Contents lists available at ScienceDirect

Technological Forecasting & Social Change

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

Discovering the potential opportunities of scientific advancement and technological innovation: A case study of smart health monitoring technology

Yung-Chi Shen[a,*], Ming-Yeu Wang[a], Ya-Chu Yang[a]

[a] *Department of BioBusiness Management, National Chiayi University, 580 Sinmin Road, Chiayi, Taiwan, ROC*

ARTICLE INFO

Keywords:
Text mining
ORCLUS
Cosine similarity of tf-idf vectors
Smart health monitoring

ABSTRACT

Detecting potential opportunities for scientific advancement and technological progress is vital for academia and technology-based firms. While patent data have become an extensively employed source for technological forecasting, scientific publications that represent advanced knowledge provide a potential for technological commercialization. Thus, the primary purpose of this study intends to discover potential opportunities for scientific advancement and technological innovation by comparing the information in scientific publications and patents. Accordingly, this study applies text mining and the arbitrarily ORiented projected CLUSter generation (ORCLUS) algorithm to cluster important scientific and technological topics. To identify potential opportunities for scientific advancement or technological commercialization, the cosine similarity of tf-idf vectors is used to detect the semantic similarity between clustered scientific fields and technological fields. Smart health monitoring technology is the case employed in this study. The results not only demonstrate the effectiveness of the text-based clustering approach and semantic similarity for exploring technological opportunities but also identify the potential patenting opportunities in the management of smart health monitoring and the potential scientific advancement in home health care systems.

1. Introduction

The primary objective of scientific research is to explore unknown frontiers and produce knowledge for the public rather than create new technologies and products for private sectors (Ogawa and Kajikawa, 2015). While Mertonian norms facilitate scientific research to benefit scientific communities by publishing scholars' findings (Merton, 1973), the fruits of basic research produce the seeds of innovation for industry (Ogawa and Kajikawa, 2015). According to Mansfield's investigation, 10% of new products and processes are generated by academic research (Mansfield, 1991). Hence, some studies have indicated that exploring potential opportunities for scientific advancement that contribute to technological progress is critical for innovation in industry (Leydesdorff et al., 1994; Kostoff and Schaller, 2001; Kostoff, 2008; Kajikawa et al., 2008; Shibata et al., 2008; Ogawa and Kajikawa, 2015).

Similarly, research and development (R&D) in industry aim to maintain competitiveness and sustain growth by commercializing new technologies for technology-based enterprises (McNamara and Baden-Fuller, 1999). Promising technological opportunities, which are defined as a set of possibilities or the potential for technological advancement or progress in general or within a specific field (Klevorick et al., 1995; Olsson, 2005; Yoon et al., 2014), influence the performance of firms and shape the landscape of industry (Wang et al., 2015). Therefore, the identification of potential technological opportunities has been a major challenges in the technology management literature (Yoon and Park, 2005; Yoon, 2008; Shibata et al., 2010; Yoon and Kim, 2012; Ma et al., 2014; Lee et al., 2014; Kim et al., 2014; Wang et al., 2015).

According to Ogawa and Kajikawa (2015), however, researchers have to engage two difficulties during the identification of academic areas in terms of the potential for innovative seeds in industry and technological opportunities. First, the innovations that are generated from academic research require years to be commercialized. Moreover, the aim of academic research is not always to contribute to industrial applications. Therefore, assessing whether the academic areas have the potential for industrial applications is difficult (Ogawa and Kajikawa, 2015). In addition, reading papers and patents is always an important means to observe and analyze R&D trends (Leydesdorff et al., 1994; Ogawa and Kajikawa, 2015). Nevertheless, researchers need to engage the second difficulty, namely, the difficulty in separately

[*] Corresponding authors.
E-mail addresses: syc@mail.ncyu.edu.tw (Y.-C. Shen), mywang@mail.ncyu.edu.tw (M.-Y. Wang), s1031504@mail.ncyu.edu.tw (Y.-C. Yang).

https://doi.org/10.1016/j.techfore.2020.120225
Received 1 November 2018; Received in revised form 14 March 2020; Accepted 21 July 2020
Available online 04 August 2020

reviewing the rapidly increasing number of papers and patents when exploring the potential opportunities for scientific advancement and technological innovation. When manually reading papers or patents, some emerging research areas are probably overlooked (Ogawa and Kajikawa, 2015).

Many scholars suggest citation-based methods to overcome the two previously mentioned difficulties (Lai and Wu, 2005; Small, 2006; Kajikawa et al., 2008; Kajikawa and Takeda, 2008; Shibata et al., 2008; Ho et al., 2014; Ogawa and Kajikawa, 2015; You et al., 2017). Citation-based methods help to identify major research clusters in a citation network without reviewing every paper or patent (Ogawa and Kajikawa, 2015). However, citation-based methods only analyze the large clusters that contain a relatively large number of papers or patents and disregard small clusters that consist of relatively fewer papers and patents. The omission of some small clusters in citation networks risks the potential loss of information and promising research areas (Wang et al., 2015). To resolve this disadvantage of citation-based methods, another stream of studies adopts text-based approaches to explore potential technological opportunities (Yoon and Park, 2005; Ma et al., 2014; Yoon et al., 2014; Lee et al., 2015; Wang et al., 2015). Furthermore, Wang et al. (2015) suggest the arbitrarily ORiented projected CLUSter generation (ORCLUS) algorithm, which is a *k*-means-like approach to cluster scientific and technological documents without missing information.

Regarding the data source, past research has usually employed a single type of data source to explore potential opportunities for scientific advancement and industrial technologies. For example, Shibata et al. (2011a) and Huang and Chang (2014) adopted the ISI Web of Knowledge as their data source to explore the research fronts in the fields of regenerative medicine and organic light emitting diodes (OLEDs). Yoon and Park (2005), Ma et al. (2014), Yoon et al. (2014), and Lee et al. (2015) employed patents to detect technological opportunities in the areas of thin film transistor-liquid crystal displays (TFT-LCDs), dye-sensitized solar cells, light emitting diodes (LEDs), and thermal management technology of LEDs, respectively. However, Watt and Porter (1997) illustrate a way for innovation forecasting by combining basic research and industrial applications. Shibata et al. (2010) indicate that R&D managers and policy makers should understand the relationship between science and technology. Moreover, Ogawa and Kajikawa (2015) indicate that detecting emerging fields in academic research is not sufficient for R&D strategic planning because academic research does not always serve to develop industrial applications. In addition, as R&D targets for firms, the technological opportunities extracted by patent analysis seem to be out-of-date because the technologies analyzed by patent analysis have been patented (Ogawa and Kajikawa, 2015).

To address the deficiency of a single type of data source, Shibata et al. (2010) and Ogawa and Kajikawa (2015) suggest employing scientific publications and patents as data sources to detect promising academic research areas and potential technological opportunities. Previous studies indicate that technological innovations are presented in scientific publications and then disseminated to patents (Watt and Porter, 1997; Martino, 2003). On the other hand, a feedback stimulus from industrial technology is able to facilitate continuous exploration and study of scientific knowledge (Klevorick et al., 1995; Meyer, 2002; Glänzel and Meyer, 2003). To detect emerging research areas, these studies provide a theoretical foundation by simultaneously using papers and patents as data sources for academic research and industrial R&D. Especially, Shibata et al. (2010) propose a model to cluster promising areas for academic research and industrial R&D by comparing papers and patents with citation network analysis. Furthermore, to uncover the potential opportunities of academic research and industrial R&D in the field of microalgal biofuels, Wang et al. (2015) adopt text mining and ORCLUS to improve the model developed by Shibata et al. (2010).

Nonetheless, the two studies conducted by Shibata et al. (2010) and Wang et al. (2015) use experts' subjective judgments to assess the similarity between science and technology. The less similar clusters of academic research and industrial R&D are identified as potential opportunities for scientific advancements and patenting, respectively. Because the experts' judgments may be restricted by their domain knowledge, the potential opportunities acknowledged by experts might not be comprehensive. Additionally, considerable time and effort is needed for experts to assess the relationships between science and technology (Wang et al., 2015). To improve the disadvantage of the subjectiveness of experts' judgments, Wang et al. (2015) suggest applying the cosine similarity of tf-idf vectors demonstrated by Shibata et al. (2011b) to estimate the similarity between the clusters of academic research and industrial R&D.

As mentioned in the previous paragraphs, despite ORCLUS and the cosine similarity of tf-idf vectors as the remedies suggested by previous studies, no study has simultaneously identified the advantages of these methods. To bridge this research gap, this study intends to optimize the methods to discover the potential opportunities for scientific advancement and technological innovation. This study integrates ORCLUS and the cosine similarity of tf-idf vectors to extract the potential opportunities for scientific advancement and technological innovation from the comparison between papers and patents. Smart health monitoring technology is adopted to demonstrate the methods applied in this study because this technology is dependent on rigorous fundamental research to provide diversified physiological signal monitoring functions. The findings can serve as a reference for future scientific advancements in academia and technological development in industry.

