Scientometrics (2012) 90:445–461
DOI 10.1007/s11192-011-0543-2

Detecting signals of new technological opportunities using semantic patent analysis and outlier detection

Janghyeok Yoon · Kwangsoo Kim

Received: 12 April 2011 / Published online: 20 October 2011

Abstract In the competitive business environment, early identification of technological opportunities is crucial for technology strategy formulation and research and development planning. There exist previous studies that identify technological directions or areas from a broad view for technological opportunities, while few studies have researched a way to detect distinctive patents that can act as new technological opportunities at the individual patent level. This paper proposes a method of detecting new technological opportunities by using subject–action–object (SAO)-based semantic patent analysis and outlier detection. SAO structures are syntactically ordered sentences that can be automatically extracted by natural language processing of patent text; they explicitly show the structural relationships among technological components in a patent, and thus encode key findings of inventions and the expertise of inventors. Therefore, the proposed method allows quantification of structural dissimilarities among patents. We use outlier detection to identify unusual or distinctive patents in a given technology area; some of these outlier patents may represent new technological opportunities. The proposed method is illustrated using patents related to organic photovoltaic cells. We expect that this method can be incorporated into the research and development process for early identification of technological opportunities.

Keywords Technological opportunity · Outlier detection · Patent mining · Subject–action–object (SAO) structure · Semantic patent similarity · Multidimensional scaling (MDS) · Research and development (R&D) planning

JEL Classification C63 · C82

J. Yoon
Korea Institute of Intellectual Property, KIPS Center, Yeoksam-dong,
Gangnam-gu, Seoul 135-980, Republic of Korea
e-mail: janghyoon@gmail.com

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

Introduction

Faced with a competitive business environment, many companies are seeking new technological opportunities, and for this purpose, patents have been widely exploited as a technological knowledge source to identify recent technological trends. Patents constitute reliable intelligence reflecting advances in technological development (Yoon et al. 2011), and thus patent analysis has been considered to be a vital tool for technology strategy formulation and research and development (R&D) planning (Mogee and Kolar 1994).

Early detection of a possible discontinuity in technological bases of products is considered to be the most important aspect of technology planning (Schuh and Grawatsch 2004). For these reasons, patent-based studies for technological opportunities have proposed approaches and systems for identifying potential technological types by exploiting keyword-based morphology analysis (Yoon and Park 2004, 2005; Yoon 2008), for identifying undeveloped technological areas by analyzing patent vacuums (Lee et al. 2009; Yoon et al. 2002), for forecasting products by temporal analysis of patent maps (Gerken et al. 2010), for generating discontinuous ideas of product improvements by exploiting system evolution patterns (Mann 2002, 2003). These previous studies focus on exploring technological opportunities that have not been explored yet, so their output are technological directions or technological areas from a broad view that are promising and undeveloped. However, according to the best knowledge of the authors, few studies have researched how to detect patents that can act as new technological opportunities at the individual patent level. Although there exist widely-used measures to identify technological importance of patents using citations (Albert et al. 1991; Karki 1997; Narin 1994), citation-based indices, i.e., current impact index, measure the 'current' impact of patents but do not address the 'future' potential of patents.

Therefore, this paper proposes a new method to detect distinctive patents that can act as technological opportunities at the individual patent level by exploiting subject–action–object (SAO)-based semantic patent analysis and outlier detection. SAO structures are syntactically ordered sentences extractable by natural language processing (NLP) of patent text; they explicitly describe the structural relationships among components in the relevant patent (Cascini et al. 2004; Mann 2002). The set of SAO structures is considered to be a detailed picture of the inventor's expertise (Moehrle et al. 2005; Yoon and Kim 2011a; Radauer and Walter 2010). Therefore, the proposed method measures the similarity of invention structures of patents by computing semantic similarities among SAO structures of patents. The ultimate goal of this method is to detect patents that differ greatly from the rest of a group. Outlier detection is used to identify such observations, so we adopt it to detect such patents in a given technology area.

The procedure proposed in this paper consists of (1) collecting patent data, (2) analyzing syntactic structure of patent text by exploiting NLP, (3) generating a patent dissimilarity matrix that codifies technological distance between pairs of patents by measuring semantic sentence similarities between their SAO structures, (4) mapping patents onto a lower-dimensional space using multidimensional scaling (MDS), and (5) identifying outlier patents from the given patent set by exploiting distance-, density- and clustering-based outlier detection methods. By measuring the degree to which a patent differs from the rest of the group in the lower-dimensional space (i.e., "outlierness"), the proposed method quantitatively identifies patents that are distinct from the others. These unusual patents may signal development of a new technology. Recently, a large amount of patents are being increasingly generated, so this increase made relying only on expert's intrinsic skills to identify technological opportunities from patent analysis almost impossible. Therefore, the automated and quantified method of this paper can assist technology expert in identifying existence or emergence of outlier patents that may provide fresh ideas for further technological development. The proposed method is illustrated using patents related to organic photovoltaic cells (OPVCs). We expect that the proposed method can be incorporated into the strategic technology planning process to identify potential technological opportunities.

Section "Theoretical background" reviews groundwork and section "Data" describes data source for analysis. Section "Proposed method and results" suggests a procedure to detect signals of new technological opportunities. Finally, section "Summary and discussions" presents conclusions and future research directions.

Theoretical background

The procedure proposed in this paper is based on SAO-based patent analysis and outlier detection. Therefore this section presents a brief overview of these techniques.

Subject–action–object based patent analysis

In this paper, each patent is represented as a set of SAO structures, which are composed of subject (noun phrase), action (verb phrase) and object (noun phrase). SAO structures that can be extracted by grammatical processing of patent text are the syntactically ordered sentences. Consider a simple sentence 'ultrasonic waves remove small particles'. In this SAO structure, the subject is 'ultrasonic waves', the action is 'remove', and the object is 'small particles'. 'Remove' explicitly represents a structural relationship between the subject 'ultrasonic waves' and the object 'small particles'. Each patent contains many SAO structures in its text.

