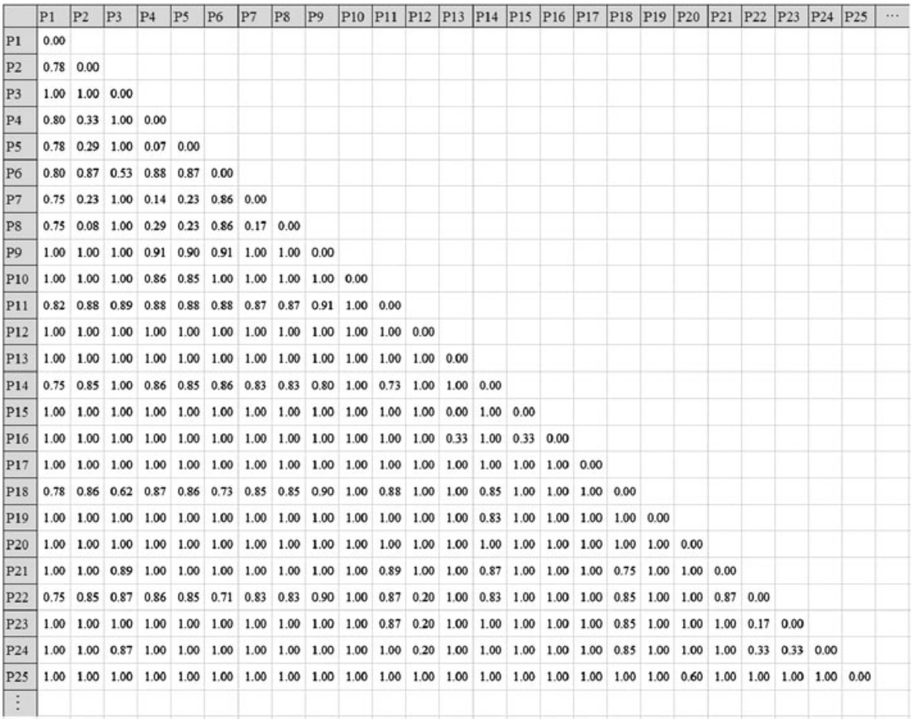
The 2-dimensional semantic patent map was generated from the measured dissimilarity matrix using the ALSCAL algorithm and Kruskal's Non-metric MDS. Since the real patent serial numbers are too long to be displayed, serial numbers from 1 to 96, sorted by the application date, were attached to the patents. The semantic patent map of prostate cancer related technology is shown in Fig. 5, and MDS quality of stress value is 0.1685.

Identifying possible patent infringement

This step interprets the generated semantic patent map and identified clusters to identify possible patent infringement. Clusters were created using an agglomerative hierarchical clustering algorithm, and the cut-off distance value (d) was set to 0.1 to group possible infringing patents into clusters. As shown in Fig. 5, nine clusters were identified and labeled in ascending numerical order of the average distance within the cluster. The identified clusters are shown in Table 2.

Table 1 A sample of extracted SAO structures from US 5,637,310

S (subject)	A (verb)	O (object)
Pharmaceutical compositions	Comprise	Pharmaceutical carrier
Method	Employ	Steroid 5-alpha-reductase
Steroid 5-alpha-reductase inhibitor	Decrease	Size of prostate tumor
Combination of inhibitors of steroid 5-alpha-reductase used in pharmaceutical composition	Treat	Prostatic adenocarcinoma
Present invention	Reside in	Discovery
Method	Treat	Human prostatic adenocarcinoma
Steroid 5-alpha-reductase inhibiting compounds	Have	Therapeutic effect