2. Literature review

2.1. Relationship between science and technology

As mentioned in the previous section, the similarity between science and technology provides a critical foundation for detecting the potential opportunities of scientific and technological progress. From the existing literature, the relationship between science and technology can be separated into several facets. First, Wang et al. (2015) indicate that the study of the relationship between science and technology can be traced back to Price's work (Price, 1965), which shows the interaction between science and technology in some cases. Many studies have noticed the relationship between science and technology. For instance, Rosenberg (1982, 1990) indicates that scientific research can stimulate technological development and innovation. On the other hand, scientific advancement needs support from advanced technologies (Nelson, 1982). According to the investigation of Pavitt (1991), a strong link exists between basic science and technology in chemical and pharmaceutical industries. The electronics industry is linked to applied research (Pavitt, 1991; Chaves and Moro, 2007). Meyer (2002) and Petrescu (2009) also argue that science and technology are interdependent. Furthermore, based on data from scientific papers and patents as proxies of science and technology, Chaves and Moro (2007) discovered that science moves technology and technology influences scientific development.

Second, the perspectives on the new production of knowledge (Gibbons et al., 1994), entrepreneurial science (Etzkowtiz, 1998), or the Triple Helix (Leydesdorff and Etzkowtiz, 1996) may explain the relationship between science and technology. Gibbons et al. (1994) propose the "Mode 2″ paradigm to describe knowledge production that relies on the collaboration among interdisciplinary teams for short periods to work on specific problems in the real world. Mode 2 knowledge production is then conceptualized in terms of university-industry-government relations, i.e., the Triple Helix model (Leydesdorff and Etzkowtiz, 1996). Etzkowtiz (1998) discusses many different forms of university-industry linkages by which scientific knowledge can be commercialized under the appropriate regime for industry.

Third, some researchers observe the relationship between science and technology as a cyclical development (Grupp & Schmoch, 1992; Schmoch, 2007). Bush (1945) proposes a science-push model, including the stages of pure basic research, applied research, experimental development, and innovation in sequence, to illustrate the innovation process. Furthermore, some researchers indicate the existence of technology or demand rather than science at the beginning of innovation and replace basic research by technological discovery or market demand (Schmoch, 2007). Thus, the science-push model is transferred to a technology-push or market-pull model (Schmoch, 2007). In addition, Schmoch (2007) analyzes patent applications, scientific publications, and market figures to conclude a double-boom development, i.e., a development in two cycles, with an initial maximum development cycle, subsequent decline and second growth stage (Schmoch and Thielmann, 2012).

Last, Murray (2002) assesses how science is commercialized via patent-paper pairs. Inspired by Murray's work, several studies adopt patent-paper pairs to analyze the knowledge flow between academia and industry (Murray and Stern, 2007; Fehder et al., 2014; Thompson et al., 2018) and the interaction between science and technology (Chang et al., 2017).

Despite abundant studies that provide explanations for the relationship between science and technology with different perspectives, only a few works acknowledge that this commensurate relationship is not effective for uncovering the potential opportunities for scientific and technological progress (Shibata et al., 2010; Wang et al., 2015; Ogawa and Kajikawa, 2015). Shibata et al. (2010) note the dissimilarity between science and technology and propose three types of relationships between science and technology, as illustrated in Fig. 1, to reveal the potential opportunities for future scientific and technological development. As presented in Fig. 1, Area A is a field that contains both technological and scientific documents, where science and technology interact and coevolve. Area B is an existing technological field without scientific research. Because this area may require advanced scientific support, an opportunity for scientific research advancement exists. Area C is a field in which scientific research has no corresponding technological developments, and it presents a gap between science and technology; this gap indicates potential technological development or commercialization opportunities (Shibata et al., 2010; Wang et al., 2015). Because Shibata et al. (2010) recognize the dissimilarity between science and technology as potential opportunities, this study adopts the three types of relationships between science and technology as the basis to identify the potential opportunities for scientific advancement and technological progress.

2.2. Analyses of scientific advancement and technological opportunities

Studies about opportunities for scientific advancement are relatively scarce. Only Ogawa and Kajikawa (2015) attempt to identify the industrial opportunities for academic research. A concept close to an estimation of the potential for scientific advancement is the research front introduced by Price (1965). Identifying research fronts helps to understand the latest developments in a particular field and not only provides insights into current focuses and future trends but also serves as an important indicator for government decision-making on technology policy (Huang and Chang, 2015). Many researchers have focused on detecting research fronts in different technology fields, such as nanotubes (Kuusi and Meyer, 2007), gallium nitride (Shibata et al., 2008), complex networks, and nano-carbon (Fujita et al., 2014). In these studies, bibliographic coupling analysis and co-citation analysis are extensively employed.

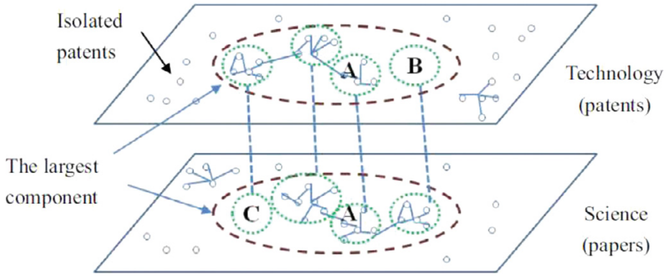

Fig. 1. Relationships between science and technology.
Source: Wang et al. (2015).

According to Ma et al. (2014), an analysis for technological opportunities was first presented in the work of Porter and Detampela (1995), which combines monitoring and bibliometric analyses to obtain effective intelligence on emerging technologies. Many studies became engaged in related topics, including relationships between technology technological opportunities and innovation (Vertova, 2001; Nieto & Quevedo, 2005) and the mining of technological opportunities (Yoon, 2008; Kim et al., 2014). The approaches for identifying technological opportunities include large database analysis (Kostoff and Schaller, 2001), morphology analysis (Yoon and Park, 2005; Yoon, 2008), patent analysis (Yoon and Kim, 2012; Ma et al., 2014; Lee et al., 2014) and text mining (Yoon and Park, 2005; Yoon et al., 2014).

Nevertheless, previous studies of technological opportunity exploration employed a single knowledge source, which overlooks the windows generated by the interaction between science and technology (Wang et al., 2015). Moreover, some studies show that the development of industrial technology continues the exploration of scientific knowledge (Meyer, 2000, 2002). Therefore, the gaps between science and technology imply the potential for technology innovations. Scientific discoveries may have industrial applications with which scientists are unfamiliar. Thus, manufacturers often do not realize which scientific discoveries can improve their technological innovation (Hellmann, 2007; Wang et al., 2015). In many studies, scientific publications and patents are regarded as the results of scientific research and technological development (Martino, 2003; Wang et al., 2015). A cross-database comparison between scientific papers and patents extracts insight into potential uncommercialized gaps and technological opportunities (Shibata et al., 2010; Wang et al., 2015). Therefore, the three types of relationships between science and technology proposed by Shibata et al. (2010), as mentioned in the previous section, are employed in this study to detect the gap between science and technology for identifying potential opportunities.

2.3. Methods for clustering technological documents

Technological documents contain important research results that are valuable to the industry, research institutions, and policy makers (Tseng et al., 2007). To extract useful information, these technological documents must be classified. Although scientific and patent databases categorize the collected papers and patents, their classification systems are too general to satisfy the needs for technological forecasting, research planning, technological positioning or strategy-making (Archibugi and Pianta, 1996; Lai and Wu, 2005). The clustering methods for grouping or classifying technological documents can be divided into two different streams, i.e., citation-based approaches and text-based approaches.

2.3.1. Citation-based approaches

The citation-based approaches are based on bibliometric analysis, which is proposed by Narin (1994), to extract valuable information from bibliometric data, such as technological document numbers, authors, assignees, titles, date of application, applicant country, and international patent classifications. Some statistical methods are usually applied to group the bibliometric data for further analysis. For instance, Huang et al. (2003) employ bibliographic coupling analysis with multidimensional scaling to graphically categorize patents. Lai and Wu (2005) use co-citation analysis to assess the similarity between patent pairs and then apply factor analysis to classify the patents. Chang et al. (2009) apply hierarchical cluster analysis to analyze a patent citation network for exploring technology diffusion and classification of business methods.

In addition, many studies use citation analysis to cluster patents for technological planning, technological forecasting or strategy-making. For example, Shibata et al. (2010) apply citation analysis to group patents for identifying the commercialization gap in solar cell technology. Gao et al. (2012) cluster a document co-citation network to analyze the interaction between science and technology. Érdi et al. (2013) analyze a patent citation network to predict the emerging technology fields. Ogawa and Kajikawa (2015) cluster a citation network of academic papers and patents to assess the industrial opportunities for fundamental research.

Nevertheless, Hsu et al. (2006) indicate that a citation analysis has the following drawbacks. First, the links between two documents could be positive or negative and the relationship may produce an inaccurate analysis (Kostoff, 1998). Second, citation analysis will be time consuming if a machine-readable form is not available (Karki, 1997). The scope of citation analysis is limited because documents without citations may be disregarded (Hsu et al., 2006). Wang et al. (2015) also suggest that citation network analysis explores large document clusters that contain a large number of citation networks but disregards small and isolated clusters without connections between a small cluster and a large cluster. Thus, an increasing number of studies seek text-based approaches to address technological document clustering.

2.3.2. Text-based approaches

Text mining extracts effective, non-trivial, hidden, previously unknown and potentially useful knowledge from non-structured or semi-structured texts (Weiss et al., 2005; Feldman and Sanger, 2007). Context information in papers and patents is an example of unstructured text. Numerous studies have adopted text mining or combined text mining with other methods to extract a large volume of unstructured text and mine hidden information from a set of documents. For example, Yoon and Park (2005) combined morphology analysis with text mining to identify technology opportunities within patent data. Wang et al. (2010) combined text mining and TRIZ to investigate technological evolutionary trends. Using patent information, Choi et al. (2013) adopted a SAO-based text mining approach for technology road mapping.