From the view of the Theory of Inventive Problem Solving (Russian acronym: TRIZ) (Altschuller 1984), SAO structures are fundamentally related to the concept of function (Yoon and Kim 2011b), which is defined as "the action changing a feature of any object" (Savransky 2000). Therefore, in SAO structures of technological documents, subjects and object refer to technological components and actions refer to functions performed by and on components (Cascini et al. 2004; Mann 2002), so SAO structures are considered to represent key findings of inventions (Choi et al. 2010; Radauer and Walter 2010). Because SAO structures include structural relationships among technological components, they are useful in analysis of the design structure of patents. Several design structure analyses have adopted the SAO-based approach to visualize design structure of patents as SAO-based networks (Cascini et al. 2004), to find inventive principles from patents (Cascini et al. 2007), and to measure similarity of invention structures using SAO-based networks (Cascini and Zini 2008). Especially, SAO structures extracted from patent claims are considered to represent the inventor's expertise (Moehrle et al. 2005) because the claims include the intensive knowledge that requires legal protection (Fujii et al. 2007). Therefore, SAO-based studies have presented patent-based merger and acquisition strategies (Moehrle and Geritz 2004), patent-based human resource decision making for R&D projects (Moehrle et al. 2005), patent infringement risk evaluation (Bergmann et al. 2008), and product forecasting based on patent maps (Gerken et al. 2010). The procedure proposed in this paper represents each patent as a set of SAO structures, and identifies the structural similarities of patents in a semantic way.

Outlier detection

In statistics, an outlier is an observation that is numerically distant from the rest of a group (Barnett et al. 1979). Popular outlier identification techniques include model-based techniques, distance-based techniques, one-class support vector machines, replicator neural networks and cluster analysis based outlier detection (Chandola et al. 2009). In some areas, outliers are due to measurement error, and researchers should discarded them or use statistical methods that are relatively unaffected by outliers (Franses et al. 1998; Park 2002). However, in some case outliers represent a deviation from an established rule, trend or pattern (Chandola et al. 2009). In this sense, outlier detection is considered to be a critical task in many safety critical environments as the outlier indicates abnormal running conditions (Hodge and Austin 2004). An outlier may pinpoint an intruder inside a system with malicious intentions, so rapid detection is critical. Furthermore, identifying outliers help detect a fault on a production line by monitoring specific features of products and comparing real-time data with the features of normal products. By identifying these unusual observations, many studies have provided methods to detect credit card fraud (Aleskerov et al. 2002), to forecast flooding attack (Siris and Papagalou 2006), to detect network intrusions (Leung and Leckie 2005; Sekar et al. 2002).

Likewise, this paper aims detecting outlier patents in a given patent set, with the goal of identifying potential technological opportunities. Technological environment of today is rapidly changing, so early identification of the outlier patents that propose novel approaches including methods (or materials) or applications (or uses) can provide technology experts with fresh insight for further technological development.

Data

An OPVC is a photovoltaic cell that uses conductive organic polymers or small organic molecules to absorb light and transport charges. OPVCs are less efficient and more fragile than are inorganic, but use raw materials that are inexpensive, flexible, and have high optical absorption coefficients. Therefore, OPVCs are considered to be a potential technology for many photovoltaic applications. Over the past decade, many inventions in this emerging field have been patented.

To illustrate the proposed method, we first collected granted patents from the United States Patent and Trademark Office (USPTO) since 2000 using a patent retrieval query composed of OPVC-related keywords, patent classification codes and date conditions (Table 1). A total of 212 patents were retrieved using the query, but generally some of them may be irrelevant to OPVCs; for example, some may be patents related to silicon-based or inorganic solar cells. After eliminating 63 irrelevant patents by checking patent titles and abstracts, finally 149 patents were prepared for analysis. The patent set ranges from US patent number 4,009,054 to 7,863,448, but these long real patent numbers make displaying maps cumbersome, so the patents were labeled from P1 to P149 in application date order.

Proposed method and results

This section describes an overall procedure for detecting signals of new technological opportunities (Fig. 1) and illustrates the procedure using the gathered OPVC-related patents.

Table 1 Retrieval query for OPVC-related patents

Retrieval query	Patents
((((((photovoltaic* solar*) adj (cell* batter* device*)) and ((bulk* adj heterojunc*) ppv* phenylenevinylen* tandem* (dye* adj sensitiz*) fluoren* fulleren* PTCBI* PTCDA* PTCDI* H2PC* ZnPc* CuPc* TPyP* TFD* NPD* CBP* PCBM* (conjugat* adj polymer*))) or (((organic* plastic* polymer* (dye* adj sensitiz*)) adj3 (photovoltaic* solar*) adj (cell* batter* device*)) DSSC*))) and (B32B* C07* H01* H05B-033*).IPC.) AND @RD>=20000101<=20101231	149 patents (after eliminating irrelevant patents)

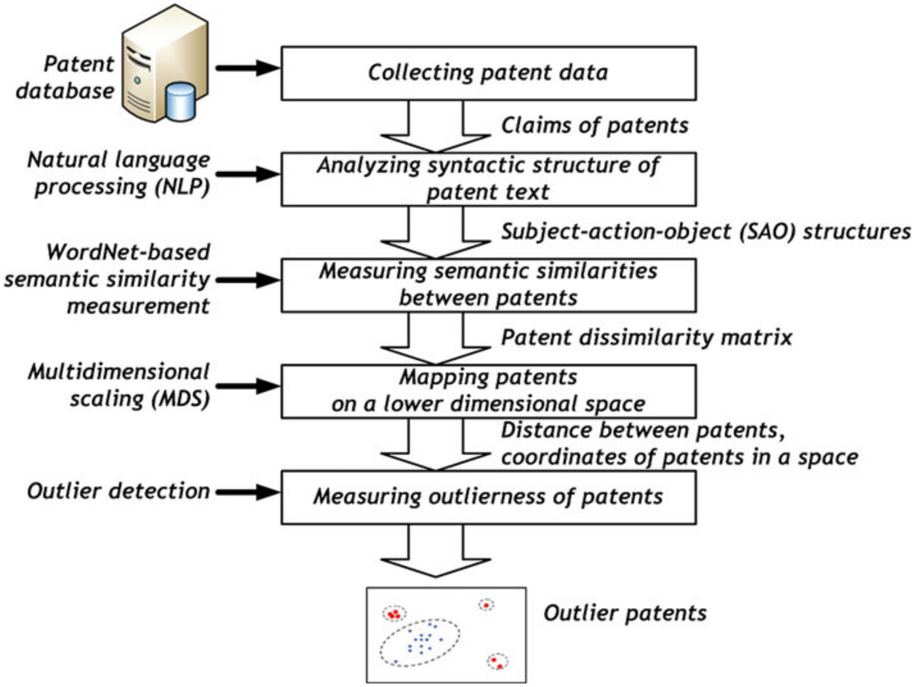

Patent database
Collecting patent data
Claims of patents
Natural language processing (NLP)
Analyzing syntactic structure of patent text
Subject-action-object (SAO) structures
WordNet-based semantic similarity measurement
Measuring semantic similarities between patents
Patent dissimilarity matrix
Multidimensional scaling (MDS)
Mapping patents on a lower dimensional space
Distance between patents, coordinates of patents in a space
Outlier detection
Measuring outlierness of patents
Outlier patents

Fig. 1 Overall procedure for detecting signals of technological opportunities

Collecting patent data

The proposed method requires a patent set to detect unusual patents in a given technology area, so in this step analysts gather patent data for analysis of a target technology area. To gather patents from patent databases, analysts can use patent retrieval queries including textual information, application dates and patent classification codes. After eliminating irrelevant patents, a final patent set for analysis can be prepared.

