Technological Forecasting & Social Change 180 (2022) 121718

Contents lists available at ScienceDirect

Technological Forecasting & Social Change

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

Technology Opportunity Discovery using Deep Learning-based Text Mining and a Knowledge Graph

MyoungHoon Lee [a,1], Suhyeon Kim [a,1], Hangyeol Kim [b], Junghye Lee [c,*]

[a] Department of Industrial Engineering, Ulsan National Institute of Science and Technology (UNIST), Ulsan 44919, Republic of Korea
[b] Department of Business Analytics, Graduate School of Interdisciplinary Management, UNIST, Ulsan 44919, Republic of Korea
[c] Department of Industrial Engineering and Artificial Intelligence Graduate School, UNIST, Ulsan 44919, Republic of Korea

ARTICLE INFO

Keywords:
Technology opportunity discovery
Text mining
Doc2vec
Knowledge graph
Logistics regression
Index

ABSTRACT

To capture emerging technologies in the fast-changing technology market, use of information concerning new technology-based firms (NTBFs) is strongly encouraged, in addition to the information about the technology itself. Especially, NTBFs rapidly respond to technological change, and their investment information is a significant criterion of technology valuation. Therefore, this study proposes a new technology opportunity discovery (TOD) framework that exploits text mining by deep learning and a knowledge graph (KG) by using three data sources: technology, NTBF, and investor data. First, a technology-classification model was developed using technical text data acquired using Doc2vec and logistic regression, and then this model assigned highly-relevant technology fields to NTBFs using NTBFs' investor relation text data. Next, a KG that considers technology, NTBF, and NTBF's investor was constructed to represent their relations for TOD by using the results of previous steps. Lastly, considering inter-connectivities of such factors, a TOD index that measures the potential of technologies was proposed. The accuracy and validity of the methods were demonstrated empirically, and an evaluation of emerging technologies identified by the analysis was provided. Our framework will be of great significance as a useful alternative to provide new insights for emerging technologies in the industry and market.

1. Introduction

Recently, new technology-based firms (NTBFs, i.e., technology-based startups) have been founded at an accelerating rate worldwide, and some world-leading companies and business angels have been actively investing in the technologies of NTBFs or carrying out mergers and acquisitions for the NTBFs (Moghaddam et al., 2015). These circumstances have formed a new technological innovation ecosystem specialized in NTBFs (i.e., the NTBF ecosystem), where an ability to preoccupy promising technologies strongly affects the quantity and flow of money in the venture investment market (Colombo et al., 2016; Tegarden et al., 2012). To be specific, when new technologies appear in this ecosystem, the NTBFs attempt to exploit and commercialize the latest promising technologies; they can generate profit by providing them to the customers, attract investment, and achieve sustainable growth (Saura et al., 2019; Yoon and Park, 2004).

Exploitation of technology opportunities is vital in an environment in which foundation, technology-driven growth, and investment related to NTBFs are central components. Therefore, the value assessment of technologies is growing in importance for various stakeholders, ranging from R&D managers to investment experts, to discovery technology fields with high growth potential in the actual market. Nevertheless, quantitative and objective measurement of market growth, sustainability, and investment potential of technology is a difficult task in a real NTBF ecosystem (Dushnitsky and Lenox, 2006; Kang et al., 2021). This difficulty arises because the investment and growth for NTBFs are incremental and phased (e.g., seed-series A-series B-unicorn); for this reason, the long-term continuity of technology growth needs to be considered for the valuation of NTBFs-related technologies.

