Technological Forecasting & Social Change xxx (2016) xxx–xxx

Contents lists available at ScienceDirect

Technological Forecasting & Social Change

Monitoring emerging technologies for technology planning using technical keyword based analysis from patent data

Junegak Joung, Kwangsoo Kim *

Department of Industrial and Management Engineering, Pohang University of Science and Technology, 77 Cheongam-ro, Nam-gu, Pohang 790-784, Republic of Korea

ARTICLE INFO

Article history:
Received 3 February 2016
Received in revised form 7 July 2016
Accepted 21 August 2016
Available online xxxx

Keywords:
Patent analysis
Keyword-based model
Technical keyword
Technology planning

ABSTRACT

This paper proposes technical keyword-based analysis of patents to monitor emerging technologies, and uses a keyword-based model in contents-based patent analysis. This study also presents methods to automatically select keywords and to identify the relatedness among them. After using text-mining tools and techniques to identify technical keywords, a technical keyword-context matrix is constructed. The relatedness between pairs of keywords is then identified in a transformation of this matrix. Patent documents are clustered by using a hierarchical clustering algorithm based on patent document vectors. As a result, emerging technologies can be monitored by identifying clusters composed of technical keywords. A case study of mechanisms of electron transfer in electrochemical glucose biosensors is given to demonstrate how the proposed method can monitor emerging technologies.

© 2016 Elsevier Inc. All rights reserved.

1. Introduction

Monitoring of emerging technologies can identify incipient technological changes quickly, and is an invaluable component of technology planning, and of development of research and development (R&D) policy by governments and companies (Ashton et al., 1991). By investing in R&D strategically in potentially-important emerging technologies, companies can become market winners or early followers of market leaders (Hamilton, 1985). Given that technology monitoring can help companies' development of new products, technologies or joint ventures, monitoring of emerging technologies provides a starting point to induce radical technological change or mergers and acquisitions (M&A) (Ashton et al., 1994). Furthermore, many methods in technology forecasting predict emerging technologies, but have limited ability to identify possible emerging technologies. Therefore, use of reliable sources (e.g., research organization reports) to comprehend emerging technologies allows examination of how they are specifically realized. To constantly monitor emergence of technologies, patents are the best source of technology information, because they contain technical details. Patent analysis has been considered as a basis for technology assessment to monitor sources of technological knowledge (Ernst, 2003).

Patent documents describe commercialized inventions, and the number of granted patents represents a company's rate of technological advance (Ernst, 2001). For these reasons, analysis of patent data has two major benefits. First, patents are written to protect the right of invention and to prevent overlapping R&D investment, so they are reliable formal sources of unique technical information. Second, patent database systems are well-organized and are supplied with retrieval systems, so large numbers of patents can be examined easily. These systems present up-to-date information, and can rapidly check new applied or granted patents and their applications; moreover, patent information is available to everyone. For example, since 1976, the United States Patent and Trademark Office (USPTO) has provided full-text patents and a retrieval system that can be used free of charge by anyone. For these reasons, stakeholders such as R&D policy makers, R&D managers, technology developers, and R&D planners have used patent information to support identification of world-wide technical evolution (Zhang, 2011; Altuntas et al., 2015b), and to support making in R&D (Thorleuchter et al., 2010; Altuntas et al., 2015a). As a result, users can examine technological trends in which competitors' R&D strategies, technological resources, and external knowledge of technology are embedded.

Many techniques have been used in patent information analysis to monitor technological trends. Engelsman and van Raan (1992) suggested co-word maps to identify declining or emerging fields of technological activities, which were identified by examining keywords at the meso-level and micro-level. Yoon et al. (2002) proposed patent maps that displayed patents in two or three-dimensional space according to similarity of keywords; an absence of patents in the map was considered as a starting point to identify emerging technologies. Yoon and Park (2004) developed keyword-based patent network and identified up-to-date trends of high technologies by using network analysis. In addition, the k-Means algorithm (Kim et al., 2008), a formal concept analysis-based approach (Lee et al., 2011), and a novelty detection technique (Geum et al., 2013) have been used in keyword-patent matrices

* Corresponding author.
E-mail address: kskim@postech.ac.kr (K. Kim).

http://dx.doi.org/10.1016/j.techfore.2016.08.020
0040-1625/© 2016 Elsevier Inc. All rights reserved.

to capture technological flows and emerging patterns. More-advanced techniques have been described, such as an approach based on subject-action-object (SAO) relationships. Choi et al. (2011) suggested using an SAO-based network to identify technology trends; the authors constructing a noun-by-verb relationship matrix, so emerging technologies could be inferred by low density and high cohesion in the network. Wang et al. (2015) conducted SAO-based technology roadmapping to comprehend developing trends.

However, previous research to monitor technological trends has some limitations in the process of keyword selection and in identification of relatedness of keywords that all keyword-based patent analysis methods share. These processes rely too much on the intervention of experts, so the reliability of the analysis can be greatly affected by the experts' opinions. Expert-dependent analysis is also expensive because it is time-consuming and laborious. Keyword selection is the most crucial factor, but the main keywords are chosen based on the subjective judgment of experts (Tseng et al., 2007; Noh et al., 2015). Although some research (Yoon and Park, 2004; Lee et al., 2009; Geum et al., 2013) suggested term frequency to guide experts in evaluating significant terms, experts must still take time to eliminate common words. Moreover, previous extraction of keywords could not effectively analyze multiple-phrase words even if they articulated a patent document well. Identification of the relatedness of keywords as synonyms, hypernyms, and hyponyms raises the quality of semantic processing when comparing patent documents, but assessment of their relatedness has been completely dependent on technical experts. Choi et al. (2011) and Wang et al. (2015) proposed using a word ontology such as WordNet (Miller, 1995) to understand the relatedness of keywords, but WordNet uses only a hierarchy of generic terms, so it does not consider technical terms; therefore, it does not produce a good solution because patents include many technical terms.

To summarize, the selection of keywords and identification of their relatedness are important to provide reliable and objective results of patent analysis, but current algorithm are not sufficient to assist experts to select significant keywords or to grasp the relatedness among them (Joung and Kim, 2015). To reduce these limitations, measures that reduce dependence on experts are required.

Therefore, this paper suggests a reliable way to monitor emerging technologies. This method uses analysis of technical keywords extracted from patents, and improves on the keyword-based patent analysis. Technical keyword based analysis allows monitoring of whether patents embody emerging technologies. For this purpose, methods to choose technical keywords and to be aware of their relatedness are presented. Technical keywords are extracted using a commercial NLP software package (*Alchemy*™), then selected using a term frequency-inverse document frequency (TF-IDF) function. The relatedness between pairs of technical keywords is established by distributional similarity. Next, a dissimilarity matrix is built as the complement of the similarity of patent documents that was identified on the basis of the relatedness of keywords. Then a hierarchical clustering algorithm clusters the patents by considering the relationships encoded by this dissimilarity matrix. Emerging technologies can be identified by monitoring clusters composed of technical keywords. To demonstrate the validity of the proposed approach, a case study of mechanisms of electron transfer in electrochemical glucose biosensors is given.

The rest of this paper is organized as follows. Section 2 briefly provides theoretical background of proposed methodology. Section 3 explains the proposed methodology. Section 4 describes a case study of mechanisms of electron transfer in electrochemical glucose biosensors to apply the proposed approach. Section 5 provides conclusions and future work.

2. Theoretical background

2.1. Keyword-based patent analysis to identify emerging technologies

Keyword-based patent analysis has been applied in various techniques such as text mining, data reduction, clustering, and network analysis to identify emerging technologies. Engelsman and van Raan (1992) provided a co-word map based on keywords extracted by experts, and used it to visualize developments in fields of technology by using co-occurrences. Similarly, Courtial et al. (1993) identified emerging technologies by using degree of density and centrality in co-word analysis. Zhu and Porter (2002) offered automatic extraction of keywords by using text mining, and developed a commercial patent-analysis program, *VantagePoint*™. Given that text mining could extract keywords automatically, Yoon et al. (2002) proposed a self-organizing feature-map-based patent map that uses a keyword-patent matrix to produce simplified images of multi-dimensional patent data, and to visualize the dynamic pattern of technological advancement. Kim et al. (2008) developed a patent map to visualize emerging technologies and to anticipate future technological trends on the basis of a semantic network of keywords by clustering patent documents that share keywords. Lee et al. (2011) suggested a concept-analysis-based approach to monitor technological changes that organizes objects with shared properties based on a keyword-patent matrix. Geum et al. (2013) also applied novelty detection techniques that identify new or unusual data that existing systems did not perceive, and detected new and emerging pattern of patents. These existing papers recommended term frequency to assist experts to select keywords, but experts were still required to remove common words; this is also time-consuming work (Yoon et al., 2002; Lee et al., 2011; Geum et al., 2013).

In addition to these studies, Yoon and Park (2004) proposed a text mining-based patent network to analyze up-to-date trends of high technologies by calculating the distance between patents on the basis of a keyword-patent matrix. They then discovered emerging technologies by developing new network analysis indexes (centrality index, cycle index) and keyword clusters. Similarly, Chang et al. (2010) concentrated on cluster network analysis to identify key technologies in carbon nanotube field-emission displays. Zhang et al. (2014) provided term clumping that includes automatic keyword selection by exploiting term frequency-inverse document frequency (TF-IDF), but did not consider many variants of TF and IDF.