The text mining process includes the preprocessing of document collections, storage of intermediate representations, and techniques to analyze the intermediate representations and visualization of the results (Feldman and Sanger, 2007). Document preprocessing involves stemming, stopword removal and extraction of representative words or terms, which are referred to as the features of the document.

The purpose of intermediate representations is to transfer each document with unstructured texts into a numerical vector that captures a document's characteristics. One class of representation methods involves the use of the vector space model (VSM) proposed by Salton et al. (1975) to present the relationship between documents and features, where the element value of vectors mainly depends on the occurrence frequencies of the features in the documents. For decades, numerous studies have shown that the method can effectively solve various text mining tasks. A major advantage of the method is that it produces intuitively interpretable document vectors. However, the method fails to preserve accurate proximity information when the number of unique features increases and disregards the impact of semantically similar features for preserving document proximity (Kim et al., 2017). Recently, with the development of neural language processing, many neural-network-based document representations are proposed, such as Word2Vec (Mikolov et al., 2013), GloVe (Pennington et al., 2014) and BERT (Devlin et al., 2018). The class of neural-network-based methods creates low dimensional vectors that successfully preserve the proximity information by using contextual information of each word and document to embed documents into a continuous vector space with a manageable dimension. However, each feature of document vectors fails to provide intuitive interpretability because its value indicates the weight of the neural network that is used to train the methods (Kim et al., 2017). This study intends to provide insights about the potential opportunities of smart health monitoring technology; therefore, this study adopts VSM-based methods to represent documents because the interpretability is essential to provide practical insights.

Clustering documents helps to categorize a collection of documents with an unknown structure to learn the hidden knowledge (Weiss et al., 2005). Cluster detection is based on the similarity among documents and is typically determined using measures of the dimensions in a vector space (Wang et al., 2015). When the number of dimension increases, however, all documents in the high-dimensional vector space will be nearly equidistant from each other (Parsons et al., 2004). Conventional clustering algorithms ineffectively detect meaningful clusters because of high-dimensional space (Beyer et al., 1999; Wang et al., 2015). In addition, a vector space generated from a set of documents, namely, a corpus, is usually very sparse and features many zero values in a matrix (Kriegel et al., 2009). Some features may be irrelevant to the themes, which may confuse clustering algorithms (Parsons et al., 2004; Wang et al., 2015).

Previous studies used feature transformation or feature selection techniques to cluster objectives in high-dimensional space (Wang et al., 2015). The two streams of techniques conduct clustering in a global space by computing only one subspace of the original data space in which the clustering can be performed (Kriegel et al., 2009; Müller et al., 2009). They ineffectively detect clusters, because each cluster may exist in a different subspace (Kriegel et al., 2009; Wang et al., 2015).

Recently, based on the concept of feature selection, some studies have developed clustering algorithms for a high-dimensional space by separately selecting relevant subspaces for each cluster. The first class of algorithms, named projected clustering or subspace clustering algorithms, focuses on finding clusters in axis-parallel subspaces. The second class of algorithms that search for subspace solutions, where a cluster may exist in any arbitrarily oriented subspace, is referred to as generalized subspace/projected clustering or correlation clustering algorithms. The third class of algorithms is referred to as pattern-based clustering algorithms, and this class falls between the previous two algorithms. The algorithms in the first class define the closeness in terms of the Euclidean distance in an axis-parallel projection. The pattern-based clustering algorithms define the closeness in terms of a common behavior of objects in an axis-parallel subspace, that is, with respect to a certain pattern in which the objects form a subset of attributes.

In the first class, the algorithms to search relevant subspaces for clusters can be divided into top-down and bottom-up approaches. The top-down approach attempts to anticipate cluster members and then determines the subspace of each cluster. Well-known algorithms that apply the top-down approach include PROCLUS (PROjected CLUStering) (Aggarwal et al., 1999). The algorithms of FINDIT (Woo et al., 2004) and SSPC (Yip et al., 2005) are variations of PROCLUS. Furthermore, the bottom-up approach attempts to anticipate the subspace of the clusters and then determines the cluster members. CLustering In QUEst (CLIQUE) is the first bottom-up subspace clustering algorithm (Aggarwal et al., 1998).

Among the three classes of clustering algorithms, based on the application viewpoint, the second class is generalized, concise, and meaningful compared with the other two classes. ORCLUS, which is an extended version of PROCLUS, is the first proposed generalized projected clustering algorithm (Aggarwal and Yu, 2000). This algorithm arose from the observation that many datasets contain inter-feature correlations and is a *k*-means-like approach. The algorithm consists of a certain number of iterations, in which each of the following three steps are applied: assigning clusters, finding subspaces, and merging clusters. During cluster assignment, the algorithm iteratively assigns each object to its closest cluster center. The distance between two points is measured in subspace E, where E is a set of orthonormal vectors in some l_c-dimensional subspace. Locating the subspaces redefines the subspace E associated with each cluster by calculating the covariance matrix for a cluster and selecting the orthonormal eigenvectors with the smallest eigenvalues. The selected eigenvectors correspond to the projected subspace, where the clustered objects exhibit high density, and hence, can exclude noisy subspaces. Calculation and selection are iteratively adapted to the current state of the updated cluster, and the dimensionality l_c reduces from iteration to iteration. As a result, each successive iteration continues to strip noisy subspaces from different clusters. The third step merges the closest pairs of clusters that have similar directions of high density. The number of clusters k and the subspace dimensionality l must be specified by the researcher. The algorithm terminates when the merging process over all the iterations has reduced the number of clusters to k. At this moment, the dimensionality l_c of the subspace E that is associated with each cluster is also equal to l. The cluster sparsity coefficient can evaluate the choice of subspace dimensionality (Aggarwal and Yu, 2000; Kriegel et al., 2009; Parsons et al., 2004). If the value approaches 1, then the chosen subspace dimensionality may be too large. A value close to 0 can be interpreted as a hint that a strong cluster structure has been identified (Szepannek, 2013).

2.4. Smart health monitoring technology

With the rapid increase in life span and a decline in fertility in recent decades, members of the older population and patients who require health monitoring have increased. Therefore, the cost of hospitalization and patient care is predicted to increase worldwide. In the US, for example, the hospitalization cost for patients who suffer sentinel events associated with incorrect medication, dosage inaccuracies, contraindications or critical delays in interventions is between 1.5 billion and 5 billion dollars annually (Williams, 2004; Baig and Gholamhosseini, 2013). Health monitoring technology can reduce hospitalization, the burden of medical staff, consultation time, waiting lists and overall healthcare costs (Baig and Gholamhosseini, 2013).

Baig and Gholamhosseini (2013) indicate that smart health monitoring, which is referred to as advanced technology or a new approach to health monitoring, usually applies smart devices or "smart" approaches to address health issues. The scope of smart health monitoring, as suggested by Baig and Gholamhosseini (2013), includes wearable health monitoring, remote health monitoring, and mobile health monitoring. Wearable health monitoring focuses on the applications of wearable devices or biosensors. Remote health monitoring refers to applications with remote access or systems that can exchange data with a remote location (Baig and Gholamhosseini, 2013). The previously described smart health monitoring systems fall within the scope of digital health, including mobile health, health information technology, wearable devices, telehealth and telemedicine, as defined by the U.S. Food & Drug Administration (FDA) (U.S. Food and Drug Administration, 2019). The World Health Organization (WHO) also defines eHealth as the use of information and communication technologies for health and acknowledges the potential major role of eHealth in improving public health (World Health Organization, 2018).

Several studies indicate that technological issues and challenges sill require fundamental research in the field of smart health monitoring technologies for assisting remote diagnosis, disease treatment (Chan et al., 2012; Baig and Gholamhosseini, 2013; Liu et al., 2016), chronic disease management, and protective systems against emerging infectious diseases (Chang and Oyama, 2018). In addition, a large variety of laboratory prototypes, test beds and industrial products are under development in academia and industry, as reported by some studies (Chan et al., 2012; Liu et al., 2016). Moreover, the American College of Cardiology (ACC) has explored state-of-the-art digital health and disclosed a road map for innovation in this area (Bhavnani et al., 2017). Thus, potential opportunities for scientific advancement and technological innovation discovered in the field of smart health monitoring technology can be validated with these abundant existing studies. The proposed methods can also be verified. Therefore, smart health monitoring technology is adopted as a case to demonstrate the methods presented in this study.