In patents, narrative sections including title, abstract, background summary, detailed description and claims can be used for textual analysis. Among various narrative sections, claims of patents are considered to be most important because they include the specific knowledge that requires legal protection (Fujii et al. 2007). Therefore the proposed method uses only claims to extract SAO structures. For syntactic analysis of patent text in the next step, collected patent data can be stored in electrical formats including Microsoft Excel files and Text files.

Analyzing syntactic structure of patent text

This step identifies SAO structures from patent text by exploiting NLP tools including the Stanford parser (Stanford 2010) and MiniPar (Lin 2010). The application programming interfaces of these tools allow identification of part-of-speech for each word in any grammatically correct sentence. After identifying noun phrases and verb phrases, SAO structures that have the grammatical relation of 'subject phrase-verb phrase-object phrase' can be extracted without human intervention. Otherwise, analysts can use commercial tools including Knowledgist2.5$^{\mathrm{TM}}$ (www.invention-machine.com) to extract SAO structures from patent text. Although many SAO structures can be extracted from each patent, some may be irrelevant or duplicated. Therefore this step filters these out by using a set of stopwords; a human expert performs additional screening. Finally, each patent can be represented as a set of SAO structures, which explicitly show the structural relationship among components in the relevant patent (Table 2). The set of SAO structures from each patent represents its uniqueness.

We used Knowledgist2.5$^{\mathrm{TM}}$ to extract SAO structures from patent text. After deleting stopwords including 'said', 'this' and 'above' from SAO structures, we eliminated irrelevant or duplicated SAO structures, then prepared 1366 SAO structures (average 9.18 per patent) for analysis.

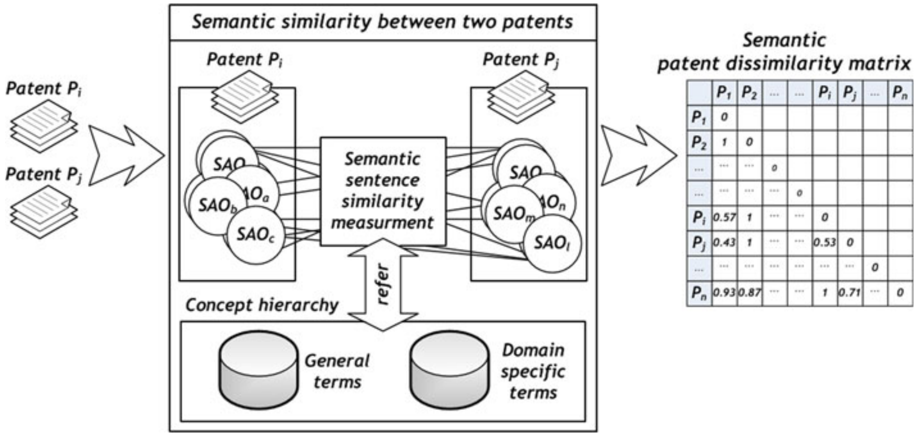
Measuring semantic similarities between patents

This step generates a semantic patent dissimilarity matrix by first measuring semantic patent similarity (Fig. 2). Because each patent contains many SAO structures which are the complete sentences, semantic patent similarities between patents can be computed using semantic sentence similarities between SAO structures of patents.

In the concept hierarchies that include 'is–a' relationships among concepts, semantic similarity between two sentences can be measured using a procedure that is composed of (1) tokenizing the sentences, (2) stemming words, (3) tagging parts of speech, (4)

Table 2 An example of SAO extraction (Patent US2002-468679)

S (subject)	A (action)	O (object)
dye-sensitized solar cell	contain	metal oxide fine particles
dispersion	contain	binder and metal oxide fine particles
action	bind to	fine particles and solvent
electrode	have	metal oxide
gas phase method	produce	metal oxide
metal oxide	contain	binder and metal oxide fine particles
metal oxide	contain	metal oxide
metal oxide fine particles	contain	titanium oxide
polymer compound	compose	binder
resin material	have	high transparency
titanium oxide	have	BET specific surface area of 10–100 m



	P_1	P_2	···	···	P_i	P_j	···	P_n
P_1	0							
P_2	1	0						
···	···	···	0					
···	···	···	···	0				
P_i	0.57	1	···	···	0			
P_j	0.43	1	···	···	0.53	0		
···	···	···	···	···	···	···	0	
P_n	0.93	0.87	···	···	1	0.71	···	0

Fig. 2 Concepts for generation of semantic patent dissimilarity matrices; semantic dissimilarities between pairs of patents are computed by measuring semantic sentence similarities of their SAO structures

determining the most likely sense of each word in each sentence, and (5) computing the similarity of the sentences based on the similarity of the pairs of words (Simpson and Dao 2005). A measure of similarity between two concepts in a concept hierarchy (Resnik 1999) is defined as follows:

$$\mathrm{sim}(c_1, c_2) = \frac{2 \times \mathrm{depth}\ (\mathrm{lcs}(c_1, c_2))}{\mathrm{depth}\ (c_1) + \mathrm{depth}\ (c_2)}, \tag{1}$$

where lcs is the lowest common subsumer of two concepts c_1 and c_2 in the concept hierarchy, and depth is the distance from a concept node c_i to the root of the concept hierarchy. The similarity of two concepts is $0 < \mathrm{sim}(c_1, c_2) \leq 1$, where 1 means that the two concepts are identical. A measure for semantic sentence similarity between two SAO structures SAO_i and SAO_j can be formulated by exploiting the matching average (Yoon and Kim 2011b; Simpson and Dao 2005):

$$\mathrm{MatAvg}(\mathrm{SAO}_1, \mathrm{SAO}_2) = \frac{2 \times \mathrm{Match}(\mathrm{SAO}_1, \mathrm{SAO}_2)}{|\mathrm{SAO}_1| + |\mathrm{SAO}_2|}, \tag{2}$$