	P1	P2	P3	P4	P5	P6	P7	P8	P9	P10	P11	P12	P13	P14	P15	P16	P17	P18	P19	P20	P21	P22	P23	P24	P25	···
P1	0.00																									
P2	0.78	0.00																								
P3	1.00	1.00	0.00																							
P4	0.80	0.33	1.00	0.00																						
P5	0.78	0.29	1.00	0.07	0.00																					
P6	0.80	0.87	0.53	0.88	0.87	0.00																				
P7	0.75	0.23	1.00	0.14	0.23	0.86	0.00																			
P8	0.75	0.08	1.00	0.29	0.23	0.86	0.17	0.00																		
P9	1.00	1.00	1.00	0.91	0.90	0.91	1.00	1.00	0.00																	
P10	1.00	1.00	1.00	0.86	0.85	1.00	1.00	1.00	1.00	0.00																
P11	0.82	0.88	0.89	0.88	0.88	0.88	0.87	0.87	0.91	1.00	0.00															
P12	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.00														
P13	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.00													
P14	0.75	0.85	1.00	0.86	0.85	0.86	0.83	0.83	0.80	1.00	0.73	1.00	1.00	0.00												
P15	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.00	1.00	0.00											
P16	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.33	1.00	0.33	0.00										
P17	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.00									
P18	0.78	0.86	0.62	0.87	0.86	0.73	0.85	0.85	0.90	1.00	0.88	1.00	1.00	0.85	1.00	1.00	1.00	0.00								
P19	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.83	1.00	1.00	1.00	1.00	0.00							
P20	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.00						
P21	1.00	1.00	0.89	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.89	1.00	1.00	0.87	1.00	1.00	1.00	0.75	1.00	1.00	0.00					
P22	0.75	0.85	0.87	0.86	0.85	0.71	0.83	0.83	0.90	1.00	0.87	0.20	1.00	0.83	1.00	1.00	1.00	0.85	1.00	1.00	0.87	0.00				
P23	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.87	0.20	1.00	1.00	1.00	1.00	1.00	0.85	1.00	1.00	1.00	0.17	0.00			
P24	1.00	1.00	0.87	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.20	1.00	1.00	1.00	1.00	1.00	0.85	1.00	1.00	1.00	0.33	0.33	0.00		
P25	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.60	1.00	1.00	1.00	1.00	0.00	
⋮																										

Fig. 4 A part of the dissimilarity matrix of prostate cancer treatment related patents

We identified that the patents intentionally added to patent set were successfully grouped into the same clusters, clusters 1, 4 and 9. Thus, empirical study focused on analysis of these three clusters, and the following is a detailed interpretation of the clusters.

(1) The technology in cluster 1 is about the treatment of hormone-sensitive cancers of the prostate and breast by administering a chemical compound called Cetrorelix acetate which is an injectable gonadotropin-releasing hormone antagonist (GnRH antagonist). Four patents (US 6,054,432(P13), US 5,998,377(P15), US 6,071,882(P16) and US 6,300,313(P26)) are included and were filed by one company, ASTA Medica, Inc. This cluster is an example of a patent portfolio which was constructed to prevent monetary and non-monetary benefits of patents from depreciating. Since a patent portfolio usually contains technologically similar or equivalent technology patents, it appears almost the same as a possible patent infringement case in the patent maps. However, a patent portfolio case can be distinguished from patent infringement by referring to the bibliographic data of the patents, especially assignee or inventor's data.

(2) Cluster 4 contains five patents, four of which (P2, P4, P5 and P7) were involved in a patent infringement lawsuit. These four patents are EP 285,383(P2) filed by Gary H. Rasumsson in Merck & Co. and US 5,300,294(P4) US 5,496,556(P5) and US 5,637,310(P7) filed by SmithKline Beecham (now GlaxoSmithKline). They deal with the technology of treating a type of prostate cancer by administering a chemical compound called Finasteride that inhibits the production of specific enzyme, 5-a-reductase, which is responsible for converting the hormone testosterone to dihydrotestosterone ("DHT").