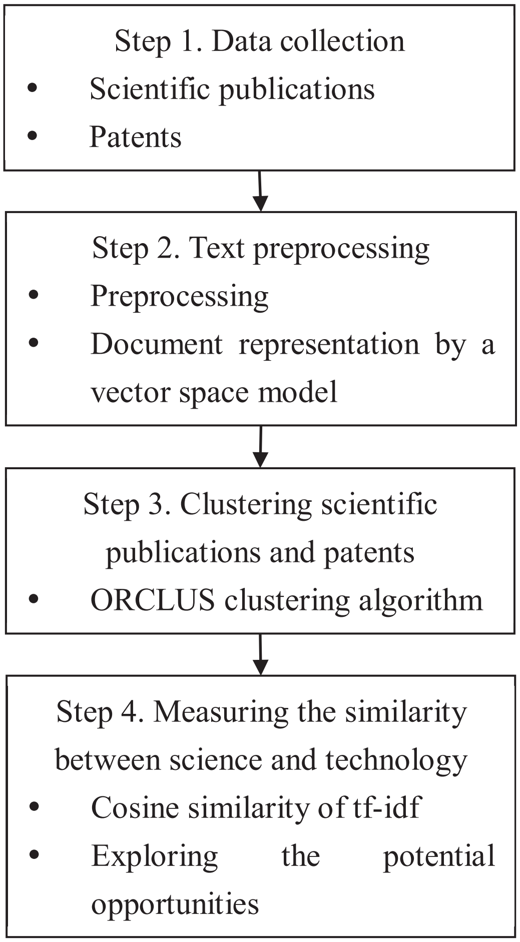
Step 1. Data collection
• Scientific publications
• Patents
Step 2. Text preprocessing
• Preprocessing
• Document representation by a vector space model
Step 3. Clustering scientific publications and patents
• ORCLUS clustering algorithm
Step 4. Measuring the similarity between science and technology
• Cosine similarity of tf-idf
• Exploring the potential opportunities

Fig. 2. Research process.

3. Research process and results

As shown in Fig. 2, the research process in this study contains four steps: data collection, text preprocessing, clustering documents for scientific and technological fields and measuring the similarity between science and technology.

3.1. Step 1. data collection

The primary research purpose of this study is to explore potential technological opportunities by comparing science and technology databases. The databases employed in this study include the Web of Science to collect Science Citation Index (SCI) publications and the patent database of the United States Patent and Trademark Office (USPTO). The SCI is a typical source for basic research. The patent database of the USPTO is employed in this study to extract information with regard to applied technology because combined with the territoriality of patent protection, the US market is important for technology transfer and international trade and lures inventors to file patent applications in the US (Lai and Wu, 2005).

To address as many smart health monitoring technology documents as possible, this study collects keywords by the literature review and consultations from two experts. The first expert is an academic researcher who has been dedicated to the field of health monitoring technology for over ten years. The second expert is an industrial practitioner who is a senior researcher at the Industrial Technology Research Institute, which is the most important nonprofit R&D organization that is engaged in applied research and technical services in Taiwan. As suggested by Baig & Gholamhosseini (2013) and the two experts, the keywords employed in this study include remote health monitoring, remote health care, remote healthcare, wearable health care, wearable healthcare, wearable health monitoring, mobile health care, mobile healthcare, mobile health monitoring, smart health care, smart healthcare, smart health monitoring, intelligent health care, intelligent healthcare, intelligent health monitoring, e-health, telehealth and telecare.

The search field in the SCI contains the abstract and keywords, and the search fields in the patent database include the title, abstract and claim. The timeframe of the retrieved papers covers the time in which the keywords first appeared in these databases through December 31, 2015. As a result, this study obtains 9865 scientific papers and 930 patents, respectively. To ensure that scientific papers in the discipline of smart health monitoring technology are employed, by consulting the two previously mentioned experts, this study manually excludes the research areas that are not related to health care, such as soil science, particle physics, nuclear physics, atomic molecular chemical physics, mineralogy and limnology. Scientific documents retain articles. Proceeding papers, book chapters and retracted publications are removed. A total of 4543 scientific papers are retained.

3.2. Step 2. text preprocessing

The second step addresses the text preprocessing and document representation by VSM. In the text preprocessing, this study eliminates word suffixes to retrieve their radicals. The preprocessing step of text mining includes the following tasks: removing stopwords, numbers and punctuation; eliminating whitespace; converting characters to lower case; and unifying synonyms. Extracted terms that are referred to as sparse terms do not appear in most documents. In this study, because of their low usefulness, the sparse terms are excluded to reduce noise. In addition, the adjectives and nouns with similar meaning are unified. For example, the adjective “diagnostic” and the noun “diagnosis” are regarded as a synonym and unified as “diagnosis.” With the assistance of the two experts’ suggestions, this study establishes two synonym lists to unify synonyms in the scientific publications and the patents.

In this stage, the document representation by VSM is used to organize a document-by-term matrix, where each cell highlights the term's importance in a given document. The document-by-term matrices are established for papers and patents. The extensively applied term frequency-inverse document frequency weighting measures is adopted in this study (Feldman and Sanger, 2007). As a result, a total of 14,428 words and 3121 words from the paper corpus and patent corpus, respectively, are obtained. To reduce noise, terms with high zero-entries, i.e., high sparsity, in the document-by-term matrix are excluded. In this study, the sparsity for the paper corpus is set to 96%, as suggested by the two experts. Terms with more than a 96% zero-entry level in the derived document-by-term matrix are excluded. A 96% sparsity-level setting retained 141 terms. For the patent corpus, the sparsity level is set to 98%, as suggested by the two experts. The sparsity settings produced a document-by-term matrix of 4543 × 139 for the paper corpus and a document-by-term matrix of 930 × 138 for the patent corpus. For the two corpuses, a TF-IDF weighting is applied to measure the term weights in the cells of the two matrices for the corresponding documents.

3.3. Step 3. clustering scientific publications and patents

In the third step, ORCLUS is used to cluster papers and patents. In the implementation of ORCLUS, four parameters need to be prespecified: the final number of clusters (k), dimensionality of subspaces where the final clusters are concentrated (l), initial number of clusters (k_0) and factor for the cluster number reduction in each iteration of the algorithm (a and $a < 1$). The area of smart health monitoring technology comprises different underlying knowledge streams and technologies, revealing that smart health monitoring technology is involved in several scientific and technological fields. Thus, Wang et al. (2015) suggest that the final number of cluster k ranges from 5 to 10 for each corpus when performing ORCLUS. The experiments performed by Aggarwal and Yu (2000) revealed that ORCLUS performs well when the specified dimensionalities are between 2 and 8. The optimal value of dimensionality is 6. This study tends to widen the range and attempts dimensionalities from 4 to 12. A large value of k_0 is chosen to enable the iterations to begin with a larger number of seeds, improving the likelihood that each cluster will be covered by at least one seed (Aggarwal and Yu, 2000; Wang et al., 2015). This study adopts *tm, Snowball* and *ORCLUS* packages in R language for preprocessing, document representation and ORCLUS algorithm implementation.

The cluster sparsity coefficients provided by ORCLUS are applied to preliminarily determine the appropriate parameter settings. Aggarwal and Yu (2000) argued that the smallest sparsity coefficient is 0.002 (Wang et al., 2015). To inherit the same ORCLUS algorithm adopted by Wang et al. (2015), the specified subspace dimensionalities l in this study are not less than ($k - 2$), where the specified final clusters k range between 5 and 10, and the specified initial number of cluster k_0 are, or approach, the greatest value that computers can handle. Most sparsity coefficients are less than 0.001 when performing ORCLUS to cluster papers and patents, according to the report of Wang et al. (2015). In addition, the two experts are consulted to determine the threshold. By comparing the reality of smart health monitoring technology, they have perceived the filtered results. The threshold that can reflect the reality of the technology development should be selected. By reviewing the titles of papers and patents within the scientific and technological clusters, this study sets a threshold of 0.001, as suggested by Wang et al. (2015) and the two experts. The parameter settings, whose cluster sparsity coefficient is higher than the threshold value, are eliminated.

Based on different parameter settings to produce parameter setting candidates, ORCLUS is performed two times. Clusters are determined after identifying terms with high centroids and reading document titles. The two experts are requested to read the titles of papers and patents within the clusters that are generated by parameter setting candidates to assess the parameter settings. The ideal parameter setting should reveal the circumstance of smart health monitoring technology recognized by the experts. After several runs of iterations, the most realistic and stable clustering results correspond to a parameter setting with the cluster number k of 10 and the subspace dimensionality l of 11 for the paper corpus and correspond to a parameter setting of k at 6 and l at 8 for the patent corpus. The cluster sparsity coefficients are 0.00099 for the paper corpus and 0.00071 for the patent corpus. As a result, the ORCLUS algorithm produces 10 clusters for the paper corpus and 6 clusters for the patent corpus. For the sequence analysis, SF1 to SF10 denotes the ten scientific fields, and TF1 to TF6 denotes the six technological fiends.

Table 1 and Table 2 present the scientific fields and technological fields identified by ORCLUS. In Tables 1 and 2, the main areas in Web of Science Categories, main International Patent Classification (IPC), and representative terms with high centroids identified by ORCLUS help to interpret the characteristics of the scientific and technological fields. The two experts help name these fields based on the main areas in the Web of Science Categories, main IPC, representative terms, and referring the titles and abstracts of papers and claims of patents within the

Table 1
Scientific fields in paper documents.