To improve the efficiency of keyword-based patent analysis, Yoon et al. (2011) suggested property-function that is composed of 'adjective + keyword' and 'verb + keyword' that are extracted using the Stanford dependency parser, and built a network based on co-occurrence matrix. Emerging properties and functions were then identified by analyzing small and highly dense sub-networks. Similarly, Yoon and Kim (2011) constructed a network on the basis of similarity matrix that compared patents by using SAO structures extracted by *Knowledgist*™, and in this way detected clusters that include up-to-date technology. Choi et al. (2011) built a network composed of noun nodes and verb nodes based on a noun-by-verb matrix, and identified emerging technologies by using density and a cohesion index to analyze sub-networks. Wang et al. (2015) proposed technology roadmapping based on SAO analysis to identify technology development trends and future directions of the technology domain. These studies suggested a thesaurus called WordNet to perform semantic processing while comparing patent documents (Yoon and Kim, 2011; Choi et al., 2011; Wang et al., 2015), but WordNet uses a glossary of generic terms and therefore cannot take technical terms into account.

Although previous research has been useful to look into emerging technologies, it has limitations, including too much dependence on expert intervention during keyword selection and identification of the relatedness of keywords. Consequently, improvement of keyword-based patent analysis requires algorithms to better secure experts' objectivity and reliability. For these reasons, algorithms to guide selection of technical keywords and to identify relatedness among them will be considered next.

2.2. Technical keywords & their relatedness

Development of text mining has resulted in ways to extract keywords automatically, so contents analysis has become possible; this process can analyze unstructured text data. The extracted keywords should represent the contents of a patent document, and the relatedness of keywords is crucial to successful sematic processing to calculate the similarities among patent documents. Several methods have been proposed to automatically extract technical keywords and to identify their relatedness.

Construction of algorithms to extract technical keywords is complicated by the fact that they are embedded in phrases, which are difficult to extract due to vagueness of their lexical boundaries (Tseng et al., 2007). To solve this problem, many software packages to extract keywords have been developed. One keyword extraction tool, *Alchemy*™ provides keyword extraction and other text analysis tools. In a comparison using documents from biomedicine, domain experts identified *Alchemy*™ as the most promising keyword extraction tool (Ramirez et al., 2010). *Alchemy*™ has been used to extract core terminology, which was then used to build a domain ontology (Missikoff et al., 2015).

Other techniques can be used to extract keywords and to evaluate extracted keywords. Term frequency-inverse document frequency (TF-IDF) functions can be used to evaluate crucial terms or to extract them among many words. The TF-IDF concept has been commonly used to weight words in information retrieval to compare user queries with documents (Niwa and Sakurai, 1999; Robertson, 2004). A term with high TF is significant; although the significance does not increase linearly with counts, an extremely high TF indicates that the term is becoming common. To alleviate the drawbacks of TF, IDF has been used to emphasize rare terms (Sparck Jones, 1972). Therefore, TF-IDF can identify important terms (Chen et al., 2008; Li et al., 2009). Application of TF-IDF functions can identify crucial terms that are not peripheral.

To identify the relatedness of terms, many proposed approaches exploit semantic-processing methods developed in artificial intelligence and computational linguistics; these methods are thesaurus-based and consider distributional similarity. Word similarity in thesaurus-based methods can be measured by a distance between pairs of word nodes, if a hierarchical knowledge base has already been established. For instance, WordNet is a large hierarchical generic database of English words; it can be used in semantic processing or for other purposes of lexical taxonomy (Miller, 1995). Many similarity measures in the thesaurus can used to estimate the distance between pairs of word nodes; one direct method computes the distance by finding the minimum length of path that connects the two word nodes (Rada et al., 1989; Budanitsky and Hirst, 2006). Resnik similarity (Resnik, 1995) is based on information theory-based hierarchical taxonomy, and only considers the information content of the lowest common subsumer; i.e. the one that is closest to the two concepts compared. Lin similarity is based on Resnick similarity, but is calculated by normalizing to a common node, assuming their independence (Lin, 1998). Lesk similarity is computed to compare analogous words in their glosses (Banerjee and Pedersen, 2003). However, these methods require construction of a hierarchy of lexical databases; this process takes a long time, and can miss many words and most phrases.

Distributional similarity is another approach to quantifying the relatedness of keywords. Similarity is computed in a term-context matrix. The concept is based on the hypothesis that words are cognate if their contexts are cognate (Harris, 1985). A term-context matrix was prepared for each term and the corresponding contexts in documents of unstructured text, and the relatedness of words was perceived by the similarity of context vectors (Pantel et al., 2009). Document similarity used in many existing studies concentrated on a column vector to compare documents in term–document matrix, but word similarity focuses on a row vector to compare keywords in transformed term-context matrix (Fig. 1); the transformed matrix uses statistical measures such as pointwise mutual information (PMI), weighed PMI, and positive PMI.

With useful extraction tool of domain terminology, *Alchemy*™, TF-IDF evaluation method for selecting main keywords, and distributional similarity for identifying the relatedness of keywords, progressive keyword selection and identification of their relatedness will be considered next.

3. Proposed method

3.1. Overall research process

The overall process to monitor emerging technologies is as follows (Fig. 2). (1) Domain technology patents are collected from the USPTO database, and *Alchemy*™ is used to extract technical keyword candidates from their contents. (2) Extracted keywords are assessed using a TF-IDF function, and technical keywords are automatically identified. (3) Relatedness between keywords is quantified using the concept of distributional similarity. (4) Given that the process of identifying relatedness can attain semantic processing, a dissimilarity matrix is constructed by comparing patent documents. (5) Agglomerative clustering is used for keyword trend analysis; emerging technologies are discovered by analyzing the clustering results.

3.2. Definition of emerging technologies

Emerging technology is characterized by radical novelty, relatively fast growth, coherence, prominent impact, and uncertainty (Rotolo et al., 2015). A technology of radical novelty indicates "novelty (or newness)" (Small et al., 2014) or that a new idea is built on different basic principles than are existing methods (Arthur, 2007). A technology of relatively fast growth has "clockspeed nature" (Srinivasan, 2008) or is likely to produce an increasing number of number of papers (e.g. publications, patents) over time. A technology of coherence means "convergence of previously separated research streams" (Day and Schoemaker, 2000) or "has already moved beyond the purely conceptual stage" (Stahl, 2011). A technology of prominent impact "creates a new industry or transforms existing ones" (Day and Schoemaker, 2000) or "exerts much enhanced economic influence" (Porter et al., 2002). A technology of uncertainty involves ambiguity regarding potential applications of the technology (Stirling, 2007).

Emerging technology is likely to pass through three phases (e.g., pre-emergence, emergence, post-emergence) that are characterized by the above attributes (Rotolo et al., 2015) (Fig. 3). During the pre-emergence phase, a technology has high levels of radical novelty and uncertainty, but low levels of growth rate, coherence, and prominent impact. During the emergence phase, the technology may show moderate relatively fast growth, coherence, and prominent impact; as a consequence, during the post-emergence phase, the technology reaches low levels of radical novelty and uncertainty.

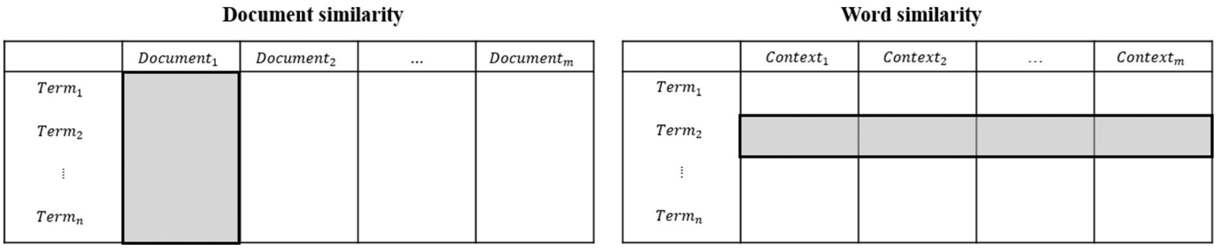
This study focuses on the emergence phase that has features of both adequate radical novelty and coherence. These emerging technologies have a high probability to become the mainstream of the market. Therefore, this paper identifies these emerging technologies by using clustering analysis to capture coherence, and keyword trend analysis to recognize radical novelty.

3.3. Collection of patents & extraction of keywords

Domain technology patents were collected from the USPTO database, taking patent dates into account. Then *Alchemy*™ was used to extract technical keyword candidates from various components including abstracts, summary of the invention, descriptions of the preferred embodiments, examples, and claims. *Alchemy*™ can extract noun phrases that have ambiguous of lexical boundaries, and is therefore useful in automatic extraction for domain terminologies.

Technical keyword candidates must then be refined to a general form, because experts or attorneys who write patents can use different terms to describe the same technology. Therefore, the WordNet lemmatizer that uses word affix information is used to process inflectional terms (e.g., to remove 's' from 'measurements' to convert it to

ARTICLE IN PRESS



	$Document_1$	$Document_2$	…	$Document_m$
$Term_1$	▨			
$Term_2$	▨			
⋮	▨			
$Term_n$	▨			

	$Context_1$	$Context_2$	…	$Context_m$
$Term_1$				
$Term_2$	▨	▨	▨	▨
⋮				
$Term_n$				

Fig. 1. Comparison of document similarity and word similarity.

the root form, 'measurement'). Phrase terms like 'special membrane', 'novel membrane', and 'inventive membrane' are processed by identifying the head noun. In the above cases, the head noun is 'membrane', so these phrases all refer to a similar concept. To standardize the terminology, chemical formulas are modified to full names: e.g., H_2O_2, NH_3, and H^+ are transformed to 'hydrogen peroxide', 'ammonia', and 'hydrogen ion', respectively.