To respond to this demand, technology opportunity discovery (TOD) is necessary, which is a way to identify emerging technology opportunities and quickly react to them (Klevorick et al., 1995; Nieto and Quevedo, 2005; Olsson, 2005). The main perspective of TOD is to assess the potential value of technology and to discover promising technologies that have great future value (i.e., emerging technologies). However, existing TOD studies have only focused on evaluating the technology by

* Corresponding author.
E-mail addresses: lmhoon012@unist.ac.kr (M. Lee), suhyeonkim@unist.ac.kr (S. Kim), hangyeol0225@unist.ac.kr (H. Kim), junghyelee@unist.ac.kr (J. Lee).
[1] These authors equally contributed

https://doi.org/10.1016/j.techfore.2022.121718
Received 16 September 2021; Received in revised form 18 February 2022; Accepted 21 April 2022
Available online 29 April 2022

quantifying the increase in the amount of bibliography of technical data (Bengisu and Nekhili, 2006; Curran and Leker, 2011; Xin et al., 2010; Yoon et al., 2014; Yoon and Park, 2005), or on searching the company's portfolio to find a technology opportunity that is suitable for the company (Cho et al., 2016; Choi et al., 2019; Lee et al., 2010). To the best of our knowledge, no TOD work has considered the growth potential of technology in the NTBF-investment perspective and how the technology is used and valued in new industries and markets by suppliers and consumers.

Therefore, in an advance over existing TOD analysis studies, this study considers the trends in NTBFs and their technological innovation ecosystem to represent the marketability in real businesses. This perspective has never been explored before, despite the important contributions on NFBFs to advances in technology. We present a framework for TOD analysis by utilizing information of NTBFs and their investment information together with technical data to represent the NTBF-related technology ecosystem and its cycle structure. To do so, we propose a new approach that fits the data by using deep learning-based text mining and a knowledge graph (KG). The framework allows people to identify technology opportunities that are of good quality. Specifically, we first develop a technology-classification model (Doc2vec-LR) that makes use of Doc2vec, a contextual document-embedding algorithm along with logistic regression (LR). Then we apply Doc2vec-LR to the textual information of NTBFs to automatically identify the technology fields of each NTBF. Third, we generate a novel KG for the TOD (TOD-KG), that contains technology, NTBF, and investing information of the NTBF. Finally, we create a new TOD index that can quantitatively evaluate technology by applying social network analysis measures on the TOD-KG from various perspectives, then use the calculated TOD index to identify the emerging technologies. We conduct a detailed TOD analysis using the proposed framework for NTBFs, and provide meaningful interpretation and evidence of TOD analysis. Our framework is the first attempted TOD analysis to represent the technology ecosystem in which NTBFs and investors are linked together.

Our Contributions. This study has made four contributions:

- This study proposed a new TOD analysis framework using NTBFs and their investment data along with technical documents and used deep learning-based text mining and KG to develop methodologies suitable for the data. Our framework can identify technology opportunities that have sustainable growth potential in the NTBF ecosystem.
- This study used Doc2vec to generate the document representation for TOD analysis to better structure the document stream by extracting document-specific embedding, including the contextual characteristics of words. Further, we presented a method to represent the hierarchical aspects of categorization levels of technical documents on Doc2vec training; the method enables embedding of technology to competently represent inherent patterns of entire technologies.
- Our TOD-KG was created by representing technology, NTBF, and investor. For the TOD-KG, the connection between nodes is composed of the similarity of text embedding between technologies or NTBFs as well as the link elements of direct technology classification and investing behaviors between technology and NTBF or between NTBF and investor.
- We proposed the new TOD index derived from the TOD-KG to quantify emerging technology opportunities. We introduced a method to calculate a TOD index by appropriately combining the representative centrality measures by considering three types of factors in the TOD-KG. The TOD index can identify the major technology that is the main crossroad in the TOD-KG and is highly linked to other entities.

This paper is organized as follows. Section 2 reviews existing TOD studies, and Section 3 describes the data used in this study. Section 4 details the proposed TOD framework. Section 5 demonstrates the results and validity of the methodology. Section 6 presents a detailed analysis of the results, and then Section 7 concludes this study.

2. Related work

In this section, we review related work on TOD analysis from technology and corporate perspectives, respectively. The former review considers studies that identify emerging technologies in specific fields by adopting TOD techniques or by developing such methods by considering technical data. The latter considers studies that focus on the target corporations by analyzing their technical documents, which aim to provide adequate R&D directions and TOD strategies on promising technologies for them, and to discover the potential of their core technologies.