SF	Count	Mean year	Main areas	Representative terms	Field name
SF1	1731	2009.6	Computer science, information systems	composition, damage, detect, ecg, estimate, frequency, material, signal, structure, technique	Healthcare applications focusing on psychological signal measurement
SF2	200	2012.1	Health care sciences & services	analysis, body, communication, connect, energy, inform, medicine, message, network, node, privacy, protocol, scheme, server, simulate, smart, telecare, wireless	Security and privacy of healthcare systems
SF3	131	2012.3	Health care sciences & services	behavior, function, interface, monitor, physiology, rate, scale, scenario, sensor, symptom, transmission, wearable	Sensors for wearable health or medical devices
SF4	143	2010.2	Computer science, information systems	architecture, assist, clinician, collect, data, database, environment, healthcare, live, mechanic, mobile, model, organ, prototype, software, system, ubiquity	Infrastructure of health information networks
SF5	134	2011.3	Computer science, information systems	active, adult, algorithm, automatic, device, home, image, individual, life, phone, platform, remote, transfer	Medical imaging platforms
SF6	138	2009.5	Computer science, information systems	compute, digit, electron, engine, facility, health, infrastructure, intelligent, medic, safety, solution, technology	Health information processing technology
SF7	148	2010.2	Health care sciences & services	administer, assess, doctor, emergency, equipment, hospital, internet, link, online, staff, telemedicine, therapy	Diagnosis, treatment or therapy through telemedicine
SF8	129	2010.3	Computer science, information systems; health care sciences & services	component, direct, evaluate, framework, methodology, source, telephone, video, web	The management of smart health monitoring
SF9	1649	2010.3	Health care sciences & services	chronic, community, heart, impact, intervention, program, report, telehealth, treatment	Telehealth for chronic diseases
SF10	140	2010.6	Health care sciences & services	accuracy, blood, clinic, comparison, control, diabetes, diagnosis, disease, measure, method, patient, physician, statistic, weight	Patients' psychological signal measurement

Table 2
Technological fields in patent documents.

TF	Count	Mean year	Main IPC	Representative terms	Field name
TF1	250	2009.9	G06Q	audio, blood, cellular, device, document, download, electric, glucose, house, machine, number, pharmacy, phone, physic, software, technique, telephone, video, visual, wearable	Health information infrastructure in hospitals
TF2	29	2009.9	H04N	body, calculate, care, circuit, clinic, computer, data, detect, determine, digit, disease, display, drug, emergency, graphic, handheld, image, material, mechanic, media, memory, method, mobile, organ, parameter, personnel, physiology, predetermine, storage, terminal, time, transfer, transmission	Medical imaging management systems
TF3	66	2010.5	G06F	action, alarm, assess, call, camera, component, electron, embody, engine, home, measure, message, module, platform, portable, safety, signal, switch, telecommunication, therapy, web	Home health care systems
TF4	45	2008.6	G06Q	analysis, apparatus, automatic, bed, biometrics, caregiver, code, communicate, database, diagnosis, evaluate, facility, healthcare, hospital, individual, monitor, person, prescript, processor, program, reader, record, remote, screen, script, smart, store, system, transmit	Health monitoring systems
TF5	120	2010.3	G06F	access, account, alert, appliance, client, environment, intelligent, link, network, node, process, protocol, sensor, structure, technology, track, transact, voice, wireless	Data communication systems
TF6	420	2009.9	G06Q	decision, equipment, health, heart, line, medic, model, patient, physician, procedure	Decision support systems for diagnosis

scientific fields and technological fields clustered by ORCLUS.

3.4. Step 4. measuring the similarity between science and technology

After clustering the scientific publications and patents in the previous step, step 4 intends to measure the similarity between science and technology. Shibata et al. (2011b) compare the Jaccard coefficient, cosine similarity of tf-idf vectors, and cosine similarity of the log-tf-idf vector to explore the semantic similarity between the clusters of scientific publications and the clusters of patents and conclude that the cosine similarity of tf-idf vectors perform the best in detecting the similarity between scientific clusters and patent clusters. Thus, this study adopts the cosine similarity of tf-idf vectors to measure the similarity between scientific clusters and patent clusters.

The cosine similarity of tf-idf vectors assumes that the greater is the number of common terms two clusters, the closer are the two document clusters. Therefore, the similarity between the scientific clusters and the patent clusters can be measured by the cosine similarity of the tf-idf vectors (Shibata et al., 2011c). The cosine similarity of the tf-idf vectors Cosine(s, t) between the scientific cluster s and the patent cluster t is defined as

$$sim_{tfidf}(s, t) \equiv Cosine(s, t) = \vec{w_s} \cdot \vec{w_t} = \sum_i w_s^{(i)} w_t^{(i)}$$

where

$$w_c^{(i)} = \left(\sum_{d \in c} tf_{i,d}\right) \times \log\left(\frac{N}{df_i}\right),$$

$td_{i,d}$ is the number of occurrences of the ith term in document d, df_i is the number of documents that contain the ith term, N is the total number of documents and $\|w\|$ is normalized as $\|w\| = 1$ to remove the effect of the difference of the document length (Shibata et al., 2011).

The similarity between science and technology is presented in Table 3. To prevent information overload for practical application, establishing a threshold to eliminate less similar relationships is necessary. A suitable threshold should reflect the reality of technology development. Thus, the two experts compare the remaining relationships between science and technology with the status quo that they recognize to assess the appropriate thresholds. As a result, the third quartile (0.555) suggested by the two experts is applied as the threshold to eliminate the less similar relationships because the filtered relationships between science and technology better match the practice perceived by the experts.

As shown in Table 3, the following scientific and technological fields are semantically similar: SF1 and TF1, SF1 and TF5, SF2 and TF5, SF2 and TF5, SF2 and TF6, SF3 and TF1, SF4 and TF1, SF4 and TF6, SF5 and TF1, SF5 and TF2, SF5 and TF6, SF6 and TF1, SF6 and TF6, SF7 and TF6, SF9 and TF6, and SF10 and TF6. Fig. 3 illustrates the correspondent relationships between the scientific fields and the technological fields according to the similarity measurements in Table 3. The lower ten circles that denote the ten scientific fields in Table 1 are located in the chart according to their published mean year. Similarly, the upper six circles represent the six technological fields in Table 2 and are located in the chart and aligned with their mean year of patent granted. The circle size presents its count of corresponding scientific field or technological field.

Table 3
Similarity between scientific field and technological field.

	TF1	TF2	TF3	TF4	TF5	TF6
SF1	0.579*	0.225	0.295	0.229	0.662*	0.533
SF2	0.541	0.197	0.297	0.246	0.626*	0.585*
SF3	0.595*	0.191	0.308	0.228	0.480	0.554
SF4	0.596*	0.237	0.349	0.284	0.538	0.623*
SF5	0.579*	0.594*	0.349	0.229	0.422	0.559*
SF6	0.589*	0.280	0.351	0.265	0.469	0.608*
SF7	0.524	0.289	0.327	0.308	0.404	0.665*
SF8	0.444	0.213	0.343	0.227	0.327	0.506
SF9	0.508	0.179	0.331	0.321	0.316	0.694*
SF10	0.490	0.297	0.278	0.266	0.307	0.605*

Note:.
* denotes the relationship between science and technology.

Shibata et al. (2010) suggest that less similar fields suggest the potential opportunities for scientific advancements or technological development. In this case, SF8 has less similarity with the technological fields. TF4 and TF5 are less similar with the scientific fields. Thus, SF8 might imply a potential opportunity for technological innovation, and TF4 and TF5 might be considered two potential opportunities for scientific advancements. To ensure that the less similar scientific field and technological fields are the opportunities, consulting experts who are familiar with the domain knowledge of smart health monitoring is necessary.

4. Discussions

Fig. 3 shows that many scientific fields are related to more than one technological field and many technological fields connect to more than one scientific field. This phenomenon may reveal the nature of cross-boundary knowledge creation process and knowledge diffusion toward transdisciplinary and heterogeneous technoscientific areas with applied prospects, such as nanotechnology and biology in silico (Verbeek et al., 2002). Past studies assume that a technology is directly descended from science in the form of a paper (Carpenterand Narin, 1983; Narin and Oilvastro, 1992, 1998). Researchers emphasize the interaction or overlap relationship between science and technology because of the increasing network-embedded structure of knowledge generation constituted by different actors, such as universities, institutes, and companies (Gibbon et al., 1994; Meyer, 2000; Verbeek et al., 2002). In addition to the applied technology produced from multidisciplinary fundamental research, the demand side from industry stimulates continuous exploration and the study of scientific knowledge (Klevorick et al., 1995; Meyer, 2002; Glänzel and Meyer, 2003). The remainder of this section discusses the practical implications based on the relationships between science and technology discussed in the previous section and attempts to validate the results with the literature, knowledge of the two external experts, and the reports of institutions that are dedicated to this field.

4.1. TF1 connected with SF1, SF3, and SF5

As shown in Table 3 and Fig. 3, TF1 (health information infrastructure in hospitals) has strong semantic similarity with SF1 (healthcare applications that focus on psychological signal measurement), SF3 (sensors for wearable health or medical devices), SF4 (infrastructure of health information networks), SF5 (medical imaging platforms), and SF6 (health information processing technology). TF1 addresses the emerging hospital information infrastructure to achieve smart health monitoring, such as psychological signal monitoring and collection, information encryption, drug administration, wearable devices, electronic medical record, etc. SF1 focuses on the measurement of psychological signals, including electrocardiogram (ECG), electroencephalography (EEG), electromyography (EMG), and sensor material. SF3 concentrates on the development of sensors for psychological signals. SF4 contains the research topics regarding data collection, data storage, software, and information system architectures with respect to the information infrastructure of health monitoring. SF5 is dedicated to medical image transmission, compression, and decompression. Owing to the big data nature of health information, SF6 focuses on information processing and data safety. This result is supported by the ACC's report (Bhavnani et al., 2017) and the work of Liu et al. (2016). The former research indicates that healthcare innovation in hospitals include big