3.4. Decision of technical keywords

TF and IDF have complementary advantages. TF can identify significant words, but has the disadvantage that it includes common words. IDF can comprehend domain-specific terms. Therefore, the combined TF-IDF method is an effective way to identify meaningful technical keywords that represent contents of technology described in a patent collection while avoiding common terms and including domain-specific terms. Python natural language toolkit package was used to apply the TF weighting method to determine the frequency of many variants of technical keyword candidates. TF is basically the number of times the term occurs in a patent collection, but this method gives a high value to common words, which may not be relevant to the technology. Therefore, the Augmented TF normalizes word weights by adding an arbitrary constant of 0.5 to terms that appear in the patent collection (Croft, 1983):

$$\text{Augmented TF} = 0.5 + \frac{0.5 \times tf_t}{max_{t(tf_t)}}, \tag{1}$$

where tf_t is the frequency at which term t occurs in a collection of patent documents.

In addition, logarithm TF is provided to reduce the effect of frequency by taking the logarithm of frequency (Robertson and Walker, 1994):

$$\text{Logarithm TF} = 1 + \log(tf_t); \tag{2}$$

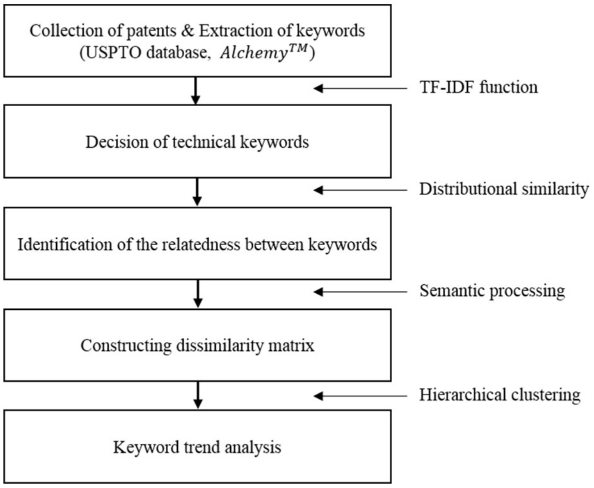

Collection of patents & Extraction of keywords (USPTO database, $Alchemy^{TM}$)
TF-IDF function
Decision of technical keywords
Distributional similarity
Identification of the relatedness between keywords
Semantic processing
Constructing dissimilarity matrix
Hierarchical clustering
Keyword trend analysis

Fig. 2. Overall framework.

this form has normally been used in information retrieval. Augmented TF restricts the TF values to a maximum value of 1.0, and therefore only settles the matter of normalizing the terms that have the highest frequencies. It is a useful way to evaluate keywords while considering long documents in which higher term frequencies have large values as others. Otherwise, Logarithm TF is normally used for normalization.

To make up for TF's limitations, IDF is calculated as (Sparck Jones, 1972)

$$\text{IDF} = \log\frac{N}{df_t}, \tag{3}$$

where N is the number documents in a collection and df_t is the number of documents that include a specified word.

Probabilistic IDF is obtained by first subtracting df_t from N in the denominator of (3), and setting a floor of 0 (Kolda, 1997):

$$\text{Probabilistic IDF} = \max\left\{0,\ log\frac{N-df_t}{df_t}\right\}; \tag{4}$$

This quantity assigns a low weight to terms that appear in more than half of the documents. Both IDF measures are useful in term evaluation, but Probabilistic IDF gives negative values to keywords that appear in more than half of the documents; for this reason probabilistic IDF is more effective than IDF to eliminate common words.

This paper suggests combinations of Logarithm TF and Probabilistic IDF (i.e., the product of a Logarithm TF and a Probabilistic IDF). The reason why select Logarithm TF is that higher term frequencies normally have no big difference to the others, and Probabilistic IDF effectively remove common words, so that technical keywords are determined by their rankings.

3.5. Identification of the relatedness between keywords

The concept of distributional similarity is useful because words are cognate if their contexts are cognate. Therefore, the concept is suitable to provide objective evidence to identify the relatedness between keywords in emerging fields that have no domain-terminology thesaurus. The relatedness between keywords is quantified by the concept of distributional similarity, and is represented as a technical keyword-context matrix. In the process (Fig. 4), if raw counts of technical keyword frequency are not processed, the results may be influenced by words that generally appear frequently, or by words that are frequent in the collection. For this reason, a raw count in its matrix is transformed to pointwise mutual information (PMI), which is a statistical measure of word association (Church and Hanks, 1989):

$$\text{PMI}(word_i, context_j) = log_2\frac{P(word_i, context_j)}{P(word_i) \times P(context_j)}. \tag{5}$$

where $P(word_i)$ is calculated by dividing the total of $word_i$ count in the contexts into the total count in the matrix, $P(context_j)$ is calculated by dividing the total of $context_j$ count in the words into the total count in

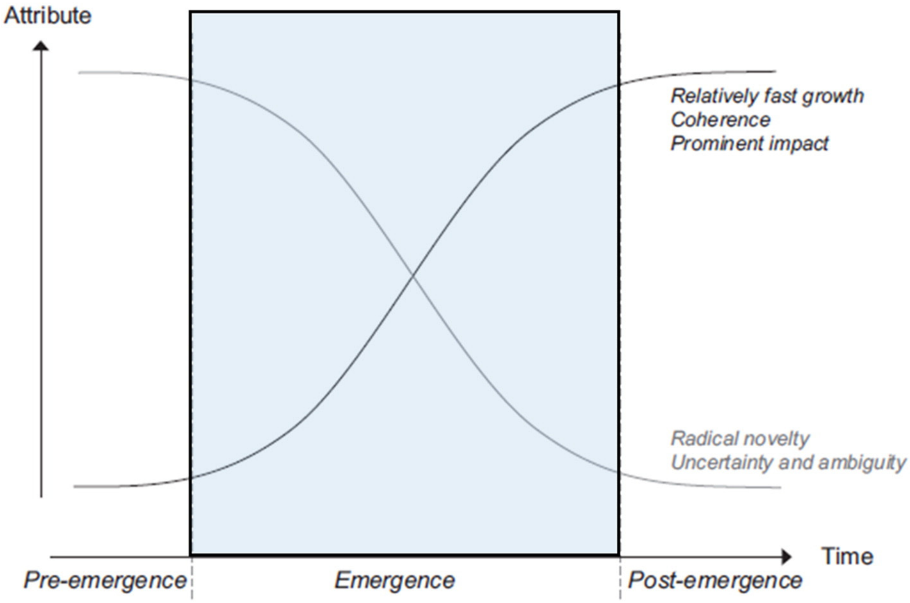


Fig. 3. Pre-emergence, emergence, and post-emergence characterized by attributes (Rotolo et al., 2015).

the matrix, and $P(word_i, context_j)$ is calculated by dividing the count in $word_i$ and $context_j$ into the total count in the matrix.

PMI can also apply various smoothing techniques, such as 'weighting PMI', which assigns weights to words that do not occur, and 'positive PMI' (PPMI), which replaces all negative PMI with '0' (Niwa and Nitta, 1994; Turney and Pantel, 2010). Consequently, our method transforms raw counts to PMI, weighting PMI, or PPMI, and calculates the probability of each word and context. The similarity of technical keywords is quantified using cosine similarity, Jaccard similarity, Dice similarity, Jensen-Shannon similarity, or some other method (Lee, 1999). As a result, with context vectors in the preceding a technical keyword-context matrix, a pair of keywords is compared by executing the python *scikit learn* package. The relatedness of two keywords that has high similarity score is assigned by domain experts by reference to WordNet Noun relations (Fig. 5) (Miller, 1995) and is stored for sematic processing.

3.6. Constructing dissimilarity matrix

A dissimilarity matrix is constructed by semantic processing of the similarity between keyword vectors that represent patents, while considering the size and characteristics of the data (Moehrle, 2010) (Fig. 6). To measure the similarity between two patents in a collection, the first step is to determine whether they share keywords. Technical keywords that have the same word surface form in both patents are judged to be identical. They also can be recognized as the same word if their relatedness databases are identical. Then the *dis*similarity of the pair of patents is calculated as

$$\text{Dissimilarity}(X, Y) = 1 - \text{Similarity}(X, Y). \tag{6}$$

The dissimilarities among all pairs of N patents are assembled in an N x N symmetric matrix.

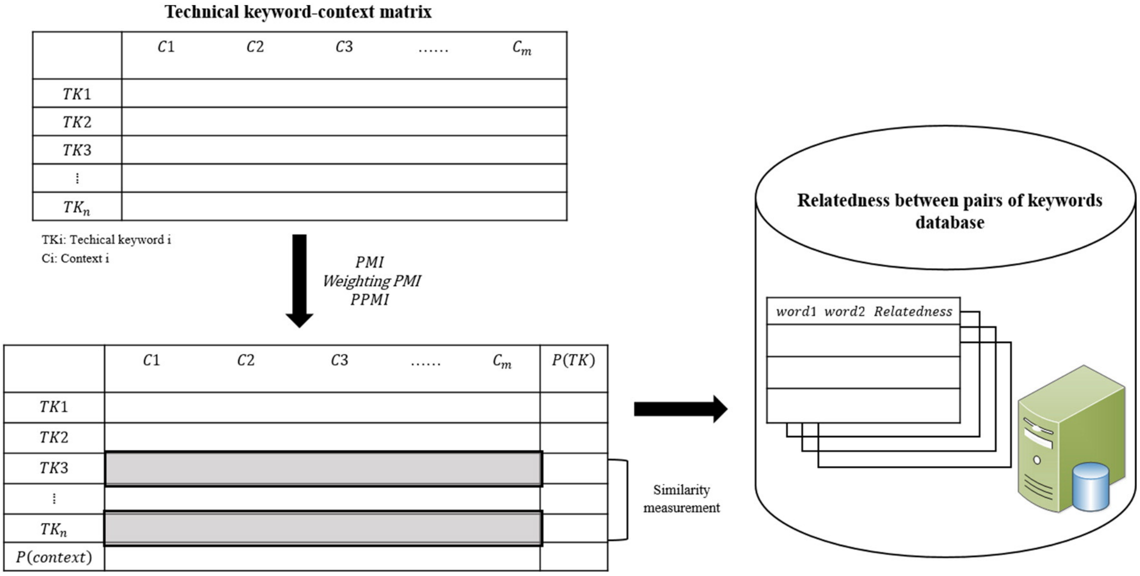


Fig. 4. Process for relatedness between pairs of keywords database.