2.1. TOD in a technology perspective

2.1.1. Text Mining Methods for the TOD

The initial TOD studies were conducted by experts by monitoring and bibliographic analysis of public databases (Porter and Detampel, 1995). The amount of data has increased rapidly and continuously, so TOD studies have moved towards automated data processing to supplement expert-oriented techniques (Lee et al., 2020; Yoon et al., 2015). Patents and scientific articles summarize R&D results and are rich sources of technology information, mainly in the form of text, so TOD studies that have used such data have attempted to identify technology opportunities by utilizing keyword-search methods such as a keyword morphological matrix and a keyword matching algorithm (Xin et al., 2010; Yoon and Park, 2004). These methods can derive potential technological opportunities by identifying semantic similarity between technologies, and create possible technology combinations by using technical keywords. Still, these methods consider only the frequency of co-words extracted from technical documents, and cannot discover their contextual and latent semantic information.

To overcome the limitations of keyword-based methods, TOD studies have used machine learning-based text mining methods. These methods can examine a large database to extract latent information related to TOD, which existing TOD methods may not find (Lee and Lee, 2019). Lee et al. (2009) converted patent data to a term-document matrix and then used principal component analysis to generate a two-dimensional patent map; vacant spaces on this map represent possible opportunities to create new technology. Wang and Chen (2019) proposed a method that uses latent semantic analysis to transform patent data to low-dimensional latent features, and then used cosine similarity to discover outlier patents, which can be interpreted as latent technology opportunities.

Advances in deep learning-based text mining algorithms have been exploited in recent TOD studies; they have focused on neural network-based text embedding methods such as Word2vec and Glove (Mikolov et al., 2013; Pennington et al., 2014), which generate word representations to capture the context of the documents. Roh et al. (2019) pproposed a way to find technical opportunities by using Word2vec and *k*-medoids to mine opinions in customer review data and patent data. They identified the technologies related to customer needs by scanning the data to find words that have similar contextual meaning. They identified the technologies related to customer needs by scanning the data to find words that have similar contextual meaning. Lee et al. (2020) proposed a product landscape analysis of a patent-product database by using Word2vec to identify product areas that can refer to the possibility of technology opportunities.

2.1.2. Network Analysis for the TOD

Network analysis is an approach to quantify interactions between actors (i.e., entities, represented as nodes on a graphical representation of a network). Several researchers have proposed this type of TOD framework, including a subject-action-object (SAO) network and patent network, because it can effectively represent the relationships among

Table 1
Basic statistics of technology and NTBF corpus

Corpus	Number of documents	Number of unique tokens	Token length/document		Minimum of token length	Maximum of token length
			Mean	Standard deviaton		
Technology	8,772	17,497	181.7	68.7	25.0	553.0
NTBF	118	2,144	346.9	183.9	35.0	2,162

technologies (Han et al., 2019; Kim et al., 2019; Von Wartburg et al., 2005; Yoon and Park, 2004; 2005).

An SAO structure consists of a subject, an action (verb), and an object, which represent interactions among the factors that constitute the technical document. The subject and object are nodes of SAO network, and indicate respectively the technology and the purpose of technology (e.g., the product); the action is a link (i.e., an edge) between the nodes, and represents the mean for purpose of the technology (Choi et al., 2012; Lee et al., 2014). Han et al. (2019) conducted a TOD study of the medical domain fields by using the abstracts of articles in medical journals to build an SAO network, and identified combinations between technologies by link prediction between nodes. However, techniques that use SAO have only considered three factors, and may have difficulty representing features of technology (e.g., specific applicable fields, advantageous effects of inventions) other than SAO factors (Kim et al., 2019).