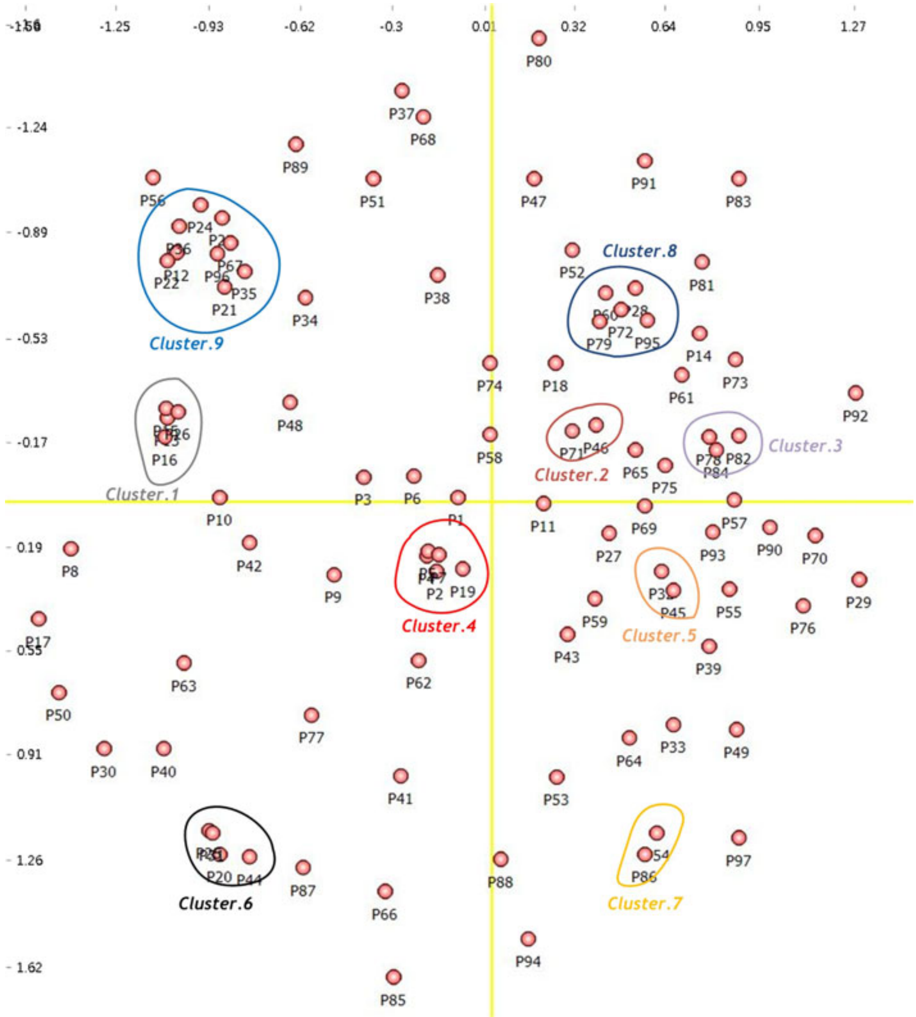


Fig. 5 Semantic patent map of prostate cancer related technology

In 2001, the USPTO declared an interference between Rasmusson's patent, US patent application No. 08/460,296 filed on June, 1995 as 9th in a series of applications on April, 1987 and SmithKline Beecham's patents; US 5,300,294 filed on June, 1990, and US 5,496,556 and US 5,637,310 which were accorded the benefit of the filing date of US 5,300,294. Rasmusson asserted that SmithKline Beecham's patents US 5,300,294, US 5,496,556 and US 5,637,310 could be anticipated by its European patent application, EP No. 285,383 filed on October, 1988. But, the Federal Circuit decided that Rasmusson was not entitled to benefit from a priority filing date according to United States patent law 35 U.S.C. § 112 Specification, because of a lack of data for demonstrating effects of Finasteride in treating prostate cancer. Although Rasmusson's European patent application, EP No. 285,383 did not sufficiently