Please cite this article as: Joung, J., Kim, K., Monitoring emerging technologies for technology planning using technical keyword based analysis from patent data, Technol. Forecast. Soc. Change (2016), http://dx.doi.org/10.1016/j.techfore.2016.08.020

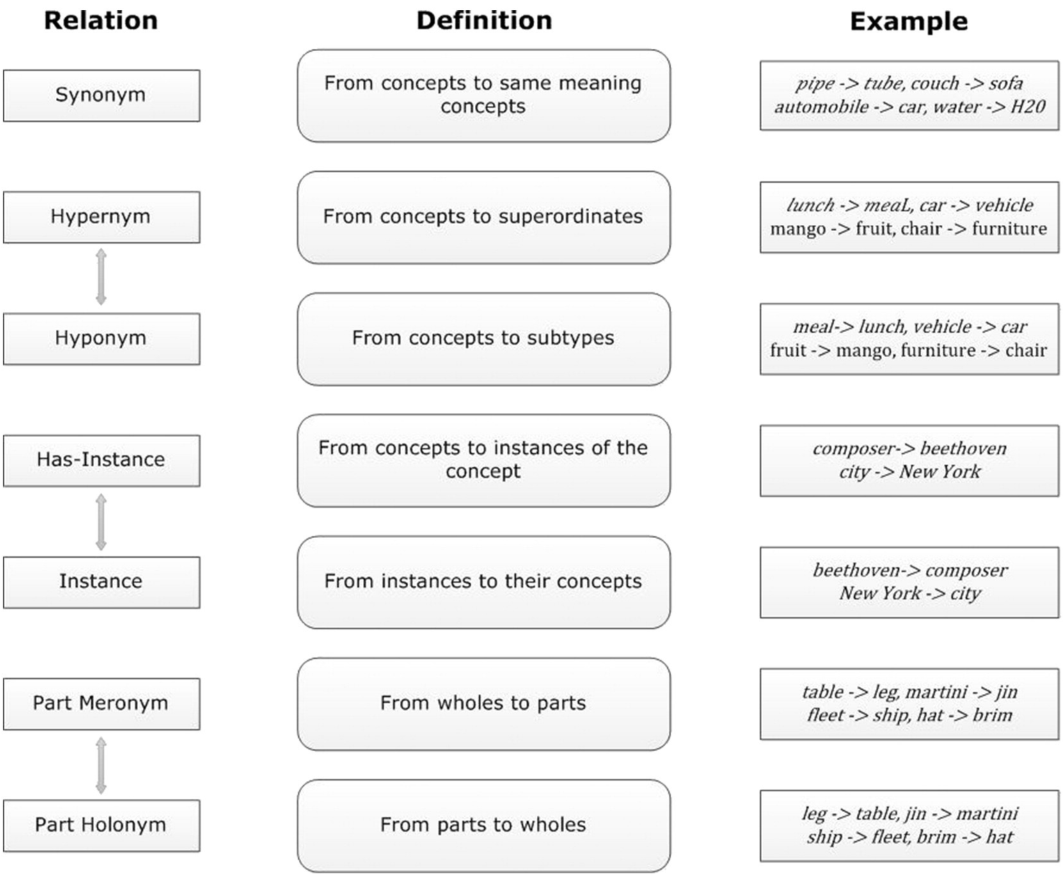

Relation
Definition
Example
Synonym
From concepts to same meaning concepts
pipe -> tube, couch -> sofa
automobile -> car, water -> H20
Hypernym
From concepts to superordinates
lunch -> meaL, car -> vehicle
mango -> fruit, chair -> furniture
Hyponym
From concepts to subtypes
meal-> lunch, vehicle -> car
fruit -> mango, furniture -> chair
Has-Instance
From concepts to instances of the concept
composer-> beethoven
city -> New York
Instance
From instances to their concepts
beethoven-> composer
New York -> city
Part Meronym
From wholes to parts
table -> leg, martini -> jin
fleet -> ship, hat -> brim
Part Holonym
From parts to wholes
leg -> table, jin -> martini
ship -> fleet, brim -> hat

Fig. 5. Example of Noun relations in WordNet (Miller, 1995).

3.7. Keyword trend analysis

The strategy of clustering patent documents is a useful way to identify the main stream of emerging technologies, and keyword trend analysis can capture radical novelty by analyzing keywords that symbolize a cluster. Therefore this strategy can identify technology during the emergence phase that has the characteristics of coherence and radical novelty.

To fulfill this, hierarchical agglomerative clustering is performed based on dissimilarities between pairs of patents. Numerous linkage methods such as average linkage, centroid linkage, complete linkage, median linkage, single linkage, and ward linkage can be used for this process. All methods can be applied in hierarchical agglomerative clustering, but clustering results should conform to the goodness of fit that represents the validity of the model in clustering results or provides

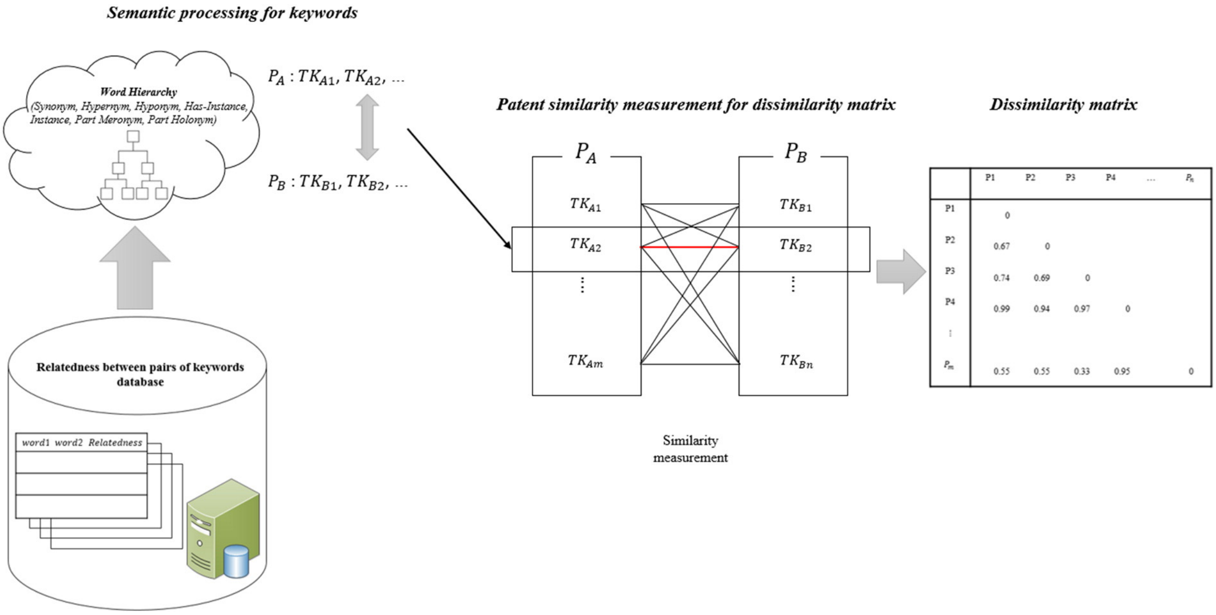

Semantic processing for keywords
Word Hierarchy
(Synonym, Hypernym, Hyponym, Has-Instance, Instance, Part Meronym, Part Holonym)
$P_A : TK_{A1}, TK_{A2}, \ldots$
$P_B : TK_{B1}, TK_{B2}, \ldots$
Relatedness between pairs of keywords database
word1 word2 Relatedness
Patent similarity measurement for dissimilarity matrix
P_A
P_B
TK_{A1}
TK_{A2}
TK_{Am}
TK_{B1}
TK_{B2}
TK_{Bn}
Similarity measurement
Dissimilarity matrix
P1 P2 P3 P4 … P_n
P1 0
P2 0.67 0
P3 0.74 0.69 0
P4 0.99 0.94 0.97 0
⋮
P_m 0.55 0.55 0.33 0.95 0

Fig. 6. Procedure for dissimilarity matrix using semantic processing.