Furthermore, some TOD studies that use patent analysis have generated a patent network that is composed of patents as nodes, and patent relationships (e.g., citations) as links. Yoon and Park (2004) constructed a patent network by extracting keyword vectors from each patent and calculating the similarity between patents to create an association matrix, and then applied the degree centrality measure to the association matrix to generate several indices that can quantify the importance, newness, and similarity of patents. In addition, a patent-citation network composed of the cross-citation information of patents can be used quantify the value a of technology opportunity by using in-degree and out-degree centrality measures adjusted to the asymmetric nature of the citation network (Von Wartburg et al., 2005). Kim et al. (2019) proposed a TOD study to generate technology network indices by using a patent network that is constructed by considering the relationship between patents and their international patent classification codes; the technology convergence capability can be measured by calculating degree centrality, betweenness centrality, and closeness centrality in this network.

Overall, previous studies that used patent networks have enabled identification of technology opportunities by identifying visual patterns, and have permitted evaluation of opportunities by measuring several network-quantification indices, but the studies have limited ability to represent external factors (e.g., corporate- and investment-related information) that can be highly correlated to trends of technologies.

2.2. TOD in a corporate perspective

Recently, several studies have conducted TOD by using corporate technical data, such as corporate portfolios and patent data owned by companies. The studies have suggested ways to discover promising technologies suitable for the characteristics of a target company for its R&D planning (Lee et al., 2017).

Choi et al. (2019) defined firms which have a large number of registered patents as precedent enterprises (PEs) that have already undergone technological changes. Then they generated sequences using patent data of PEs, representing the temporal change in their technology fields. They proposed a TOD method that uses using a sequential-pattern mining algorithm, PrefixSpan, to find important sequences for the target firm similar to the PE sequence, by considering the technology similarity, business stability, and recency from the PE sequence. Lee et al. (2014) identified new technology opportunities for SMEs by combining an expert's analysis of keywords, with an SAO network derived from the SMEs' patent data. Park and Yoon (2017) used firms' technology portfolios to analyze TOD that supports the firms' decision-making tailored to their technology preferences. They applied collaborative filtering to calculate the similarity between the target company and other companies, and then recommended the technology with the highest preference among the technologies that the target company has not explored yet.

In similar contexts as those of prior studies, our proposed framework defines an effective TOD index that represents the corporate perspective by using deep learning-based text mining and network analysis (i.e., graph analysis). However, no study has analyzed TOD for NTBFs and their investors, despite the importance of such factors.

3. Material

Our TOD framework considers three datasets, technology text data, NTBF data, and data on investors who have invested in NTBFs. First, we collected the technical data, the technology roadmap (TRM), which is a report that summarizes trends in promising technology sectors such as artificial intelligence (AI), big data, and future automobiles in South Korea from 2016 to 2019 (Smeroadmap, 2019). TRM data can be divided into small documents at the level of a paragraph, and the documents have technology labels with a hierarchical structure from upper-level technologies (i.e., category) to lower-level technologies (i.e., subcategory); the technology labels are organized into a two-level tree structure, of which a category covers multiple subcategories exclusively. For example, two documents with the same category 'AI' can have different subcategories such as 'Human-AI collaboration system' and 'AI software', and do not fall into other categories than AI. The final technology text data contained 8,772 documents with 24 categories and 269 subcategories, and the examples of TRM data are shown in Table A1.

The data were preprocessed. We first removed special characters, numbers, and whitespace. Then we tokenized the corpus into words (i.e., tokens), and then to minimize the use of redundant words unrelated to technology objects, we extracted the nouns by extracting the morpheme of each word. Then we eliminated stop-words and applied Zipf's law to the corpus of nouns. Zipf's law means that the frequency of a word is proportional to its rank, and the frequency and rank of a word follow a probability distribution in the form of a power law (Li, 1992; Powers, 1998). According to Zipf's law, we can remove very common words or rare words that are not necessary to understand the core meaning of the sentence, which degrade the accuracy of the text embeddings models. Finally, we obtained a total of 8,772 preprocessed technology text data entries composed of 17,497 unique noun tokens.