Table 2 A brief summary of clusters

Cluster number	Cluster members (patents)	Number of members	Average distance
1	US 6,054,432(P13), US 5,998,377(P15), US 6,071,882(P16), US 6,300,313(P26)	4	0.069
2	US 7,041,844(P46), US 2009-0215853(P71)	2	0.070
3	US 2007-0244110(P78), US 2007-0238647(P84), US 2009-0175868(P82)	3	0.072
4	EP 285,383(P2), US 5,300,294(P4), US 5,496,556(P5), US 5,637,310(P7), US 6,165,504(P19)	5	0.076
5	US 6,670,392(P32), US 2003-0082188(P45)	2	0.079
6	US 6,174,858(P20), US 6,093,722(P25), US 2001-0041713(P31), US 2003-0133927(P44)	4	0.093
7	US 7,470,431(P54), US 7,858,325(P86)	2	0.094
8	US 2008-0200393(P72), US 6,409,664(P28), US 2007-0196346(P60), US 7,666,584(P79), US 2009-0311716(P95)	5	0.12
9	US 6,107,090(P12), US 6,360,116(P21), US 7,666,425(P22), US 7,112,412(P24), US 7,112,412(P23), US 6,767,711(P35), US 2003-0003101(P36), US 7,858,323(P67), US 2010-0092477(P96)	9	0.138

describe the effects of Finasteride in treating prostate cancer, the US Court of Appeals for the Federal Circuit (CAFC) sided with Rasmusson according to United States patent law 35 U.S.C. § 102 Conditions for patentability; novelty and loss of right to patent.

In conclusion, even though Rasmusson's US patent application No. 08/460,296 was rejected, they achieved another favorable decision that SmithKline Beecham's patents were anticipated by Rasmusson's European patent application, EP No. 285,383 (see CAFC case 413 F.3d 1318, June 27, 2005).

(3) Cluster 9 includes nine patents, eight of which are about treating and diagnosing prostate cancer with antibodies to the extracellular prostate-specific membrane antigen (PSMA) using immunoglobulin G (IgG), an antibody molecule. Six patents in the cluster; US 6,107,090(P12), US 7,666,425(P22), US 7,112,412(P23), US 6,290,956(P24), US 6,767,711(P35) and US 2003-0003101(P36) are owned by a single company, Cornell Research Foundation, and form a patent portfolio like cluster 1. The other two patents; US 7,858,323(P67) and US 2010-0092477(P96) filed by The Regents of the University of Michigan and Hovens, Christopher respectively are also methods for prostate cancer treatment or diagnostics with IgG.

Although two patents (US 7,858,323 and US 2010-0092477) and a patent portfolio (US 6,107,090, US 7,666,425, US 7,112,412, US 6,290,956, US 6,767,711 and US 2003-0003101) in this cluster hasn't produced infringement litigation yet, these patents have high possibilities of infringement in that these patents were not only included in the same cluster on the generated patent map but were also actually identified as highly similar technologies by qualitative analysis of domain experts.

Conclusion

Identifying possible patent infringement beforehand and reacting to potential risks is significant since, if companies were involved in a lawsuit, it could cause tremendous financial damage. To identify the possibility of patent infringement, the technological similarity among patents is an important factor and can be adopted as a criterion. A keyword-based technological similarity has been widely exploited for various purposes such as technology trend or opportunity analysis. However, since a keyword-based approach cannot reflect specific technological key-findings and structural relationships among technological components in inventions, this approach is inadequate for identifying patent infringement.

Thus, we proposed a method to identify possible patent infringement using SAO based semantic similarity of technology. Since an SAO structure explicitly describes the structural relationships among technological components in the patent and the set of SAO structures is considered to be a detailed picture of the inventor's expertise which are the specific key-findings in the patent, an SAO based semantic technological similarity can adequately identify patent infringement. Technological similarity among patents was measured by SAO based semantic similarity using WordNet instead of using expert-based analysis. Furthermore, a clustering algorithm automatically suggested possible patent infringement on the patent map, making it possible to handle large patent sets with minimal effort of human experts. The usefulness of the proposed method was demonstrated by identifying real case of patent infringement in prostate cancer treatment technology. Although this paper analyzed a case, in which patents were infringed by other patents, the proposed method can be applied to the identification of patent infringement of products or processes by examining the detailed functions of products or processes by experts, transcribing them into the form of SAO structure, and then treating such documentation of the product or process as a 'patent' like entity.