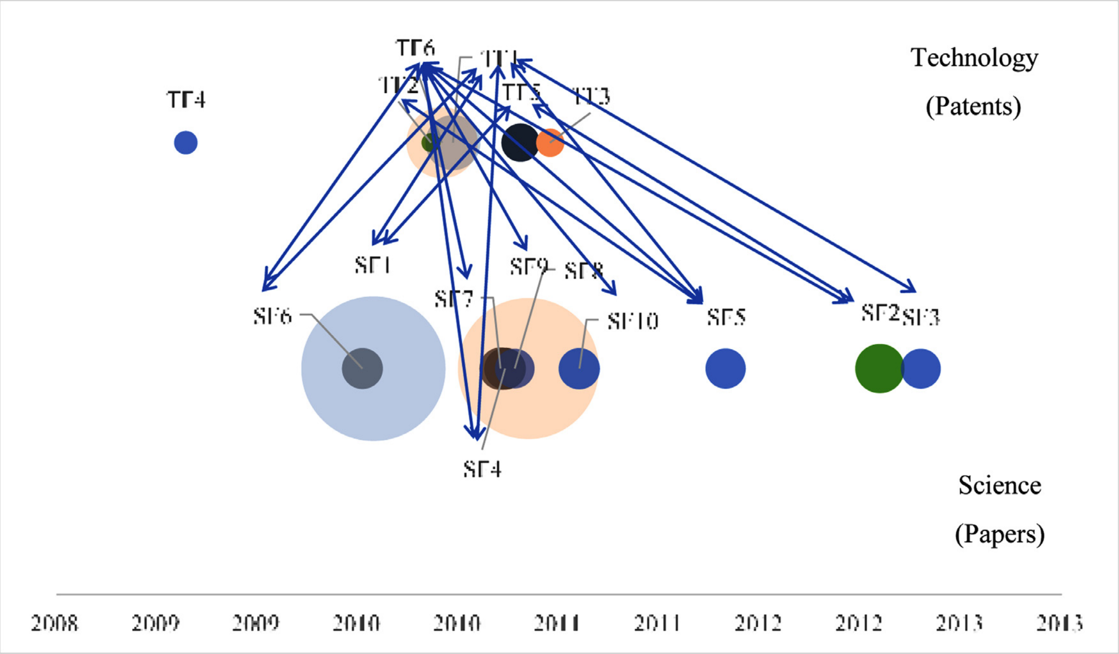

Fig. 3. Linkage between scientific fields and technological fields.

data, sensors, and biophysical measurements, and the latter research reports that imaging technology is one of the keys for smart health monitoring.

4.2. TF2 connected with SF5

TF2 (medical imaging management systems) and SF5 share high semantic relatedness, as presented in Table 3. Many studies have demonstrated the potential of imaging technology for smart health monitoring (Aldaz et al., 2015; Pan et al., 2015; Liu et al., 2016; Bhavnani et al., 2017). This result shows that both academic researchers and practitioners are dedicated to this area, including basic research and applied research. Thus, this outcome echoes prior studies.

4.3. TF5 connected with SF1 and SF2

TF5 (data communication systems) has high similarity with SF1 and SF2 (security and privacy of healthcare systems), as shown in Table 3. Leveraging nearly ubiquitous internet connectivity, information and data communication technologies covered in TF5 can provide real-time data transfer of psychological signal measurement supported by SF1 (Bhavnani et al., 2017). Moreover, privacy and security during data transfer or data sharing are considered one of the critical issues in the basic research regarding smart health monitoring (Chan et al., 2012; Baig et al., 2013; Liu et al., 2016; Bhavnani et al., 2017).

4.4. TF6 connected with SF2, SF4, SF5, SF6, SF7, SF9, and SF10

TF6 (decision support systems for diagnosis) connects SF2, SF4, SF5, SF6, SF7 (diagnosis, treatment or therapy through telemedicine), SF9 (telehealth for chronic diseases), and SF10 (patients' psychological signal measurement), as presented in Table 3. The theme of SF7 involves delivering a diagnosis, treatment or therapy via telemedicine technology. SF9 focuses on exploring chronic disease management for patients diagnosed with dementia, Parkinson's disease, and Alzheimer's disease. SF10 involves psychological signal measurements, especially for patients. Smart health monitoring systems integrate multi-parameter physiological sensors, measurement systems, digital devices, and services toward real-time decision support for disease prevention, symptom detection, and diagnosis (Chan et al., 2012; Bhavnani et al., 2017). To achieve decision support for disease diagnosis, therefore, smart health monitoring systems involve SF2 for patients' privacy and security (Chan et al., 2012; Baig et al., 2013; Liu et al., 2016; Bhavnani et al., 2017), SF4 for data collection, data storage, software, and information system architectures (Liu et al., 2016; Bhavnani et al., 2017), SF5 for medical imaging (Liu et al., 2016; Bhavnani et al., 2017), SF6 for medical information processing and data safety (Liu et al., 2016; Bhavnani et al., 2017), SF7 for telediagnosis and teletherapy (Chan et al., 2012), SF9 for chronic disease telehealth (Chan et al., 2012; Baig et al., 2013; Bhavnani et al., 2017), and SF10 for measuring patients' psychological signals (Chan et al., 2012; Baig et al., 2013; Liu et al., 2016).

4.5. Potential opportunity of technological innovation

According to Table 3 and Fig. 3, the potential opportunity of technological innovation might exist in SF8 (management of smart health monitoring) owing to its less similarity with technological fields. This scientific field tends to involve the evaluation of smart health monitoring systems and usability studies. The performance and users' perception of smart health monitoring systems are critical for patients or elderly users (Chan et al., 2012; Liu et al., 2016). Despite the flourishing development of smart health monitoring applications, as presented in the related patents, the future product, device, or service design should reflect the voice of the customer to facilitate the adoption of smart health monitoring.

4.6. Potential opportunities of scientific advancement

In addition, TF3 (home health care systems) and TF4 (health monitoring systems) might imply the potential opportunities of scientific advancement due to their less semantic similarity with the scientific fields, as shown in Table 3 and Fig. 3. TF3 involves the software, components, systems, and platforms for home use. Liu et al. (2016) indicate that additional research is needed to address the conditions of disability prediction, chronic obstructive pulmonary disease (COPD), health related quality of life, and fail prevention at home. Therefore, TF3 could be considered a potential opportunity of scientific advancement.

The patents in TF4 contain health monitoring apparatuses, network systems for remote health monitoring, medical record systems, and diagnosis and treatment systems that are related to health monitoring at the system level. TF4 seems to imply a potential opportunity for technological innovation because of its weak similarity with scientific fields. However, the experts suggest that TF4 should not be regarded as a potential opportunity of scientific advancement, considering that its mean granted year is 2008.6, which is earlier than the mean published year of all of ten scientific fields. Thus, the related fundamental theories should appear in the larger number of later academic publications. The patent applicants tend to seek vagueness in their patent claims (Arinas, 2012), whereas academic language is designed to be concise, precise, and authoritative (Snow, 2010). This speculation by the experts explains why TF4 is less similar to all of the scientific fields.

5. Conclusion and limitations

This study integrates text mining, ORCLUS, and the cosine similarity of the tf-idf vector to determine the advantages of these methods, which are retaining information and reducing the experts' subjectiveness for extracting potential opportunities for scientific advancement and technological innovation from the comparison between scientific publications and patents. Smart health monitoring technology is adopted to verify the process proposed by this study. For the administrators in firms or research institutions, the methods and process applied in this study can be employed to quickly and thoroughly detect the potential scientific or technological prospects or build a technology roadmap. In terms of smart health monitoring, SF8 (management of smart health monitoring) and TF3 (home health care systems) are considered potential opportunities of technological innovation and scientific advancement, respectively.

This study has some limitations that suggest avenues for future research. Two limitations are inherent in ORCLUS for clustering and the cosine similarity of tf-idf for similarity detection. First, the number of scientific and technological fields clustered by ORCLUS may not reflect the real world even if the results are validated with two experts and the literature. The validity of the clustering results is severely affected by the preciseness of document wordings (Wang et al., 2015). Especially, the vagueness nature of patent wording (Arinas, 2012) increases the challenge to the text-based ORCLUS approach and then affects the clustering. Second, the effective linkages between the scientific fields and the technological fields rely on a threshold value that is determined by expert assessment, while the cosine similarity of tf-idf measures the similarity between science and technology. However, the expert judgment might be deficient because their experience may not address all of the research themes in smart health monitoring technology. Future research should experiment with other promising approaches, especially the blooming natural language processing (NLP) methods in recent years, such as word2vec, global vectors (GloVe), bidirectional encoder representations from transformers (BERT), embeddings from language models (ELMo), enhanced representation from knowledge integration (ERNIE), etc., to better reflect the clusters of science and technology and their relationship.

Third, this study only collects SCI journal papers as science and patents from the USPTO database as technology. In many areas, conference proceedings are important for containing relevant advanced science. Future research could include conference proceedings to uncover the potential opportunity for technological progress. Regarding technological documents, the patents retrieved from the USPTO database may only present the application development in the US market. Future work could collect patents from other databases, such as the Espacenet of European Patent Office (EPO) or Patentscope of World Intellectual Property Organization (WIPO), to learn the overall picture of the technological development of smart health monitoring technology.

The whole process presented in this study still needs assistance from experts to select keywords for searching and collecting documents, unify the synonyms, establish a threshold value for screening important linkages between science and technology, and determine the potential opportunities for scientific advancement and technological innovation. The results are affected by the experts' domain knowledge and subjectiveness. Future research could attempt to construct a process or approach based on NLP to automatically identify potential technological opportunities without experts' supports.