Next, we collected NTBF data from the Rocket Punch website (Rocketpunch, 2019). Rocket Punch is one of the large business networking platforms in Korea, where numerous companies, including NTBFs, participate and share their reliable corporate information to promote themselves to investors and job seekers. Rocket Punch data includes the company's various information such as its technical investor relation (IR) text, news texts, the year of establishment, and funding information. Among the companies that were registered on Rocket Punch, 446 startup ventures that were established less than five years ago were selected as NTBFs for this study. Of these 446 NTBFs, those that have been shut down were excluded, and then we constructed the NTBF corpus by merging the IR texts with news articles. Then we conducted text preprocessing for the NTBF corpus in the same way as for

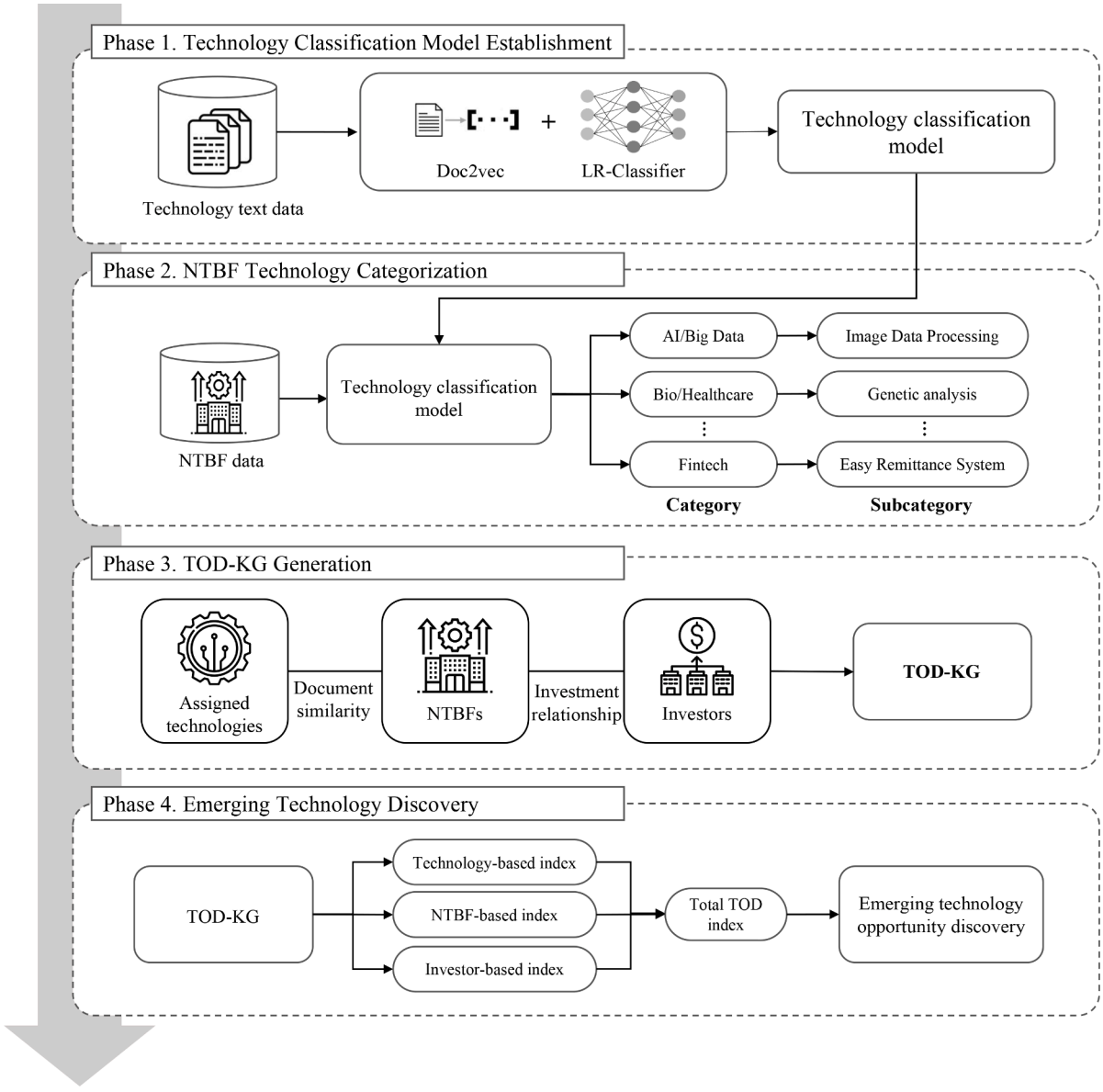

Fig. 1. Overview of the TOD framework

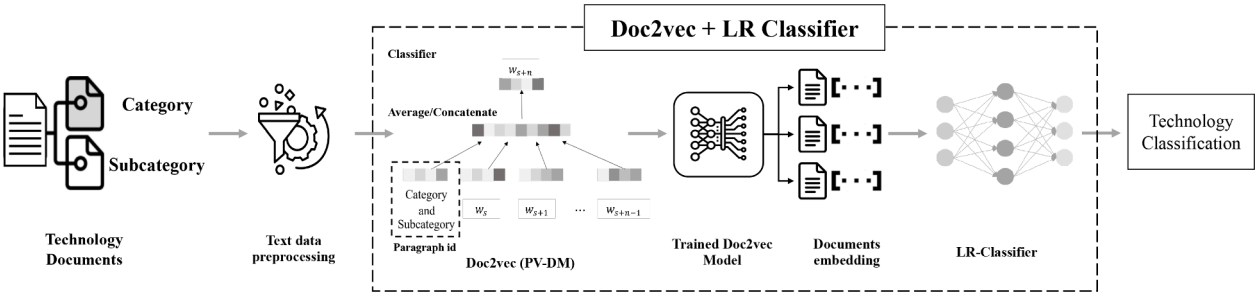

Fig. 2. Technology classification process

the technology text data. Text that is too short can have insufficient information, so we removed text that had fewer than 30 tokens. Finally, 118 NTBF samples that contained 2,144 unique noun tokens for NTBF technology categorization were prepared. The variable description of NTBF data is described in Table A2 and the basic statistics of technology and NTBF corpus are shown in Table 1.