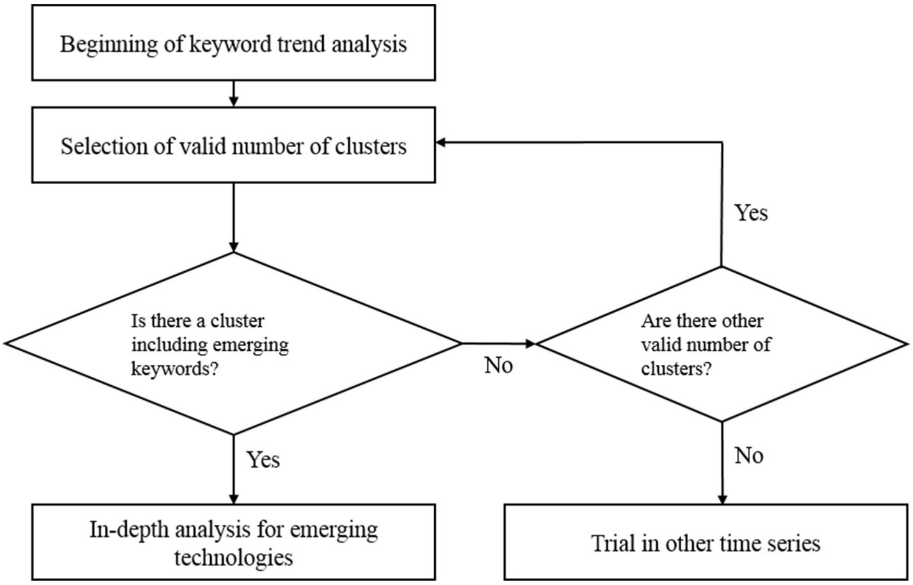

Beginning of keyword trend analysis
Selection of valid number of clusters
Yes
Is there a cluster including emerging keywords?
No
Are there other valid number of clusters?
Yes
No
In-depth analysis for emerging technologies
Trial in other time series

Fig. 7. Flow chart of keyword trend analysis.

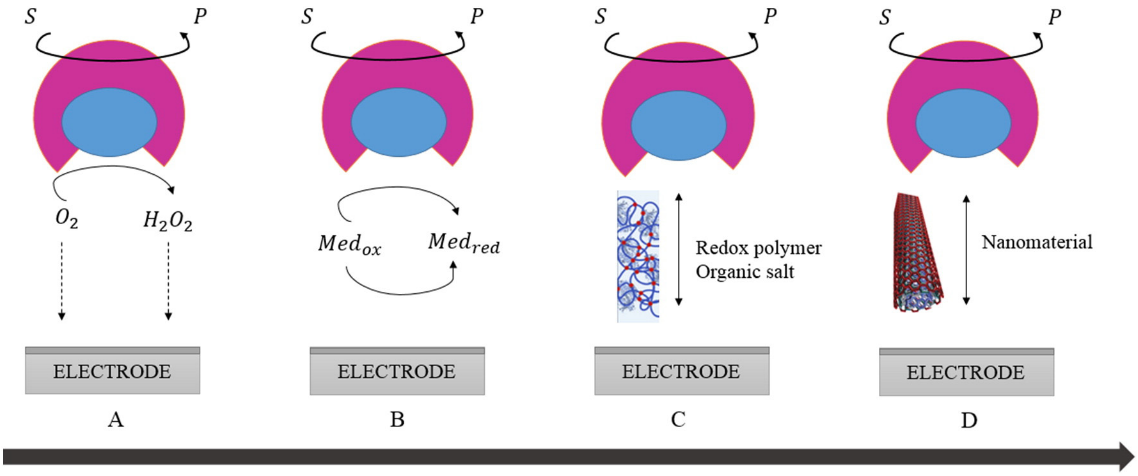


Fig. 8. Evolution of electron transfer mechanisms in glucose biosensor.

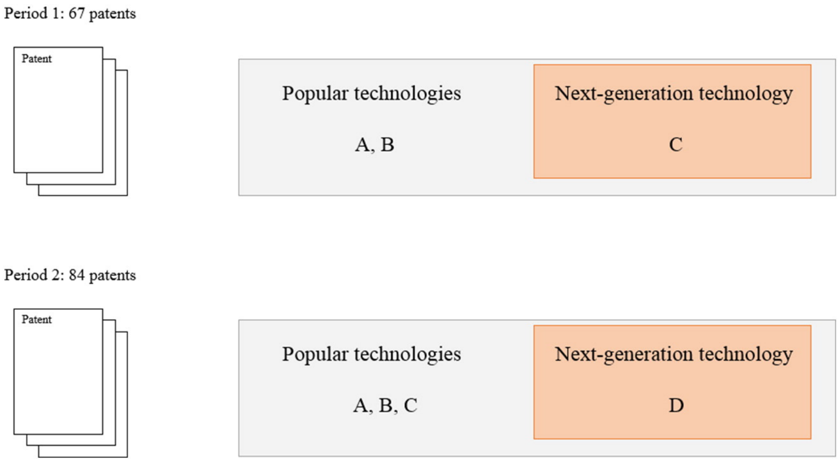


Fig. 9. Trend of electrochemical glucose biosensors based on mechanisms of electron transfer.

Please cite this article as: Joung, J., Kim, K., Monitoring emerging technologies for technology planning using technical keyword based analysis from patent data, Technol. Forecast. Soc. Change (2016), http://dx.doi.org/10.1016/j.techfore.2016.08.020

Table 1
Technical keywords in period 1.

Keyword	Logarithmic TF	Prob IDF	Rank (TF × IDF)
Direct electron transfer	2.26	1.51	1 (3.41)
TCNQ	2.62	1.20	2 (3.14)
PQQGDH	2.61	1.20	3 (3.13)
HRP	2.81	1.09	4 (3.06)
⋮	⋮	⋮	⋮
Analyte	3.20	0.01	321 (0.03)

analysts with a reasonable understanding of them. The clustering process proceeds as follows. First, one of many linkage methods is chosen to assess the distance between clusters and one patent is assigned to one cluster. Second, after the distances between all pairs of clusters are measured using a linkage method based on the dissimilarity matrix, the pair of clusters that is separated by the shortest distance is agglomerated. The process continues until all clusters have been combined.

Next, a keyword trend analysis using a flow chart (Fig. 7) is conducted to identify clustering results. First, clusters for which the best-cut value >1.25 are identified; this value indicates that a clustering result is statistically significant (Everitt et al., 2001). Normally, several clusters meet this criterion; each cluster is characterized by technical keywords that occur in it but not in other groups. The cluster with the highest best-cut value is examined first. If it contains emerging keywords, they can be used to identify emerging technologies. Emerging keywords are recognized as terms that have appeared in academic papers during the pre-emergence phase, and that are identified by domain researchers; these terms are determined in terms of radical novelty (Section 3.2). If it contains no emerging keywords, the conclusion is that its keyword set cannot identify emerging technologies. Therefore, the cluster with the next-highest best-cut value is evaluated in the same way. This process of sequential examination and (if necessary) rejection is continued until a cluster with emerging keywords is identified. If no cluster contains emerging keywords, a technical keyword-based patent analysis should be performed in other time series. By this process, patents that include emerging technical keywords are identified, and emerging technologies can be found by in-depth analysis of the patents in the cluster.

4. Case study: mechanisms of electron transfer in electrochemical glucose biosensors

A case study of ‘mechanisms of electron transfer in electrochemical glucose biosensors’ was conducted to demonstrate how the proposed method works.

4.1. Data

The proposed method was applied to patents related to mechanisms of electron transfer in glucose biosensors, and that were collected from the USPTO database by searching the terms ‘glucose biosensors’ and ‘electron transfer’ in abstracts. After removing irrelevant patents, 151 patents were used to identify emerging technologies. All parts of the patent (title, abstract, claims, summary of the invention, and examples) were used for analysis except the background, which was not considered because keywords in it can misrepresent the content of the patent. Electrochemical glucose biosensors are used to monitor blood sugar levels in diabetes patients, and occupy a large portion of the biosensor market. The devices must measure blood glucose level quickly and exactly. In these circumstance, a new technology that recognizes blood sugar in a better way than do existing devices can provide a company's developers with innovative detection strategies.

Table 2
Technical keywords in period 2.

Keyword	Logarithmic TF	Prob IDF	Rank (TF×IDF)
SWNT	2.34	1.61	1 (3.77)
GDH	2.15	1.61	2 (3.46)
metal ion	2.04	1.61	3 (3.28)
nanorod	2.04	1.61	3 (3.28)
⋮	⋮	⋮	⋮
counter electrode	3.12	0.02	496 (0.06)

Table 3
Relatedness of pairs of keywords in period 1.

Word1	Word2	Cosine similarity	Relatedness
Direct electron transfer	HRP	0.78	Hyponym
TCNQ	Redox electrode	0.82	Part holonym
Redox membrane	Polymer matrix	0.91	Part meronym
Subcutaneous tissue	Subcutaneous glucose sensor	0.76	Hypernym
Cathode	Cathode wire	1	Part meronym
Drug	Insulin	0.79	Has-instance
Epidermis	Dermis	0.99	hypernym
Polymer matrix	Tetracyanoquinodimethane	0.76	Part meronym
⋮	⋮	⋮	⋮

Four generations of electrochemical glucose biosensors were identified according to their mechanisms of electron transfer (Fig. 8). (A) First-generation glucose biosensors exploited the chemical reaction between natural oxygen cosubstrate and hydrogen peroxide. (B) Second-generation glucose biosensors replaced oxygen with a mediator that exploits a redox process to shuttles electrons to the electrode. (C) Third-generation sensors were reagentless: the mediator was removed, and electrons were transferred directly to the electrode through a redox polymer or organic salt. (D) Fourth-generation glucose biosensors used various nanomaterials to achieve this direct electrical connection.