where $|\mathrm{SAO}_i|$ is the number of set concepts in the relevant SAO structure, and Match(SAO_i, SAO_j) is the sum of similarity of the matching concepts (Eq. 1) between the sentences. The semantic similarity between two SAO structures is $0 < \mathrm{MatAvg}(\mathrm{SAO}_i, \mathrm{SAO}_j) \leq 1$, where 1 means that the two SAO structures are identical. The WordNet semantic dictionary (Miller 1995) can be used as a concept hierarchy for most terms, and user-defined concept hierarchies can be defined for domain specific terms. Although WordNet defines the concept hierarchies comprising synonyms, hypernyms and hyponyms by word meaning for most words, it does not contain abbreviations or terms that are very domain specific; for example, OPVC experts use specialized acronyms including 'PTCBI' for 'perylenetetracarboxylic bisbenzimidazole' and 'CVD' for 'chemical vapor deposition'. For this reason, we defined a synonym set that can be grouped by investigating the gathered SAO structures; the semantic sentence similarity measurement (Eq. 2) referred to this synonym set. To determine whether or not two SAO structures are identical, we use a threshold value p. Two SAO structures are considered to be semantically the same if the similarity score of two SAO structures is larger than p; for example, when $p = 0.90$ and Match(SAO$_i$, SAO$_j$) = 0.93, then the two SAO structures can be considered identical:

$$\mathrm{SAO}_{ij} = \begin{pmatrix} 1 & \text{if MatAvg}(\mathrm{SAO}_i, \mathrm{SAO}_j) \geq p \\ 0 & \text{otherwise} \end{pmatrix}. \tag{3}$$

By exploiting how many SAO structures the two patents share, a measure for semantic similarity between patents P_i and P_j can be simply formulated (Yoon and Kim 2011b):

$$\mathrm{SIM}(P_1, P_2) = \frac{2 \times N_{\mathrm{SAO}}(P_1, P_2)}{N_{\mathrm{SAO}}(P_1) + N_{\mathrm{SAO}}(P_2)}, \tag{4}$$

where $N_{\mathrm{SAO}}(P_i)$ is the number of SAO structures in patent P, and $N_{\mathrm{SAO}}(P_i, P_j)$ is the number of identical SAO structures that the patents share. Because duplicated SAO structures are eliminated in section "Analyzing syntactic structure of patent text", the semantic similarity score between patents is $0 < \mathrm{SIM}(A, B) \leq 1$, where 1 means that two patents are identical. Finally, semantic dissimilarity between two patents P_i and P_j can be directly computed:

$$\mathrm{DSIM}(P_1, P_2) = 1 - \mathrm{SIM}(P_1, P_2). \tag{5}$$

By computing the semantic dissimilarity between all pairs of patents (5), this step generates a patent dissimilarity matrix that encodes the technological distances between pairs of patents.

In this paper, the semantic patent similarities between all pairs of patents were computed by measuring semantic sentence similarities (the threshold value $p = 0.6$, 0.7, 0.8 and 0.9) between SAO structures of patents, so a semantic patent similarity matrix in the form of a lower triangular matrix was obtained (Table 3). Next, by calculating dissimilarities between patents, a 149 × 149 semantic patent dissimilarity matrices were obtained.

Mapping patents on a lower dimensional space

The patent dissimilarity matrix includes only semantic dissimilarities between all pairs of patents, so this step maps the patents on a lower dimensional space, and then identifies the coordinates of patents. For this, this step exploits MDS; MDS is the most proper among layout algorithms because patents consisting of a set of SAO structures do not contain feature dimensions. MDS is a set of related statistical techniques often used in information visualization for exploring similarities or dissimilarities in data. Currently many MDS algorithms are available, including PREFSCAL, PROXCAL and ALSCAL; also, various commercial statistical software tools including NetMiner3.0 (www.netminer.com) and SPSS 17.0 (www.spss.com) support these algorithms. The quality of MDS can be identified using the stress value V ($0 \leq V \leq 1$). In general, MDS with $V < 0.2$ is accepted to avoid degeneration (Kruskal 1964); the tolerance levels of stress are 'poor' ($0.2 \leq V < 0.4$), 'fair' ($0.1 \leq V < 0.2$), 'good' ($0.05 \leq V < 0.1$) and 'excellent' ($0.025 \leq V < 0.005$). By applying the patent dissimilarity matrix to the MDS algorithms, we can identify the coordinates of patents and they are used as data to quantify the outlierness of patents.

In this step, we used a patent dissimilarity matrix of $p = 0.8$ to choose a patent map for detecting outlier patents. To determine a proper threshold value p, we applied sensitivity analysis with respect to the results of patent mapping. In fact, when $p = 0.6$ or $p = 0.7$, too many SAO structures between two patents were computed identical, so patents were very evenly plotted on the patent map. Interestingly, when $p = 0.9$, we found that patents were

Table 3 A part of the semantic patent similarity matrix ($p = 0.8$); Blanks in the lower triangular matrix indicate that similarities between relevant patents are 0

	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	P26	P27
P1																											
P2																											
P3		0.1																									
P4		0.3	0.1																								
P5			0.4																								
P6			0.1		0.1																						
P7																											
P8			0.2																								
P9						0.1																					
P10	0.4																										
P11	0.2																										
P12			0.2																								
P13	0.1	0.1	0.1		0.1	0.1	0.1																				
P14																											
P15	0.1		0.1			0.2	0.1			0.1	0.1		0.1														
P16							0.1								0.1												
P17						0.1							0.1														
P18	0.2		0.1												0.1												
P19		0.1	0.1			0.1							0.1														
P20																	0.4										
P21	0.1	0.1		0.1			0.1				0.1		0.2	0.6	0.1	0.1											
P22	0.1	0.1		0.1		0.1	0.1				0.1		0.3	0.1	0.1	0.3	0.5	0.2	0.1	0.1	0.2						
P23	0.2		0.1					0.2							0.1	0.1					0.1						
P24				0.1			0.2			0.1					0.2	0.1	0.3		0.1	0.1	0.1	0.1	0.1				
P25			0.1	0.1	0.1	0.1							0.3		0.1		0.1	0.1			0.1	0.3		0.1			
P26						0.1											0.1					0.1					
P27		0.1	0.3	0.1	0.2	0.3					0.1		0.3		0.2		0.1				0.2	0.1	0.2		0.1		
P28		0.1	0.2		0.1			0.1				0.2				0.1			0.1				0.1	0.1		0.1	0.1