Finally, we collected investment data about 121 investors who invested in the NTBFs from Rocket Punch. These data include the funding information such as the total number of investments and the investment amount (see Table A3).

Table 2
Technology classification results for category and subcategory

Label type	Classifier	Accuracy (10-fold CV)	F1-score (10-fold CV)
catagory	**Doc2vec-LR**	**0.9254±0.014**	**0.9249±0.001**
	Doc2vec-MLP	0.9180±0.017	0.9175±0.012
	Doc2vec-RF	0.6977±0.017	0.6714±0.017
	Word2vec-LR	0.8471±0.015	0.8427±0.016
	Word2vec-MLP	0.8655±0.015	0.8646±0.015
	Word2vec-RF	0.7662±0.015	0.7521±0.017
	FastText-LR	0.8687±0.011	0.8657±0.011
	FastText-MLP	0.8809±0.013	0.8800±0.013
	FastText-RF	0.8072±0.007	0.7975±0.008
	Glove-LR	0.7309±0.026	0.7256±0.029
	Glove-MLP	0.7068±0.023	0.6971±0.024
	Glove-RF	0.6612±0.022	0.6406±0.023
subcatagory	**Doc2vec-LR**	**0.8899±0.012**	**0.8780±0.013**
	Doc2vec-MLP	0.8503±0.016	0.8361±0.019
	Doc2vec-RF	0.5640±0.023	0.5033±0.022
	Word2vec-LR	0.6800±0.015	0.6331±0.018
	Word2vec-MLP	0.7364±0.017	0.7165±0.019
	Word2vec-RF	0.5767±0.016	0.5292±0.017
	FastText-LR	0.7177±0.016	0.6750±0.020
	FastText-MLP	0.7741±0.019	0.7568±0.020
	FastText-RF	0.6217±0.018	0.5788±0.020
	Glove-LR	0.4713±0.026	0.4452±0.024
	Glove-MLP	0.4023±0.017	0.4023±0.018
	Glove-RF	0.3628±0.012	0.3182±0.011