CRediT authorship contribution statement

Yung-Chi Shen: Conceptualization, Formal analysis, Investigation, Writing - original draft. **Ming-Yeu Wang:** Methodology, Writing - review & editing. **Ya-Chu Yang:** Data curation.

Acknowledgement

This research is supported by the Ministry of Science and Technology Taiwan, under grant MOST 105–2410-H-415–038. The insightful comments from reviewers are gratefully acknowledged.

References

Aggarwal, C.C., Yu, P.S., 2000. Finding generalized projected clusters in high dimensional spaces. ACM SIGMOD Record 29 (2), 70–81.

Aggarwal, C.C., Wolf, J.L., Yu, P.S., Procopiuc, C., Park, J.S., 1999. Fast algorithms for projected clustering. ACM SIGMOD Record 28 (2), 61–72.

Aggarwal, R., Gehrke, J., Gunopulos, D., Raghavan, P., 1998. Automatic subspace clustering of high dimensional data for data mining applications. ACM SIGMOD Record 27 (2), 94–105.

Aldaz, G., Shluzas, L.A., Pickham, D., Eris, O., Sadler, J., Joshi, S., Leifer, L., 2015. Hands-free image capture, data tagging and transfer using Google Glass: a pilot study for improved wound care management. PLoS ONE 10 (4).

Archibugi, D., Pianta, M., 1996. Measuring technological change through patents and innovation survey. Technovation 16 (9), 451–468.

Arians, I., 2012. How vague can your patent be? Vagueness strategies in U.S. patents. HERMES – Journal of Language and Communication in Business 48, 55–74.

Baig, M.M., Gholamhosseini, H., 2013. Smart health monitoring systems: an overview of design and modeling. J Med Syst 37 (2), 1–14.

Beyer, K., Goldstein, J., Ramakrishnan, R., Shaft, U., 1999. When is "nearest neighbor" meaningful? In: Database Theory—ICDT'99: 7th International Conference Jerusalem, Israel, January 10–12, 1999 Proceedings, pp. 217–235.

Bhavnani, S.P., Parakh, K., Atreja, A., Druz, R., Graham, G.N., Hayek, S.S., Krumholz, H.M., Maddox, T.M., Majmudar, M.D., Rumsfeld, J.S., Shah, B.R., 2017. 2017 roadmap for innovation-ACC health policy statement on healthcare transformation in the era of digital health, big data, and precision health: a report of the American College of Cardiology task force on health policy statements and systems of care. J. Am. Coll. Cardiol. 70 (21), 2696–2718.

Bush, V., 1945. Science, The Endless Frontier. Public Affairs Press, Washington, DC.

Carpenter, M.P., Narin, F., 1983. Validation study: patent citations as indicators of science and foreign dependence. World Patent Information 5 (3), 180–185.

Chan, M., Estève, D., Fourniols, J.-.Y., Escriba, C., Campo, E., 2012. Smart wearable systems: current status and future challenges. Artif Intell Med 56 (3), 137–156.

Chang, C.K., Oyama, K., 2018. Guest editorial: a roadmap for mobile and cloud services for digital health. IEEE Transactions on Services Computing 11 (2), 232–235.

Chang, S.B., Lai, K.K., Chang, S.M., 2009. Exploring technology diffusion and classification of business methods: using the patent citation network. Technol Forecast Soc Change 76 (1), 107–117.

Chang, Y.W., Yang, H.W., Huang, M.H., 2017. Interaction between science and technology in the field of fuel cells based on patent paper analysis. Electronic Library 35 (1), 152–166.

Chaves, C.V., Moro, S., 2007. Investigating the interaction and mutual dependence between science and technology. Res Policy 36 (8), 1204–1220.

Choi, S., Kim, Ho., Yoon, J., Kim, K., Lee, J.Y., 2013. An SAO-based text-mining approach for technology roadmapping using patent information. R&D Management 43 (1), 52–74.

Devlin, J., Chang, M.W., Lee, K. & Toutanova, K. (2018). BERT: pre-training of deep bidirectional transformers for language understanding. Retrieved fromhttps://arxiv.org/abs/1810.04805.

Érdi, P., Makovi, K., Somogyvári, Z., Strandburg, K., Tobochnik, J., Volf, P., Zalányi, L., 2013. Prediction of emerging technologies based on analysis of the US patent citation network. Scientometrics 95 (1), 225–242.

Etzkowitz, H., 1998. The norms of entrepreneurial science: cognitive effects of the new university-industry linkages. Res Policy 27 (8), 823–833.

Fehder, D.C., Murray, F., Stern, S., 2014. Intellectual property rights and the evolution of scientific journals as knowledge platforms. International Journal of Industrial Organization 36 (SI), 83–94.

Feldman, R., Sanger, J., 2007. The Text Mining handbook: Advanced Approaches in Analyzing Unstructured Data. Cambridge University Press, Cambridge.

Fujita, K., Kajikawa, Y., Mori, J., Sakata, I., 2014. Detecting research fronts using different types of weighted citation networks. Journal of Engineering and Technology Management 32, 415–423.

Gao, J.P., Ding, K., Teng, L., Pang, J., 2012a. Hybrid documents co-citation analysis: making sense of the interaction between science and technology in technology diffusion. Scientometrics 93 (2), 459–471.

Gibbons, M., Limoges, C., Nowotny, H., Schwartzman, S., Scott, P., Trow, M., 1994. The New Production of Knowledge: The Dynamics of Science and Research in Contemporary Societies. SAGE Publications, London.

Glänzel, W., Meyer, M., 2003. Patents cited in the scientific literature: an exploratory study of 'reverse' citation relations. Scientometrics 58 (2), 415–428.

Hellmann, T., 2007. The role of patents for bridging the science to market gap. J Econ Behav Organ 63 (4), 624–647.

Ho, J.C., Saw, E.C., Lu, L.Y.Y., Liu, J.S., 2014. Technological barriers and research trends in fuel cell technologies: a citation network analysis. Technol Forecast Soc Change 82, 66–79.

Hsu, F.C., Trappey, A.J.C., Trappey, C.V., Hou, J.L., Liu, S.J., 2006. Technology and knowledge document cluster analysis for enterprise R&D strategic planning. International Journal of Technology Management 36 (4), 336–353.

Huang, M.H., Chang, C.P., 2014. Detecting research fronts in OLED field using bibliographic coupling with sliding window. Scientometrics 98 (3), 1721–1744.

Huang, M.H., Chang, C.P., 2015. A comparative study on detecting research fronts in the organic light-emitting diode (OLED) field using bibliographic coupling and co-citation. Scientometrics 102 (3), 2041–2057.

Huang, M.H., Chiang, L.Y., Chen, D.Z., 2003. Constructing a patent citation map using bibliographic coupling: a study of Taiwan's high-tech companies. Scientometrics 58 (3), 489–506.

Kajikawa, Y., Takeda, Y., 2008. Structure of research on biomass and bio-fuels: a citation-based approach. Technol Forecast Soc Change 75 (9), 1349–1359.

Kajikawa, Y., Yoshikawa, J., Takeda, Y., Matsushima, K., 2008. Tracking emerging technologies in energy research: toward a roadmap for sustainable energy. Technol Forecast Soc Change 75 (6), 771–782.

Karki, M., 1997. Patent citation analysis: a policy analysis tool. World Patent Information 19 (4), 11–20.

Kim, H.K., Kim, H., Cho, S., 2017. Bag-of-concepts: comprehending document representation through clustering words in distributed representation. Neurocomputing 266, 336–352.

Klevorick, A.K., Levin, R.C., Nelson, R.R., Winter, S.G., 1995. On the sources and significance of interindustry differences in technological opportunities. Res Policy 24 (2), 185–205.

Kostoff, R.N., Schaller, R.R., 2001. Science and technology roadmaps. IEEE Transactions on Engineering Management 48 (2), 132–143.

Kostoff, R.N., 1998. The use and misuse of citation analysis in research evaluation. Scientometrics 43 (1), 27–43.

Kostoff, R.N., 2008. Literature-related discovery (LRD): introduction and background. Technol Forecast Soc Change 75 (2), 165–185.

Kriegel, H.-.P., Kröger, P., Zimek, A., 2009. Clustering high-dimensional data: a survey on subspace clustering, pattern-based clustering, and correlation clustering. ACM Trans Knowl Discov Data 3 (1), 1–58.

Kuusi, O., Meyer, M., 2007. Anticipating technological breakthroughs: using bibliographic coupling to explore the nanotubes paradigm. Scientometrics 70 (3), 759–777.

Lai, K.K., Wu, S.J., 2005. Using the patent co-citation approach to establish a new patent classification system. Inf Process Manag 41 (2), 313–330.

Ledydesdorff, L., Cozzens, S., Van den Besselaar, P., 1994. Tracking areas of strategic importance using scientometric mappings. Res Policy 23 (2), 217–229.

Lee, C., Kang, B., Shin, J., 2015. Novelty-focused patent mapping for technology opportunity analysis. Technol Forecast Soc Change 90, 355–365.

Leydesdorff, L., Etzkowitz, H., 1996. Emergence of a triple helix of university-industry-government relations. Science and Public Policy 23 (5), 279–286.

Liu, L., Stroulia, E., Nikolaidis, I., Miguel-Cruz, A., Rios Rincon, A., 2016. Smart homes and home health monitoring technologies for older adults: a systematic review. Int J Med Inform 91, 44–59.