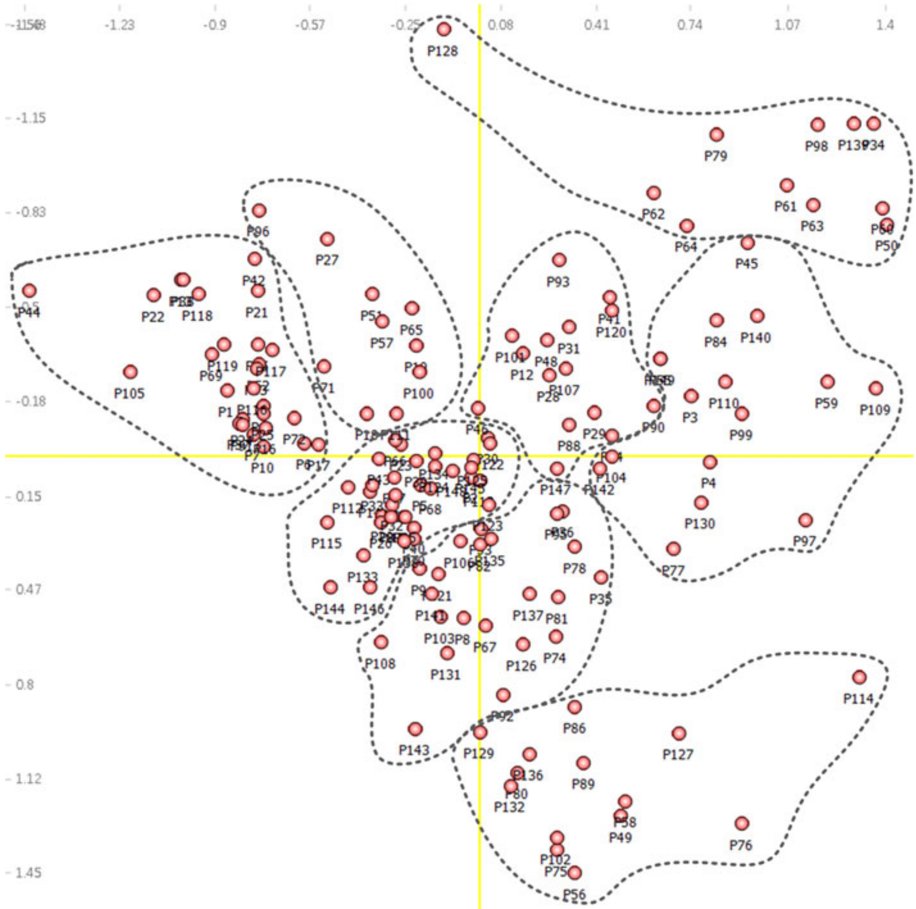


Fig. 3 OPVC-related patent map; nodes = 149, dimension = 2, stress value = 0.157

also evenly distributed on the map. The large threshold value p tightly computes similarity between two SAO structures, so most SAO structures between two patents were determined to be different. Therefore, we identified that $p = 0.9$ made patents evenly distributed on the patent map. Through this sensitivity analysis, we finally concluded that a patent map of $p = 0.8$ showed the best mapping result in the case study.

Using the dissimilarity matrix of $p = 0.8$, we mapped the patents on a two-dimensional (2D) space with 'fair' MDS quality of stress value $V = 0.157$ by exploiting ALSCAL of NetMiner3.0 (Fig. 3); therefore, we could obtain coordinates of each patent on the space.

Measuring outlierness of patents

Visualizing patents on a lower dimensional space provides an overall understanding of the given patent set. However, as a patent set becomes larger, the identification of outlier is likely to become increasingly more complex and subjective. Therefore, this step uses distance-based, density-based and clustering-based outlier detection to quantitatively identify outlier patents, which are far from the others on the 2D space produced using MDS.

For distance-based outlier detection, we first use k nearest neighbors (k-NNs) which are the k objects closest to a given object in a space. The distance-based outlierness of patents can be defined as the average distance of patent x from its k-NNs using the coordinates of patents:

$$\mathrm{Outlierness}_{\mathrm{DIST}}(x,k) = \frac{\sum_{y\in N(x,k)} \mathrm{distance}(x,y)}{|N(x,k)|}, \qquad (6)$$

where k is the number of k-NNs adjacent to patent x for outlier detection, $N(x, k)$ is the set of k-NNs of patent x, $|\cdot|$ is the size of a set, distance(x, y) is the distance between two patents x and y. Patents with a large distance-based outlierness are considered to be distinct from the others.

In density-based outlier detection, outlier patents are those which are located in or near the low density areas. The density-based outlierness of a patent can formulated using the relative density of a set that is composed of k-NNs adjacent to a patent:

$$\mathrm{Outlierness}_{\mathrm{DENS}}(x,k) = \frac{[\mathrm{Outlierness}_{\mathrm{DIST}}(x,k)]^{-1}}{\sum_{y\in N(x,k)} [\mathrm{Outlierness}_{\mathrm{DIST}}(y,k)]^{-1} \Big/ |N(x,k)|}, \qquad (7)$$

where k is the number of k-NNs adjacent to patent x for outlier detection, $N(x, k)$ is the set of k-NNs of patent x, $|\cdot|$ is the size of a set. Because patents with a small density-based outlierness are located in or near the relatively low density areas, they can be considered to be distinct from other patents.

In clustering-based outlier detection, outlier patents are those which do not belong strongly to their clusters. Because patents have their coordinates in a lower space, we can identify patent clusters using clustering algorithms. To identify how strongly a patent belongs to its cluster, we measure the relative distance between the patent and the centroid of its cluster using k-means clustering. Therefore, the clustering-based outlierness of patents is defined as follows:

$$\mathrm{Outlierness}_{\mathrm{CLUS}}(x) = \frac{\mathrm{distance}(x,\ \mathrm{centroid}(C(x)))}{\sum_{y\in C(x)} \mathrm{distance}(y,\ \mathrm{centroid}(C(x))) \Big/ |C(x)|}, \qquad (8)$$

where $C(x)$ is the set of patents in the patent cluster of patent x, $|\cdot|$ is the size of a set, centroid is the coordinates of the centroid of a patent cluster, distance(x, y) is the Euclidean distance between two objects x and y on a space. Because patents with a large clustering-based outlierness are located relatively far from its cluster than other patents in the cluster, they can be considered to be distinct from other patents.