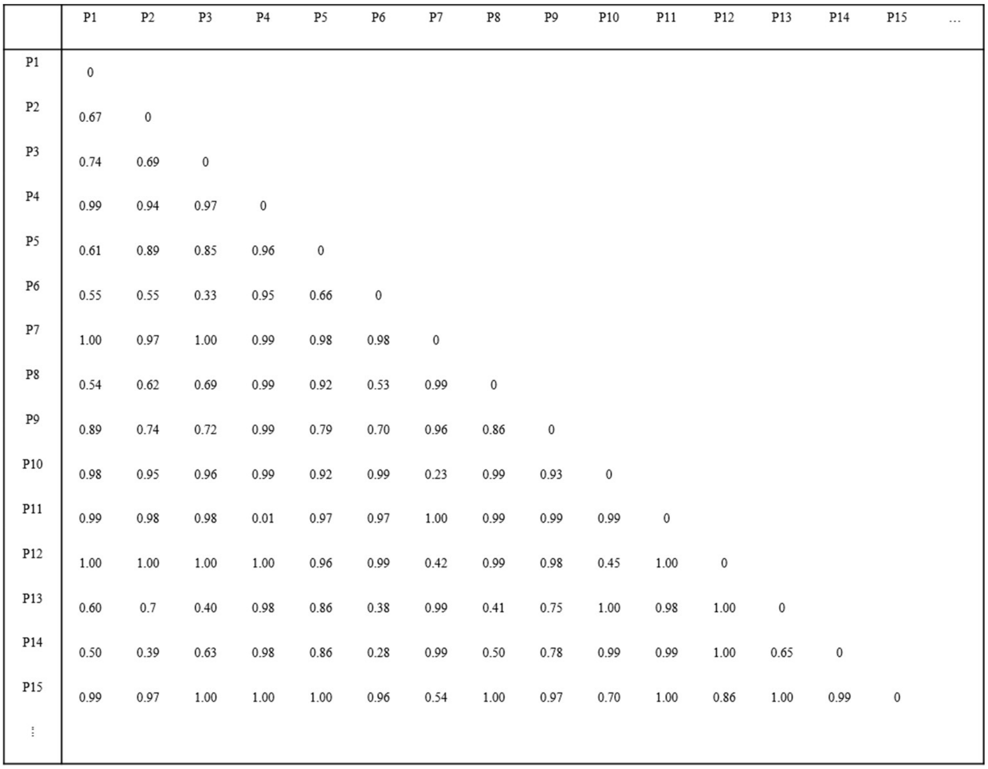
The dataset was divided into two periods to demonstrate that the proposed method can monitor emerging technologies, in this case the next-generation technology (Fig. 9). The time series data ran from 1991 to 2010; 151 patents were used. The first period contained 67 patents that identify ‘C technology’ as an emerging technology in the period 1991–2000, assuming that next-generation technology was not known during that period. In the same manner, the second period included 84 patents that detect ‘D’ as an emerging technology in 2001–2010.

4.2. Identifying technical keywords

After technical keyword candidates extracted using *Alchemy™* were refined, they were identified using a TF-IDF function during periods 1 and 2. To weight the keywords, the logarithm TF method was used, and combined with probability IDF because it eliminates common

Table 4
Relatedness of pairs of keywords in period 2.

Word1	Word2	Cosine similarity	Relatedness
HEMT	High electron mobility transistor	1	Synonym
Suitable non-releasable	Redox species	0.84	Hyponym
Saliva	Specific analyte	1	Instance
Electrical device	Component	0.77	Part meronym
SWCNT	Single-walled carbon nanotube	0.84	Synonym
Carbon ink	Ink	0.71	Instance
Analyte-specific signal	Measurement signal	0.74	Hypernym
Interferant eliminating layer	Subcutaneous glucose sensor	0.99	Part holonym
⋮	⋮	⋮	⋮



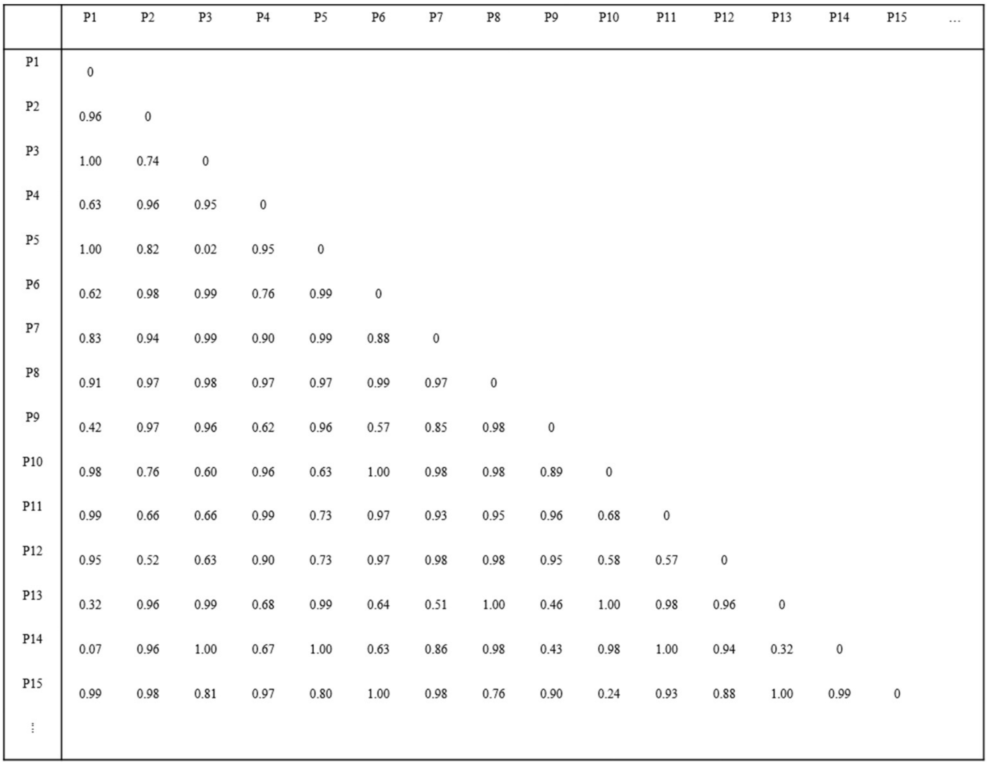
	P1	P2	P3	P4	P5	P6	P7	P8	P9	P10	P11	P12	P13	P14	P15	…
P1	0															
P2	0.67	0														
P3	0.74	0.69	0													
P4	0.99	0.94	0.97	0												
P5	0.61	0.89	0.85	0.96	0											
P6	0.55	0.55	0.33	0.95	0.66	0										
P7	1.00	0.97	1.00	0.99	0.98	0.98	0									
P8	0.54	0.62	0.69	0.99	0.92	0.53	0.99	0								
P9	0.89	0.74	0.72	0.99	0.79	0.70	0.96	0.86	0							
P10	0.98	0.95	0.96	0.99	0.92	0.99	0.23	0.99	0.93	0						
P11	0.99	0.98	0.98	0.01	0.97	0.97	1.00	0.99	0.99	0.99	0					
P12	1.00	1.00	1.00	1.00	0.96	0.99	0.42	0.99	0.98	0.45	1.00	0				
P13	0.60	0.7	0.40	0.98	0.86	0.38	0.99	0.41	0.75	1.00	0.98	1.00	0			
P14	0.50	0.39	0.63	0.98	0.86	0.28	0.99	0.50	0.78	0.99	0.99	1.00	0.65	0		
P15	0.99	0.97	1.00	1.00	1.00	0.96	0.54	1.00	0.97	0.70	1.00	0.86	1.00	0.99	0	
⋮																

Fig. 10. Part of the dissimilarity matrix in period 1.

terms by assigning a negative value to the most-common term in the data. Finally, the crucial technical term that had the top TF-IDF score among the keywords was identified. The numbers of technical keywords identified were 321 in period 1 (Table 1) and 496 in period 2 (Table 2).

4.3. Identifying the relatedness between keywords

Words that shared intimate relationships (e.g., synonym, hypernym, hyponym) were identified in the data. After the technical keyword-context matrix was altered to a PPMI matrix, the relatedness



	P1	P2	P3	P4	P5	P6	P7	P8	P9	P10	P11	P12	P13	P14	P15	…
P1	0															
P2	0.96	0														
P3	1.00	0.74	0													
P4	0.63	0.96	0.95	0												
P5	1.00	0.82	0.02	0.95	0											
P6	0.62	0.98	0.99	0.76	0.99	0										
P7	0.83	0.94	0.99	0.90	0.99	0.88	0									
P8	0.91	0.97	0.98	0.97	0.97	0.99	0.97	0								
P9	0.42	0.97	0.96	0.62	0.96	0.57	0.85	0.98	0							
P10	0.98	0.76	0.60	0.96	0.63	1.00	0.98	0.98	0.89	0						
P11	0.99	0.66	0.66	0.99	0.73	0.97	0.93	0.95	0.96	0.68	0					
P12	0.95	0.52	0.63	0.90	0.73	0.97	0.98	0.98	0.95	0.58	0.57	0				
P13	0.32	0.96	0.99	0.68	0.99	0.64	0.51	1.00	0.46	1.00	0.98	0.96	0			
P14	0.07	0.96	1.00	0.67	1.00	0.63	0.86	0.98	0.43	0.98	1.00	0.94	0.32	0		
P15	0.99	0.98	0.81	0.97	0.80	1.00	0.98	0.76	0.90	0.24	0.93	0.88	1.00	0.99	0	
⋮																

Fig. 11. Part of the dissimilarity matrix in period 2.

between pairs of keywords was quantified using their cosine similarity expressed on context vectors; to reduce the dataset, domain experts in period 1 and period 2 investigated all pairs with cosine similarity >0.7; i.e., which were closely related. A total of 1258 word relations were quantified in period 1 (Table 3) and 1906 word relations were quantified in period 2 (Table 4).