However, the proposed method needs to be improved in some aspects parts: the hierarchical ontology for semantic similarity measurement and factors considered as criteria for judging patent infringement. First, even though the proposed method can be applied to any technology domain through the strengths of the powerful hierarchical ontology, WordNet, WordNet also has limitations in that certain unusual domain-specific terms may be absent from its collection of words, especially chemical and bio-technology related terms. Without supplement of undefined domain-specific terms to WordNet, credibility problem can occur or in appropriate analysis result can be generated. Thus, insufficient domain-specific terms should be added to WordNet to obtain reliable result from semantic similarity measurement and furthermore, to achieve full automation system. Recently, research for automated generation methods of local concept hierarchy has been actively conducted in Computer Science. These methods can generate the concept hierarchy of technology specific terms. Thus, these automated generation methods can make WordNet more robust and allow fully automated semantic similarity measurement. Second, although this paper only considered technological similarity as a criterion for judging possible patent infringement, other valuable factors for identifying the possibility of patent infringement such as patentability and claims need to be considered to improve the reliability and practicality of the method.

Acknowledgments This work was supported by a National Research Foundation of Korea (NRF) grant funded by the Korean government (MEST) (No. 2009-0088379).

References

Arundel, A. (2001). The relative effectiveness of patents and secrecy for appropriation. *Research Policy, 30*(4), 611–624.

Bergmann, I., Moehrle, M. G., Walter, L., Butzke, D., Erdmann, V. A., & Furste, J. P. (2007). The use of semantic maps for recognition of patent infringements: A case study in biotechnology. *Zeitschrift fur Betriebswirtschaft—Special issue,* (4), 69–86.

Bergmann, I., Butzke, D., Walter, L., Fuerste, J. P., Moehrle, M. G., & Erdmann, V. A. (2008). Evaluating the risk of patent infringement by means of semantic patent analysis: The case of DNA chips. *R&D Management, 38*(5), 550–562.

Boslaugh, S., & Watters, P. A. (2008). *Statistics in a nutshell*. Sebastopol, CA: O'Reilly Media, Inc.

Budanitsky, A., & Hirst, G. (2001). Semantic distance in WordNet: An experimental, application-oriented evaluation of five measures. In Proceedings of the Workshop on WordNet and Other Lexical Resources NAACL.

Buja, A., Swayne, D. F., Littman, M. L., Dean, N., Hofmann, H., & Chen, L. (2008). Data visualization with multidimensional scaling. *Journal of Computational and Graphical Statistics, 17*(2), 444–472.

Carree, M. A., Klomp, L., & Thurik, A. R. (2000). Productivity convergence in OECD manufacturing industries. *Economics Letters, 66*(3), 337–345.

Carroll, J. D., & Green, P. E. (1997). Psychometric methods in marketing research: Part II, multidimensional scaling. *Journal of Marketing Research, 34*(2), 193–204.

Cascini, G., & Zini, M. (2008). Measuring patent similarity by comparing inventions functional trees. *Computer-Aided Innovation* (*CAI*), *277*, 31–42.

Cascini, G., Fantechi, A., & Spinicci, E. (2004). Natural language processing of patents and technical documentation. *Document Analysis Systems, VI*, 89–92.

Chen, R. (2009). Design patent map visualization display. *Expert Systems with Applications, 36*(10), 12362–12374.

Crampes, C., & Langinier, C. (2002). Litigation and settlement in patent infringement cases. *The RAND Journal of Economics*, *33*(2), 258–274.

Dao, T. N., & Simpson, T. (2002). Measuring similarity between sentences. http://www.codeproject.com/KB/string/semanticsimilaritywordnet.aspx.

Davidson, I., & Ravi, S. (2005). Agglomerative hierarchical clustering with constraints: Theoretical and empirical results. *Knowledge Discovery in Databases: PKDD, 2005*, 59–70.

Durham, A. L. (2004). *Patent law essentials: A concise guide*. Westport, CT: Praeger Publishers.

Ernst, H. (1998). Patent portfolios for strategic R&D planning. *Journal of Engineering and Technology Management, 15*(4), 279–308.

Franzosi, R. (1994). From words to numbers: A set theory framework for the collection, organization and analysis of narrative data. *Sociological methodology, 24*, 105–136.