We measured outlierness of patents using the three outlier detection methods (Eqs. 6, 7, 8) (Table 4). Because each outlier detection method has both strengths and weakness (Hodge and Austin 2004), of the 20 patents that were ranked as outliers by each detection techniques, we considered only the eight that were identified as outlier patents by all methods: P44, P45, P76, P96, P109, P114, P128 and P143. In the proposed procedure based on semantic structural patent similarities, they are considered distinct from other patents.

These outlier patents have a strong possibility of being unusual types of inventions because they are distinctive from the view of structural similarity. However, because not

Table 4 Patents high-ranked by three outlier detection techniques

Outlier patents by (6) (k-NNs = 6)		Outlier patents by (7) (k-NNs = 6)		Outlier patents by (8) (# of clusters = 8)	
Label (patent #)	Outlierness	Label (patent #)	Outlierness	Label (patent #)	Outlierness
P128 (7,655,860)[a]	0.912	P128 (7,655,860)[a]	0.294	P128 (7,655,860)[a]	6.752
P114 (7,569,728)[a]	0.727	P44 (6,166,320)[a]	0.334	P114 (7,569,728)[a]	4.034
P44 (6,166,320)[a]	0.536	P115 (7,851,699)	0.457	P44 (6,166,320)[a]	3.077
P76 (7,632,576)[a]	0.506	P105 (7,642,449)	0.460	P45 (6,555,840)[a]	2.842
P109 (7,626,117)[a]	0.457	P114 (7,569,728)[a]	0.470	P96 (7,888,584)[a]	2.169
P97 (7,411,223)	0.443	P46 (6,391,471)	0.474	P76 (7,632,576)[a]	2.151
P127 (7,763,727)	0.378	P76 (7,632,576)[a]	0.478	P144 (7,799,989)	2.013
P59 (6,614,057)	0.350	P71 (6,747,203)	0.492	P109 (7,626,117)[a]	1.816
P62 (6,664,137)	0.337	P144 (7,799,989)	0.526	P143 (7,524,367)[a]	1.778
P79 (7,074,501)	0.335	P133 (7,626,115)	0.532	P62 (6,664,137)	1.775
P105 (7,642,449)	0.324	P112 (7,638,706)	0.605	P146 (7,884,218)	1.773
P143 (7,524,367)[a]	0.324	P1 (4,009,054)	0.608	P42 (6,274,804)	1.570
P77 (7,592,074)	0.312	P93 (7,612,367)	0.609	P147 (7,830,584)	1.568
P50 (6,537,688)	0.308	P45 (6,555,840)[a]	0.620	P6 (4,328,389)	1.518
P27 (5,385,615)	0.296	P143 (7,524,367)[a]	0.626	P46 (6,391,471)	1.503
P96 (7,888,584)[a]	0.296	P10 (4,795,501)	0.629	P129 (7,592,540)	1.502
P64 (6,657,378)	0.295	P109 (7,626,117)[a]	0.629	P97 (7,411,223)	1.485
P140 (7,528,003)	0.294	P96 (7,888,584)[a]	0.630	P108 (7,820,908)	1.483
P45 (6,555,840)[a]	0.292	P127 (7,763,727)	0.633	P64 (6,657,378)	1.470
P34 (6,447,879)	0.273	P108 (7,820,908)	0.641	P22 (5,123,968)	1.469

OPVC-related patent map (Fig. 3) showed the best clustering result when # of clusters = 8

[a] Outlier identified by all methods

all outlier patents deliver new technological opportunities, the quantitative procedure of this paper requires a final review process by humans. After investigating all the identified outlier patents, we observed some implications for potential technological development from P45, P96, and P143.

P45 that (granted in 2003) is a method of fabricating charge-transport structures by printing low molar mass dyes on a substrate to form a charge-transport polymer layer. The conventional method of fabricating such structures in OPVCs has been vapor deposition in which a substrate is exposed to one or more volatile precursors, which react or decompose, or both, on the substrate surface to produce the desired deposit. However, P45 presents a method of using ink printed on an exposed surface to form a multicolor pattern that diffuses into the layer to form charge-recombination and emission regions within the layer. This indicates that P45 suggested a quite different method from the previous and later patents. Because P45 was granted in 2003, it is an early technology related to OPVCs. Although P45 is relatively old in the patent set, the number of its citations (48 times) was very high; 62.5% of the citations have been made for the last 3 years and 16.7% of the citations had been made for the first 3 years. Considering average citations of patents (8.6 times), P45 has been cited much more than other patents, and accordingly seems to be an important patent that has been more applied by other later patents in other technological domains including semiconductor technology and display technology; for example, the patent was cited by semiconductor-related patents including organic logic circuit elements (7589553, 7678857 and 7875975) and display technology-related patents including organic light emitting elements (7442963, 7598519 and 7939835). In fact, the patent has the potential in many applications including flexible solar cell products and design of future buildings because its method allows electroplating on a flexible substrate by exploiting roll-to-roll processing. Therefore, we concluded that P45 with this novel technological method could be an outlier patent in the gathered OPVCs-related patent set.

P96 (granted in 2011) is an invention about ‘solar networks and power grids’; it uses a large number of solar cells that can be configured to have the appearance of natural foliage including a palm tree, a deciduous tree and an evergreen tree. Such a network of solar arrays can extend unobtrusively for many miles alongside roads, high-ways, railways, or canals, and can further include means for storing and transmitting electric power. Furthermore, the network of solar arrays can be connected to recharging stations for use by electric and hybrid transportation vehicles, so ideas in this patent can be used as a basis for a solar power grid. The patent could not be cited by other later patents because it was granted in 2011. However, most patents in the patent set propose technological methods or materials, while interestingly P45 presents a novel solar energy environment as a business application. Although this patent focuses on the future potential application rather than on the development of original technology, the creative idea of this invention was considered to have future business potential in environment-friendly urban design and construction.

P143 (granted in 2009) describes use of a ‘poly cross linked phthalocyanine compound’, which absorbs near-infrared (NIR) light (wavelengths $\sim$750 to $\sim$1,100 nm). Conventional compounds for OPVCs are usually sensitized to visible light, so most OPVCs can transform solar energy into electricity only during the day. However, the high NIR-absorption efficiency of poly cross linked phthalocyanine means that solar modules based on this invention allow electricity production both during the day using infrared rays from the sun and during the night using infrared rays from the environment. Therefore, it is expected that this new concept will increase solar energy conversion by a factor of two compared to solar energy conversion using conventional solar cells. In sum, the method of P143 was considered to be a technological quantum jump by employing a new compound that can be differentiated from existing and later patents, and thus we considers the patent as an outlier patent.