4.4. Constructing dissimilarity matrix by semantic processing

In this step, dissimilarity of patents was semantically measured using the relations database. Dissimilarities among patents were calculated on a technical keyword-patent matrix for every pair of patents; the result was a 67 × 67 dissimilarity matrix with a zero diagonal in period 1 (Fig. 10) and an 84 × 84 dissimilarity matrix with a zero diagonal in period 2 (Fig. 11).

4.5. Results of keyword trend analysis

The final step analyzed uses hierarchical clustering based on the dissimilarity matrix to monitor emerging technologies. Hierarchical agglomerative clustering was conducted by using the Ward linkage method (Ward, 1963); the numbers of valid clusters according to best-cut value were 2 to 6 in period 1 and 2 to 7 in period 2.

4.5.1. Period 1

Keyword trend analysis identified three clusters based on valid best-cuts, and technical keywords to present characteristics of cluster were extracted (Table 5). Cluster 3 contained several keywords that were closely linked to keywords in technology C; e.g., redox polymer, organic salts, wired redox enzyme. Patents that correspond to cluster 3 were checked, and they indicated the next-generation technology 'C' (Table 6). In-depth analysis of these patents identified emerging technologies that were promising future technology. Moreover, examination revealed that clusters 1 and 2 included technical keywords in popular technologies A and B. As a result, the proposed method identified technology C, which was the emerging technology at that time.

4.5.2. Period 2

Four clusters were chosen for analysis. They included numerous technical keywords (Table 7). Cluster 4 included keywords connecting with emerging technology D; e.g., single-walled carbon nanotube, carbon nanotubes, nanomaterials. Therefore, this cluster was examined to determine whether it actually included the next-generation technology D. Some patents did include this technology (Table 8); therefore the proposed method identified technology D that was an emerging technology at in period 2.

Table 5
Technical keywords of clusters in period 1.

Cluster 1	Cluster 2	Cluster 3
Electrolyte	Non-flowing manner	SINGLE calibration sample
Reservoir	Distal end	Convenient rate
Porosity	Sorbent material	Genetically engineered glucose
Microprocessor	Potentiometric techniques	Electroreactive species
External receiver	Measurement period	Reagent
Data	Site	TCNQ
Microneedle	Film	Wide variety
LIQUID sample	Mediator effective diffusion	Graphite powder
Delivery	Counter	Redox enzyme
Catheter	Region	Sample solution
⋮	⋮	⋮

Table 6
Summary of patents to detect technology C, emerging technologies.

Patent no.	Title	Contents of "technology C"
5378332	Amperometric flow injection analysis biosensor for glucose based on graphite paste modified with tetracyanoquinodimethane	Use of organic metals containing salts of tetracyanoquinodimethane (TCNQ) for the direct electron exchange
5779867	Dry chemistry glucose sensor	Use of redox salts about TCNQ-TTF complex characterized by a burgundy red coloration and having an ultraviolet absorption spectrum with broad absorption from about 340 nm to about 550 nm
6284478	Subcutaneous glucose electrode	Use of redox polymer comprising a redox enzyme and a redox compound which are non-leachable by fluids in the body at a pH of between about 6.5 and about 7.8
6103509	Modified glucose dehydrogenase	Use of modified pyrrolo-quinolone quinone glucose dehydrogenase (PQQGDH) in determining glucose concentration for electron transfer
6134461	Electrochemical analyte	Use of quaternized osmium-containing redox polymer or 2-aminoethylferrocene
5225064	Peroxidase colloidal gold oxidase biosensors for mediatorless glucose determination	Use of horseradish peroxidase(HRP) without an electron transfer mediator

5. Conclusion

This paper presents an approach to identify emerging technologies, and applied it to a case of mechanisms of electron transfer in electrochemical glucose biosensors. The proposed method uses advanced algorithms to automatically select keywords by applying a TF-IDF function, and to quantify relatedness between pairs of keywords, rather relying on domain experts as did previous keyword-based patent analysis. The algorithm to automatically select keywords replaces previous subjective judgments of domain experts with decision of high rank based on TF-IDF value, and contributes to reduction of domain experts' work by just choosing a cutoff of TF-IDF values. The algorithm to quantify relatedness provides objective evidence for keyword relatedness by identifying pairs of keywords that have high similarity values, and improves semantic processing to compare patent documents. Due to the complexity and fusion of emerging technology, two or three principal keywords are not sufficient to characterize this technology. For example,

Table 7
Technical keywords of clusters in period 2.

Cluster 1	Cluster 2	Cluster 3	Cluster 4
Stability	NAD	Sensor reader	Electrically insulating layer
Sensing step	Reagent	Analyte responsive enzyme	HEMT
Operative contact	Insulating substrate	Electrolytic contact	Thioglycolic acid
Signal component	Subcutaneous measurement	Layered structure	Nanowire
Microprocessor	Polyamide insulated gold	Electroreduction	Electric field
Measurement cycle	GDH activity	Ruthenium dioxide	Sensor array
Collection reservoir	Enzymatic activity	Boundary area	SWCNT
Phosphate	Amino acid	Fingertip	Nanorod
Tissue	Test strip	Micrometer	Immobilized enzyme
Detectable signal	Flavin adenine dinucleotide	Nerve end density	CNT
⋮	⋮	⋮	⋮

Please cite this article as: Joung, J., Kim, K., Monitoring emerging technologies for technology planning using technical keyword based analysis from patent data, Technol. Forecast. Soc. Change (2016), http://dx.doi.org/10.1016/j.techfore.2016.08.020

Table 8
Summary of patents to detect technology D, emerging technologies.

Patent No.	Title	Contents of “technology D”
7118881	Micro/nano-fabricated glucose sensors using single-walled carbon nanotubes	Use of single walled carbon nanotubes (SWCNTs) assembled on microelectrodes for the hydrogen-specific gas sensing
7452452	Carbon nanotube nanoelectrode arrays	Use of carbon nanotube (CNT) materials comprising arrays of linear carbon nanotubes with controllable site densities
7955483	Carbon nanotube-based glucose sensor	Use of peptide nanostructures composed of self-assembled peptides for electrode coating
8039909	Semiconductor nanowires charge sensor	Use of semiconductor nanowire bonding with the functional material in the chemical coating layer
8232584	Nanoscale sensors	Use of n-doped semiconductor nanoscale wire and at least one p-doped semiconductor nanoscale wire, each having a reaction entity immobilized thereon for electrical sensor array device
8835984	Sensors using high electron mobility transistors	Use of nanorods for high electron mobility transistors(HEMTs)

glucose biosensors entail biotechnology, electronics, and nanotechnology, so emerging technologies could not be expressed using two or three main keywords. For this reason, automatic technical keyword selection that conveys contents of patent is useful, and the relatedness between keywords database allows semantic processing of the similarity of patent documents. Consequently, our algorithms are expected to enhance reliable and objective keyword-based patent analysis to identify emerging technologies.

Monitoring of emerging technologies by analyzing patent information provides answers to questions related to technology planning, such as: ‘Which contents are described as emerging technologies?’ and ‘Who is active in developing those technologies?’ In practice, researchers and developers can identify the contents of emerging technologies in detail, and in this way may give information about core technologies that may develop innovative products in the future. Strategically, the information can help entrepreneurs and decision/policy makers to prioritize project support for emerging technologies. Technology managers and strategic planners can obtain information about main agents that have patent rights to emerging technologies, and may therefore be able to draw up plans for cross-licensing, patent purchase, or M&A in advance.

However, the proposed method should be improved in future research. First, constructing patent analysis systems to identify emerging technologies may be difficult to automate completely, but the suggested algorithms decreased dependence on domain experts by offering automatic keyword selection and identification of their relatedness. Second, the relatedness of keywords (e.g., hypernym, hyponym, synonym) is sometimes ambiguous; even experts can disagree about how words are related. Finally, this method is works at the keyword level; a more-intelligent method would use information that is more meaningful than keywords.

Acknowledgments

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Ministry of Science, ICT & Future Planning (MSIP) (No. 2013R1A2A2A03068496).

References

Altuntas, S., Dereli, T., Kusiak, A., 2015a. Analysis of patent documents with weighted association rules. Technol. Forecast. Soc. Chang. 92 (3), 249–262.

Altuntas, S., Dereli, T., Kusiak, A., 2015b. Forecasting technology success based on patent data. Technol. Forecast. Soc. Chang. 96, 202–214.

Arthur, W.B., 2007. The structure of invention. Res. Policy 36 (2), 274–287.

Ashton, W.B., Kinzey, B.R., Gunn Jr., M.E., 1991. A structured approach for monitoring science and technology developments. Int. J. Technol. Manag. 6, 91–111.

Ashton, W.B., Johnson, A.G., Stacey, G.S., 1994. Monitoring science and technology for competitive advantage. Compet. Intell. Rev. 5 (1), 5–16.

Banerjee, S., Pedersen, T., 2003. Extended Gloss Overlap as a Measure of Semantic Relatedness. Proceedings of the 18th International Joint Conference on Artificial Intelligence. Acapulco, Mexico, pp. 805–810.

Budanitsky, A., Hirst, G., 2006. Evaluating WordNet-based measures of lexical semantic relatedness. J. Comput. Linguis. 32 (1), 13–47.

Chang, P.L., Wu, C.C., Leu, H.J., 2010. Using patent analyses to monitor the technological trends in an emerging field of technology: a case of carbon nanotube field emission display. Scientometrics 82, 5–19.

Chen, R., Liang, J., Pan, R., 2008. Using recursive ART network to construction domain ontology based on term frequency and inverse document frequency. Expert Syst. Appl. 34 (1), 488–501.