Ma, T., Porter, A.L., Guo, Y., Ready, J., Xu, C., Gao, L., 2014. A technology opportunities analysis model: applied to dye-sensitised solar cells for China. Technology Analysis & Strategic Management 26 (1), 87–104.

Mansfield, E., 1991. Academic research and industrial innovation. Res Policy 20 (1), 1–12.

Martino, J.P., 2003. A review of selected recent advances in technological forecasting. Technol Forecast Soc Change 70 (8), 719–733.

McNamara, P., Baden-Fuller, C., 1999. Lessons from the Celltech case: balancing knowledge exploration and exploitation in organization renewal. British Journal of Management 10 (4), 291–307.

Merton, R.K., 1973. The Sociology of Science: Theoretical and Empirical Investigations. University of Chicago Press, Chicago.

Meyer, M., 2000. Does science push technology? Patents citing scientific literature. Res Policy 29 (3), 409–434.

Meyer, M., 2002. Tracing knowledge flows in innovation systems—An informetric perspective on future research science-based innovation. Economic Systems Research 14 (4), 323–344.

Mikolov, T., Chen, K., Corrado, G. & Dean, J. (2013). Efficient estimation of word representations in vector space. Retrieved from https://arxiv.org/abs/1301.3781.

Müller, E., Günnemann, S., Assent, I., Seidl, T., 2009. Evaluating clustering in subspace projections of high dimensional data. Proceedings of the VLDB Endowment 2 (1), 1270–1281.

Murray, F., Stern, S., 2007. Do formal intellectual property rights hinder the free flow of scientific knowledge? An empirical test of the anti-commons hypothesis. J Econ Behav Organ 63 (4), 648–687.

Murray, F., 2002. Innovation as co-evolution of scientific and technological networks: exploring tissue engineering. Res Policy 31 (8–9), 1389–1403.

Narin, F., Oilvastro, D., 1992. Status report: linkage between technology and science. Res Policy 21, 237–249.

Narin, F., Oilvastro, D., 1998. Linkage between patents and papers: an interim EPO/US comparison. Scientometrics 41 (1–2) 51-49.

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

Nelson, R.R., 1982. The role of knowledge in R&D efficiency. Q J Econ 97 (3), 453–470.

Ogawa, T., Kajikawa, Y., 2015. Assessing the industrial opportunity of academic research with patent relatedness: a case study on polymer electrolyte fuel cells. Technol Forecast Soc Change 469–475.

Olsson, O., 2005. Technological opportunity and growth. Journal of Economic Growth 10 (1), 35–57.

Pan, C.Y., Huang, C.S., Horng, G.J., Peng, P.L., Jong, G.J., 2015. Infrared image processing for a physiological information telemetry system. Wireless Personal Communications 83 (4), 3181–3208.

Parsons, L., Haque, E., Liu, H., 2004. Subspace clustering for high dimensional data: a review. ACM SIGKDD Explorations Newsletter 6 (1), 90–105.

Pavitt, K., 1991. What makes basic research economically useful? Res Policy 20 (2), 109–119.

Pennington, J., Socher, R., Manning, C., 2014. Glove: global vectors for word representation. In: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), pp. 1532–1543.

Petrescu, A.S., 2009. Science and technology for economic growth. New insights from when the data contradicts desktop models. Review of Policy Research 26 (6), 839–880.

Price, D.J.D., 1965. Is technology historically independent of science? A study in statistical historiography. Technol Cult 6 (4), 553–568.

Rosenberg, N., 1982. How exogenous is science? In: Rosenberg, N. (Ed.), Inside the Black Box: Technology and Economics. Cambridge University, Cambridge, pp. 141–160.

Rosenberg, N., 1990. Why do firms do basic research (with their money)? Res Policy 19 (2), 165–174.

Salton, G., Wong, A., Yang, C.S., 1975. A vector space model for automatic indexing. Commun ACM 18 (11), 613–620.

Schmoch, U., Thielmann, A., 2012. Cyclical long-term development of complex technologies—Premature expectations in nanotechnology? Res Eval 21 (2), 126–135.

Schmoch, U., 2007. Double-boom cycles and comeback of science-push and market-pull. Res Policy 36 (7), 1000–1015.

Shibata, N., Kajikawa, A., Sakata, I., 2011c. Measuring relatedness between communities in a citation network. Journal of the American Society for Information Science and Technology 62 (7), 1360–1369.

Shibata, N., Kajikawa, Y., Sakata, I., 2010. Extracting the commercialization gap between science and technology—Case study of a solar cell. Technol Forecast Soc Change 77 (7), 1147–1155.

Shibata, N., Kajikawa, Y., Sakata, I., 2011b. Detecting potential technological fronts by comparing scientific papers and patents. foresight 13 (5), 51–60.

Shibata, N., Kajikawa, Y., Takeda, Y., Matsushima, K., 2008. Detecting emerging research fronts based on topological measures in citation networks of scientific publications. Technovation 28 (11), 758–775.

Shibata, N., Kajikawa, Y., Takeda, Y., Sakata, I., Matsushima, K., 2011a. Detecting emerging research fronts in regenerative medicine by the citation network analysis of scientific publications. Technol Forecast Soc Change 78 (2), 274–282.

Small, H., 2006. Tracking and predicting growth areas in science. Scientometrics 68 (3), 595–610.

Snow, C.E., 2010. Academic language and the challenge of reading for learning about science. Science 328 (5977), 450–452.

Szepannek, G. (2013). Package 'orclus'. Retrieved from https://cran.r-project.org/web/packages/orclus/orclus.pdf.

Thompson, N.C., Ziedonis, A.A., Mowery, D.C., 2018. University licensing and the flow of scientific knowledge. Res Policy 47 (6), 1060–1069.

Tseng, Y.H., Lin, C.J., Lin, Y.I., 2007. Text mining techniques for patent analysis. Inf Process Manag 43 (5), 1216–1247.

U.S. Food & Drug Administration (2019). Digital health. Retrieved from https://www.fda.gov/medical-devices/digital-health.

Verbeek, A., Debackere, K., Luwel, M., Andries, P., Zimmermann, E., Deleus, F., 2002. Linking science to technology: using bibliographic references in patents to build linkage schemes. Scientometrics 54 (3), 399–420.

Wang, M.Y., Chang, D.S., Kao, C.H., 2010. Identifying technology trends for R&D planning using TRIZ and text mining. R&D Management 40 (5), 491–509.

Wang, M.Y., Fang, S.C., Chang, Y.H., 2015. Exploring technological opportunities by mining the gaps between science and technology: microalgal biofuels. Technol Forecast Soc Change 92, 182–195.

Watts, R.J., Porter, A.L., 1997. Innovation forecasting. Technol Forecast Soc Change 56 (1), 25–47.

Weiss, S.M., Indurkhya, N., Zhang, T., Damerau, F.J., 2005. Text Mining: Predictive Methods For Analyzing Unstructured Information. Springer-Verlag, New York.

Williams, J., 2004. Wireless in Healthcare: A Study Tracking the Use of RFID, Wireless Sensor Solutions, and Telemetry Technologies By Medical Device Manufacturers and Healthcare Providers. The FocalPoint Group, USA.

Woo, K.-.G., Lee, J.-.H., Kim, M.-.H., Lee, Y.-.J., 2004. FINDIT: a fast and intelligent subspace clustering algorithm using dimension voting. Inf Softw Technol 46 (4), 255–271.

World Health Organization, 2018. Ehealth At WHO. Retrieved from. https://www.who.

int/ehealth/about/en/.

Yip, K.Y., Cheung, D.W., Ng, M.K., 2005. On discovery of extremely low-dimensional clusters using semi-supervised projected clustering. In: ICDE '05 Proceedings of the 21st International Conference on Data Engineering, pp. 329–340.

Yoon, B., Park, Y., 2005. A systematic approach for identifying technology opportunities: keyword-based morphology analysis. Technol Forecast Soc Change 72 (2), 145–160.

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

Yoon, B., Park, I., Coh, B., 2014. Exploring technological opportunities by linking technology and products: application of morphology analysis and text mining. Technol Forecast Soc Change 86, 287–303.

Yoon, J., Kim, K., 2012. Detecting signals of new technological opportunities using semantic patent analysis and outlier detection. Scientometrics 90 (2), 445–461.

You, H., Li, M., Hipel, K.W., Jiang, J., Ge, B., Duan, H., 2017. Development trend forecasting for coherent light generator technology based on patent citation network analysis. Scientometrics 111 (1), 297–315.

Yung-Chi Shen is an associate professor of the Department of BioBusiness Management, National Chiayi University, Taiwan. He received his master degree from the Graduate Institute of Industrial Education and Technology, National Changhua University of Education and the Ph.D degree in management of technology from National Chiao Tung University in Taiwan. His-works have appeared in Technological Forecasting & Social Change, Entrepreneurship & Regional Development, Computers in Industry, and Energy Policy. His-current research focuses on university technology transfer, new product development, and technology assessment.

Ming-Yeu Wang is a professor of the Department of BioBusiness Management, National Chiayi University, Taiwan. She received her Master degree in management of technology and the Ph.D. degree in industrial engineering and management from the National Chiao Tung University, Taiwan. Her works have appeared in Technological Forecasting & Social Change, R & D Management, International Journal of Technology Management and Journal of Engineering and Technology Management. Her current research focuses on technological forecasting, patent analysis and R&D management.

Ya-Chu Yang is a master student of the Institute of BioBusiness Management, National Chiayi University, Taiwan. Her research interests include patent analysis, text mining, and health care sector.