To test the validity of the proposed method, we investigated novelty scores of outlier patents and randomly selected patents. To this end, we picked three subsets with eight randomly selected patents, and subjected those sample sets to the same human judgment process that was used with the outliers. To avoid potential bias, those performing the judgments were not told that these are random samples, but the exact same protocol and background were used with these random samples. The “novelty” of the patents was examined from the aspects of methods (or materials) and applications (or uses). In this paper, novelty score of methods or materials ($\mathrm{NS_M}$) of a patent indicates how differentiated the patent is from other patents in proposing methods or using materials; novelty score of application ($\mathrm{SN_A}$) of a patent shows how differentiated the patent is from others in exploiting its relevant technology. As each novelty score approaches 10, the patent is more differentiated from the others in the given patent set. If total novelty score ($\mathrm{NS_M} + \mathrm{NS_A}$) of a patent is larger than 10, this paper considered the patent highly novel in an overall sense. With respect to the outlier patent set and the three randomly selected patent sets, $\mathrm{NS_M}$ and $\mathrm{NS_A}$ ranging from 0 to 10 were evaluated by the review process of two experts including a patent attorney and a domain expert (Table 5). In brief, average total novelty score of outlier patents (11.63) was larger than average total novelty score of random

Table 5 Novelty scores of outlier patents and randomly selected patents by review process of experts

Outlier patents				Random set A				Random set B				Random set C			
P#	NS_M	NS_A	$NS_M + NS_A$	P#	NS_M	NS_A	$NS_M + NS_A$	P#	NS_M	NS_A	$NS_M + NS_A$	P#	NS_M	NS_A	$NS_M + NS_A$
P44	5.00	3.50	8.50	P97[a]	6.00	5.50	11.50	P39	3.50	2.75	6.25	P44	5.00	3.50	8.50
P45[a]	9.00	6.00	15.00	P23	4.25	3.50	7.75	P91[a]	8.50	7.25	15.75	P29	4.00	4.25	8.25
P76[a]	7.25	3.50	10.75	P81	5.75	3.50	9.25	P101	6.25	3.75	10.00	P75	4.00	3.00	7.00
P96[a]	4.50	9.00	13.50	P110[a]	5.00	6.75	11.75	P133[a]	7.50	4.00	11.50	P66	3.00	6.00	9.00
P109	6.00	3.75	9.75	P14	4.00	2.50	6.50	P122	5.75	3.00	8.75	P89	5.50	3.25	8.75
P114[a]	7.00	3.75	10.75	P56	4.75	3.00	7.75	P88	3.50	5.50	9.00	P129	4.50	4.00	8.50
P128	6.00	3.00	9.00	P92	5.50	3.00	8.50	P96[a]	4.50	9.00	13.50	P39	3.50	3.50	7.00
P143[a]	9.50	6.25	15.75	P29	3.50	2.75	6.25	P18	4.00	4.00	8.00	P105[a]	4.00	7.25	11.25
Avg.	6.78	4.84	11.63	Avg.	4.84	3.81	8.66	Avg.	5.44	4.91	10.34	Avg.	4.19	4.34	8.53

Novelty score of methods or materials (NS_M): 0–10, novelty score of applications (NS_A): 0–10

[a] Patents are highly novel patents of which $NS_M + NS_A$ is larger than 10

patent sets (9.17). In addition, 62.5% of the outlier patents were considered highly novel because total novelty scores of 5 patents were larger than 10, while 25% of the randomly selected patent sets were identified to be highly novel. Therefore, we could conclude that the outlier patents in an overall sense were more novel than non-outlier patents.

Although not all outlier patents deliver new approaches to technological development, some of them provide fresh or unusual signals for further technological development. In the competitive technological environment, early grasp of potential technological opportunities is important to develop technological items that can increase the competitiveness of a business competition. Therefore, the automated and quantitative method proposed in this paper will assist experts including researchers and R&D policy makers to explore technological opportunities for technology strategy formulation and product development planning.

Summary and discussions

In the recent technological environment, technological quantum jumps are especially considered more market-disruptive for technological competitiveness than incremental improvements (Christensen and Leslie 1997). In response, this paper was aimed at early and quantitatively identifying the potential patents that may act as technological jumps. To this end, this paper presented a new method of detecting patents that can act as signals of new technological opportunities by combining SAO-based patent analysis and outlier detection. SAO structures are the syntactically ordered sentences composed of 'subject', 'action' and 'object'. Therefore, SAO structures can be automatically extracted by syntactic analysis of patent text; they are considered to be key findings of inventions or expertise of inventors because they explicitly represent the structural relationships among technological components in the relevant patents. As a method to identify extreme observations from the rest of a group, this paper adopted outlier detection techniques to identify distinctive patents in a given technology area. The fundamental of the proposed method assumes method (materials) or applications (uses) presented by outlier patents may have a strong possibility of being radical innovation. The procedure for identifying technological opportunities is composed of (1) collecting patent data, (2) analyzing syntactic structure of patent text by exploiting NLP, (3) generating a semantic patent dissimilarity matrix that encodes technological dissimilarities of all patent pairs by exploiting semantic sentence similarities of SAO structures, (4) mapping patents onto a lower dimensional space, and (5) identifying outlier patents from a given patent set by exploiting distance-, density-, and clustering-based outlier detection methods. The proposed method was illustrated using OPVC-related patents.

Recently, new patents are being increasingly generated, so the early detection of outlier patents that can be achieved by the proposed method will help experts detect chances to develop new types of intellectual properties that are different from existing inventions. Specifically, building on semantic structural similarities of patents, the proposed method conducted patent mapping and outlier detection in an automated way. Furthermore, previous studies for technological opportunities mainly addressed technological directions or technological areas from a broad view that are promising and have not been unexplored. Differently from the previous studies, this paper proposed a new method to detect patents that can act as fresh or unusual signals for further technological development by exploiting semantic structural dissimilarity of patents. Therefore, as a complementary approach of the previous studies, we expect that the proposed method can be incorporated into the strategic technology planning process for exploration of new potential technological opportunities.