Choi, S., Yoon, J., Kim, K., Lee, J.Y., Kim, C.-H., 2011. SAO network analysis of patents for technology trends identification: a case study of polymer electrolyte membrane technology in proton exchange membrane fuel cells. Scientometrics 88, 863–883.

Church, K.W., Hanks, P., 1989. Word Association Norms, Mutual Information and Lexicography. Proceedings of the 27th Annual Conference of the Association of Computational Linguistics, pp. 76–83.

Courtial, J.P., Callon, M., Sigogneau, A., 1993. The use of patent titles for identifying the topics of invention and forecasting trends. Scientometrics 26 (2), 231–242.

Croft, W.B., 1983. Experiments with representation in a document retrieval system. Inf. Technol. Res. Dev. 2, 1–21.

Day, G.S., Schoemaker, P.J.H., 2000. Avoiding the pitfalls of emerging technologies. Calif. Manag. Rev. 42 (2), 8–33.

Engelsman, E.C., van Raan, A.F.J., 1992. A patent-based cartography of technology. Res. Policy 23 (1), 1–26.

Ernst, H., 2001. Patent applications and subsequent changes of performance: evidence from time-series cross-section analyses on the firm level. Res. Policy 30, 143–157.

Ernst, H., 2003. Patent information for strategic technology management. World Patent Inf. 25, 233–242.

Everitt, B.S., Landau, S., Leese, M., 2001. Clustering Analysis. Arnold, London.

Geum, Y., Jeon, J., Seol, H., 2013. Identifying technological opportunities using the novelty detection technique: a case of laser technology in semiconductor manufacturing. Tech. Anal. Strat. Manag. 25 (1), 1–22.

Hamilton, W., 1985. Corporate strategies for managing emerging technologies. Technol. Soc. 7, 197–212.

Harris, Z., 1985. Distributional structure. Philos. Linguist. 24–47.

Joung, J., Kim, K., 2015. Detecting technological opportunities using technical keywords from patents: a case of electrochemical glucose biosensor. Proceedings of 12th ICMIT International Conference, Singapore.

Kim, Y.G., Suh, J.H., Park, S.C., 2008. Visualization of patent analysis for emerging technology. Expert Syst. Appl. 34, 1804–1812.

Kolda, T.G., 1997. Limited-Memory Matrix Methods with Applications. Applied Mathematics Program. University of Maryland at College Park, pp. 59–68.

Lee, L., 1999. Measures of Distributional Similarity. Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics (ACL), College Park, MD, USA, pp. 25–32.

Lee, S., Yoon, B., Park, Y., 2009. An approach to discovering new technology opportunities: keyword-based patent map approach. Technovation 29, 481–497.

Lee, C., Jeon, J., Park, Y., 2011. Monitoring trends of technological changes based on the dynamic patent lattice: a modified formal concept analysis approach. Technol. Forecast. Soc. Chang. 78 (4), 690–702.

Li, Y.-R., Wang, L.-H., Hong, C.-F., 2009. Extracting the significant-rare keywords for patent analysis. Expert Syst. Appl. 36 (3), 5200–5204.

Lin, D., 1998. An Information-Theoretic Definition of Similarity. Proc. Int'l Conf. Machine Learning.

Miller, G.A., 1995. WordNet: a lexical database for English. Commun. ACM 38 (11), 39–41.

Missikoff, M., Smith, F., Taglino, F., 2015. Ontology building and maintenance in collaborative virtual environments. Concurr. Comput. Pract. Exp. 27 (11), 2796–2817.

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

Niwa, Y., Nitta, Y., 1994. Co-occurrence Vectors from Corpora Vs. Distance Vectors from Dictionaries. Proceedings of COLING, pp. 304–309.

Niwa, Y., Sakurai, H., 1999. Document retrieval-assisting method and system for the same and document retrieval service using the same with document frequency and term frequency, USPTO patent.

Noh, H., Jo, Y., Lee, S., 2015. Keyword selection and processing strategy for applying text mining to patent analysis. Expert Syst. Appl. 42, 4348–4360.

Pantel, P., Crestan, E., Borkovsky, A., Popescu, A.M., Vyas, V., 2009. Web-Scale Distributional Similarity and Entity Set Expansion. Proceeding of the 2009 Conference on Empirical Methods in Natural Language Processing Vol. 2, pp. 938–947.

Porter, A.L., Roessner, J.D., Jin, X.-Y., Newman, N.C., 2002. Measuring national emerging technology capabilities. Sci. Public Policy 29 (3), 189–200.

Rada, R., Mili, H., Bichnell, E., Blettner, M., 1989. Development and application of a metric on semantic nets. IEEE Trans. Syst. Man Cybern. 9 (1), 17–30.

Ramirez, R., Iversen, J., Ouimet, J., Kobti, Z., 2010. A Survey of Text Extraction Tools for Intelligent Healthcare Decision Support Systems. Advances in Intelligent Decision Technologies. SIST Vol. 4, pp. 393–402.

Resnik, P., 1995. Using Information Content to Evaluate Semantic Similarity in a Taxonomy. Proc. 14th Int'l Joint Conf. Artificial Intelligence.

Please cite this article as: Joung, J., Kim, K., Monitoring emerging technologies for technology planning using technical keyword based analysis from patent data, Technol. Forecast. Soc. Change (2016), http://dx.doi.org/10.1016/j.techfore.2016.08.020

Robertson, S., 2004. Understanding inverse document frequency: on theoretical arguments for IDF. J. Doc. 60 (5), 503–520.
Robertson, S.E., Walker, S., 1994. Some Simple Effective Approximations to the 2 Poisson Model for Probabilistic Weighted Retrieval. Proceedings of the Seventeenth Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 232–241.
Rotolo, D., Hicks, D., Martin, B., 2015. What is an emerging technology? Res. Policy 44 (10), 1827–1843.
Small, H., Boyack, K.W., Klavans, R., 2014. Identifying emerging topics in science and technology. Res. Policy 48 (8), 1450–1467.
Sparck Jones, K., 1972. A statistical interpretation of term specificity and its application in retrieval. J. Doc. 28 (1), 11–21.
Srinivasan, R., 2008. Sources, characteristics and effects of emerging technologies: research opportunities in innovation. Ind. Mark. Manag. 37 (6), 633–640.
Stahl, B.C., 2011. What Does the Future Hold? A Critical View on Emerging Information and Communication Technologies and their Social Consequences. In: Chiasson, M., Henfridsson, O., Karsten, H., DeGross, J.I. (Eds.), Researching the Future in Information Systems: IFIP WG 8.2 Working Conference, Future IS 2011. Turku, Finland, June 6–8, 2011. Proceedings. Springer, Heidelberg, pp. 59–76.
Stirling, A., 2007. Risk, precaution and science: towards a more constructive Policy debate. Talking point on the precautionary principle. EMBO Rep. 8 (4), 309–315.
Thorleuchter, D., Poel, D.V., Prinzie, A., 2010. A compared R&D-based and patent-based cross impact analysis for identifying relationships between technologies. Technol. Forecast. Soc. Chang. 77 (7), 1037–1050.
Tseng, Y.H., Lin, C.J., Lin, Y.I., 2007. Text mining techniques for patent analysis. Inf. Process. Manag. 43, 1216–1247.
Turney, P.D., Pantel, P., 2010. From frequency to meaning: vector space models of semantics. J. Artif. Intell. Res. 37, 141–188.
Wang, X., Qiu, P., Zhu, D., Mitkova, L., Lei, M., Porter, A.L., 2015. Identification of technology development trends based on subject–action–object analysis: the case of dye-sensitized solar cells. Technol. Forecast. Soc. Chang. 98, 24–46.
Ward, J.H., 1963. Hierarchical grouping to optimize an objective function. J. Am. Assoc. 58 (301), 236–244.
Yoon, J., Kim, K., 2011. Identifying rapidly evolving technological trends for R&D planning using SAO-based semantic patent networks. Scientometrics 88, 213–228.
Yoon, B., Park, Y., 2004. A text-mining-based patent network: analytical tool for high-technology trend. J. High Technol. Manag. Res. 15, 37–50.
Yoon, B., Yoon, C., Park, Y., 2002. On the development and application of a self-organizing feature map-based patent map. R&D Manag. 32 (4), 291–300.
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.
Zhang, L., 2011. Identifying key technologies in Saskatchewan, Canada: evidence from patent information. World Patent Inf. 33 (4), 364–370.
Zhang, Y., Porter, A.L., Hu, Z., Guo, Y., Newman, N.C., 2014. "term clumping" for technical intelligence: a case study on dye-sensitized solar cells. Technol. Forecast. Soc. Chang. 85, 26–39.
Zhu, D., Porter, A.L., 2002. Automated extraction and visualization of information for technological intelligence and forecasting. Technol. Forecast. Soc. Chang. 69 (5), 495–506.

Junegak Joung is currently PhD candidate in industrial engineering from Pohang University of Science and Technology (POSTECH). He received a BS on industrial engineering from POSTECH in 2013. His main research interests include technology management, technology planning, and patent mining.

Kwangsoo Kim received PhD degree in industrial engineering from University of Central Florida. He was researcher from Rochester Institute of Technology. He is currently a Professor in the Department of Industrial and Management Engineering at POSTECH. His research interests include business opportunity detection, tech-mining, and customer driven innovation.

Please cite this article as: Joung, J., Kim, K., Monitoring emerging technologies for technology planning using technical keyword based analysis from patent data, Technol. Forecast. Soc. Change (2016), http://dx.doi.org/10.1016/j.techfore.2016.08.020