Gerken, J., Moehrle, M., & Walter, L. (2010). Patents as an information source for product forecasting: Insights from a longitudinal study in the automotive industry. In *The R&D management conference 2010*, Manchester, England.

Hall, B. H., & Ziedonis, R. H. (2001). The patent paradox revisited: An empirical study of patenting in the US semiconductor industry, 1979–1995. *The RAND Journal of Economics, 32*(1), 101–128.

Huang, J. J., Ong, C. S., & Tzeng, G. H. (2006). Interval multidimensional scaling for group decision using rough set concept. *Expert Systems with Applications, 31*(3), 525–530.

Johnson, S. C. (1967). Hierarchical clustering schemes. *Psychometrika, 32*(3), 241–254.

Kim, Y., Suh, J., & Park, S. (2008). Visualization of patent analysis for emerging technology. *Expert Systems with Applications, 34*(3), 1804–1812.

Kruskal, J. B. (1964). Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis. *Psychometrika, 29*(1), 1–27.

Lai, Y. H., & Che, H. C. (2009). Modeling patent legal value by Extension Neural Network. *Expert Systems with Applications, 36*(7), 10520–10528.

Lanjouw, J. O., & Schankerman, M. (2001). Characteristics of patent litigation: A window on competition. *The RAND Journal of Economics, 32*(1), 129–151.

Lin, D. (2003). Dependency-based evaluation of MINIPAR. In *Treebanks*: *Building and using parsed corpora* (Vol. 20, pp. 317–332). Springer Netherlands.

Majewski, S. E., & Williamson, D. V. (2004). Incomplete contracting and the structure of R&D joint venture contracts. In Professor G. Libecap (ed.), *Intellectual property and entrepreneurship (advances in the study of entrepreneurship*, innovation & economic growth, Vol. 15, pp. 201–228). Emerald Group Publishing Limited.

Manning, C. D., Schutze, H., & MITCogNet. (1999). *Foundations of statistical natural language processing* (Vol. 59). Cambridge, MA: MIT Press.

Mead, A. (1992). Review of the development of multidimensional scaling methods. *The Statistician, 41*(1), 27–39.

Miller, G. A. (1995). WordNet: A lexical database for English. *Communications of the ACM, 38*(11), 39–41.

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

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

Richardson, R., & Smeaton, A. F. (1995). *Using WordNet in a knowledge-based approach to information retrieval*. Dublin City University School of Computer Applications Working Paper CA-0395.

Soo, V. W., Lin, S. Y., Yang, S. Y., Lin, S. N., & Cheng, S. L. (2006). A cooperative multi-agent platform for invention based on patent document analysis and ontology. *Expert Systems with Applications, 31*(4), 766–775.

Stanford (2011). The Stanford Parser: A statistical parser http://nlp.stanford.edu/software/lex-parser.shtml. Accessed Feb 2011.

Tsourikov, V. M., Batchilo, L. S., & Sovpel, I. V. (2000). Document semantic analysis/selection with knowledge creativity capability utilizing subject-action-object (SAO) structures. United States Patent No. 6167370.

Wallerstein, M. B., Mogee, M. E., & Schoen, R. A. (1993). *Global dimensions of intellectual property rights in science and technology*. Washington, DC: National Academies Press.

Wickelmaier, F. (2003). *An introduction to MDS*. Aalborg: Aalborg Universitetsforlag.

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

Yoon, B. (2008). On the development of a technology intelligence tool for identifying technology opportunity. *Expert Systems with Applications, 35*(1–2), 124–135.

Yoon, J., & Kim, K. (2011a). Generation of patent maps using SAO-based semantic patent similarity. *Entrue Journal of Information Technology, 10*(1), 19–27.

Yoon, J., & Kim, K. (2011b). Identifying rapidly evolving technological trends for R&D planning using SAO-based semantic patent networks. *Scientometrics*. doi:10.1007/s11192-011-0383-0.

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.

Yoon, J., Choi, S., & Kim, K. (2011). Invention property-function network analysis of patents: A case of silicon-based thin film solar cells. *Scientometrics, 86*(3), 687–703.