Despite those advantages, some research challenges remain. First, although a semantic dictionary for general concepts, WordNet, was exploited for measuring semantic patent similarities, we should have used a separate synonym set because WordNet does not include very specialized terms including abbreviations. Therefore, further research will measure semantic patent similarities by defining domain-specific concept hierarchies. Second, the final step of the proposed method requires a review process by human experts to determine whether outlier patents can deliver chances for new technological opportunities or not. To reduce the amount of human effort, a future topic will be to investigate a way to identify technological importance of outlier patents by incorporating patent bibliographic information including citations and patent family relations. Third, this paper individually measured outlierness of patents with respect to each outlierness measure and the method considered patents highly ranked by all the measures as outlier patents. However, a future research will investigate an integrated outlierness index by incorporating other outlierness measures. Finally, this paper only applied the proposed method to OPVC-related patents although the method could identify some distinctive patents for technological opportunities. Therefore, future research will apply the method to other technological fields to more identify the practicality.

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

References

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.

Aleskerov, E., Freisleben, B., & Rao, B. (2002). Cardwatch: A neural network based database mining system for credit card fraud detection. In *IEEE* (pp. 220–226).

Altschuller, G. (1984). *Creativity as an exact science: The theory of the solution of inventive problems*. New York: Gordon and Breach.

Barnett, V., Lewis, T., & Abeles, F. (1979). Outliers in statistical data. *Physics Today, 32*, 73.

Bergmann, I., Butzke, D., Walter, L., Fuerste, J., Moehrle, M., & Erdmann, V. (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.

Cascini, G., Fantechi, A., & Spinicci, E. (2004). Natural language processing of patents and technical documentation. *Document Analysis Systems, VI*, 508–520.

Cascini, G., Russo, D., & Zini, M. (2007). Computer-aided patent analysis: Finding invention peculiarities. In N. Leon-Rovira (Ed.), *Trends in computer aided innovation* (pp. 167–178). Boston: Springer.

Cascini, G., & Zini, M. (2008). Measuring patent similarity by comparing inventions functional trees. *IFIP International Federation for Information Processing, 277*, 31–42.

Chandola, V., Banerjee, A., & Kumar, V. (2009). Anomaly detection: A survey. *ACM Computing Surveys (CSUR), 41*(3), 1–58.

Choi, S., Lim, J., Yoon, J., & Kim, K. (2010). Patent function network analysis: A function based approach for analyzing patent information. In *IAMOT2010*, Cairo, Egypt.

Christensen, C., & Leslie, D. (1997). *The innovator's dilemma*. Boston: Harvard Business School Press.

Franses, P., Kloek, T., & Lucas, A. (1998). Outlier robust analysis of long-run marketing effects for weekly scanning data. *Journal of Econometrics, 89*(1–2), 293–315.

Fujii, A., Iwayama, M., & Kando, N. (2007). Introduction to the special issue on patent processing. *Information Processing & Management, 43*(5), 1149–1153.

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.

Hodge, V., & Austin, J. (2004). A survey of outlier detection methodologies. *Artificial Intelligence Review, 22*(2), 85–126.

Karki, M. (1997). Patent citation analysis: A policy analysis tool. *World Patent Information, 19*(4), 269–272.

Kruskal, J. (1964). Nonmetric multidimensional scaling: A numerical method. *Psychometrika, 29*(2), 115–129.

Lee, S., Yoon, B., & Park, Y. (2009). An approach to discovering new technology opportunities: Keyword-based patent map approach. *Technovation, 29*(6–7), 481–497.

Leung, K., & Leckie, C. (2005). *Unsupervised anomaly detection in network intrusion detection using clusters* (pp. 333–342). Sydney: Australian Computer Society, Inc.

Lin, D. (2010). Minipar. http://webdocs.cs.ualberta.ca/~lindek/minipar.htm. Accessed 1 Oct 2011.

Mann, D. (2002). *Hands-on systematic innovation*. Leper: CREAX Press.

Mann, D. (2003). Better technology forecasting using systematic innovation methods. *Technological Forecasting and Social Change, 70*(8), 779–795.

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

Moehrle, M., & Geritz, A. (2004). Developing acquisition strategies based on patent maps. In *Proceedings of the 13th international conference on management of technology* (pp. 1–9), Washington, DC, USA.

Moehrle, M., 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.

Mogee, M., & Kolar, R. (1994). International patent analysis as a tool for corporate technology analysis and planning. *Technology Analysis & Strategic Management, 6*(4), 485–504.

Narin, F. (1994). Patent bibliometrics. *Scientometrics, 30*(1), 147–155.

Park, B. (2002). An outlier robust GARCH model and forecasting volatility of exchange rate returns. *Journal of Forecasting, 21*(5), 381–393.

Radauer, A., & Walter, L. (2010). Elements of good practice for providers of publicly funded patent information services for SMEs-selected and amended results of a benchmarking exercise. *World Patent Information, 32*(3), 237–245.

Resnik, P. (1999). Semantic similarity in a taxonomy: An information-based measure and its application to problems of ambiguity in natural language. *Journal of Artificial Intelligence Research, 11*(95), 130.

Savransky, S. (2000). *Engineering of creativity: Introduction to TRIZ methodology of inventive problem solving*. Boca Raton: CRC.

Schuh, G., & Grawatsch, M. (2004). TRIZ-based technology intelligence. *In Proceedings of the 13th international conference on management of technology*, Washington, DC, USA.

Sekar, R., Gupta, A., Frullo, J., Shanbhag, T., Tiwari, A., Yang, H., et al. (2002). Specification-based anomaly detection: A new approach for detecting network intrusions. *In Proceedings of the 9th ACM conference on computer and communications security*, New York, USA.

Simpson, T., & Dao, T. (2005). Wordnet-based semantic similarity measurement. http://www.codeproject.com/KB/string/semanticsimilaritywordnet.aspx. Accessed 1 Oct 2011.

Siris, V., & Papagalou, F. (2006). Application of anomaly detection algorithms for detecting SYN flooding attacks. *Computer Communications, 29*(9), 1433–1442.

Stanford. (2010). The Stanford parser: A statistical parser. http://nlp.stanford.edu/software/lex-parser.shtml. Accessed 1 Oct 2011.

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., 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.

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, 88*(1), 213–228.

Yoon, B., & Park, Y. (2004). Morphology analysis approach for technology forecasting. In *2004 IEEE international engineering management conference* (Vol. 2, pp. 566–570), Singapore.

Yoon, B., & Park, Y. (2005). A systematic approach for identifying technology opportunities: Keyword-based morphology analysis. *Technological Forecasting and Social Change, 72*(2), 145–160.

Yoon, B., Yoon, C., & Park, Y. (2002). On the development and application of a self-organizing feature map-based patent map. *R&D Management, 32*(4), 291–300.