4. Proposed framework

In this section, we propose a new TOD framework that consists of four phases (Fig. 1): (1) Technology classification: we construct the Doc2vec-LR classifier by combining Doc2vec and LR; specifically, we use Doc2vec to extract the document embeddings for technology text data, and then build the LR classifier for the extracted embeddings. (2) NTBF technology categorization: we assign relevant technologies to NTBFs by applying the Doc2vec-LR to NTBF data. (3) TOD-KG generation: we create the TOD-KG by connecting technology, NTBF, and investor by considering relationship extracted from the results of phase 2. (4) Emerging technology discovery: the TOD indices are calculated to discover the emerging technologies on top of the TOD-KG. Each phase of the proposed framework is detailed as follows.

4.1. Phase 1. Technology classification

Technology classification is a vital process that can allocate suitable technology for technical text data even if the input is composed of technical text data without specific labels. In this phase, we construct the Doc2vec-LR classifier for technology classification in two steps: (1) Doc2vec training and (2) LR training (Fig. 2). We use Doc2vec to transform each technology document to a vector (i.e., a document embedding). Doc2vec is an extension model of Word2vec to learn embeddings for documents by introducing a paragraph vector (e.g., a vector of a sentence, document, or chapter) (Le and Mikolov, 2014). It has two learning ways, a distributed memory model of paragraph vectors (PV-DM) and a distributed bag of words version of paragraph vector (PV-DBOW). The PV-DM maps the paragraph and context words into unique vectors to predict a target word (i.e., the next word in a context) while the PV-DBOW maps only the paragraph to the embedding to predict words randomly sampled from the paragraph. Doc2vec has three user-defined parameters: vector size δ, window size λ, and minimum frequency of words τ. This study implements the PV-DM of Doc2vec since it has demonstrated better accuracy than the PV-DBOW model in general (Le and Mikolov, 2014).

In Doc2vec training, we use the technology label (e.g., category or subcategory) of each technology document as the paragraph ID of Doc2vec; this approach is unlike the general method that uses the unique identification (ID) of each document (i.e., paragraph ID). Specifically, technology documents have two technology labels (i.e., category and subcategory), so we set the paragraph IDs as pairs of category and subcategory to represent the hierarchical structure of technology labels (Fig. 2). Then we extract the unique embeddings of technology categories and subcategories from the trained Doc2vec model, and use them to infer the document embeddings. This method to train Doc2vec has two merits. (1) The document embeddings can represent both technology label information and contexts of documents simultaneously, by learning the corpus with the category and subcategory; it has the effect of applying supervised learning to Doc2vec, which is an unsupervised text-embedding model. (2) The method can automatically extract the embeddings of technology categories and subcategories (i.e., technology embeddings), which would be embedded in the contexts of multiple documents. These technology embeddings will be used to define relationships between technologies in the TOD-KG of Phase 3.

After extraction of document embedding, we assign technology labels to each document by using the multinomial LR model that presented the best accuracy and F1 score in preliminary experiments (Table 2). Consequently, we obtain a technology-classification model called Doc2vec-LR trained with technology-relevant text data.

4.2. Phase 2. NTBF's technology categorization

To analyze the TOD effectively by representing NTBF information, unified technology category labeling for the NTBFs is essential because it enables intuitive identification of their technical fields, but generally the NTBF data has no clear and consistent technology category (label) information. Thus, the goal of this phase is to assign each NTBF to relevant technologies by applying the Doc2vec-LR model to the NTBF corpus. The method of allocating technology categories and subcategories to NTBF documents is shown in Fig. 3.

Next, we apply the Doc2vec-LR classifier to the pre-processed NTBF corpus. For this purpose, we extract the embedding of each NTBF document based on pre-trained Doc2vec, and then use the LR classifier to automatically assign the technology categories to the document. Here, we establish a hierarchical technology category allocation method based on the classification probability by slightly modifying the existing way of the Doc2vec-LR model. The modified method can identify the concurrent technical characteristics of NTBF, which cannot be covered

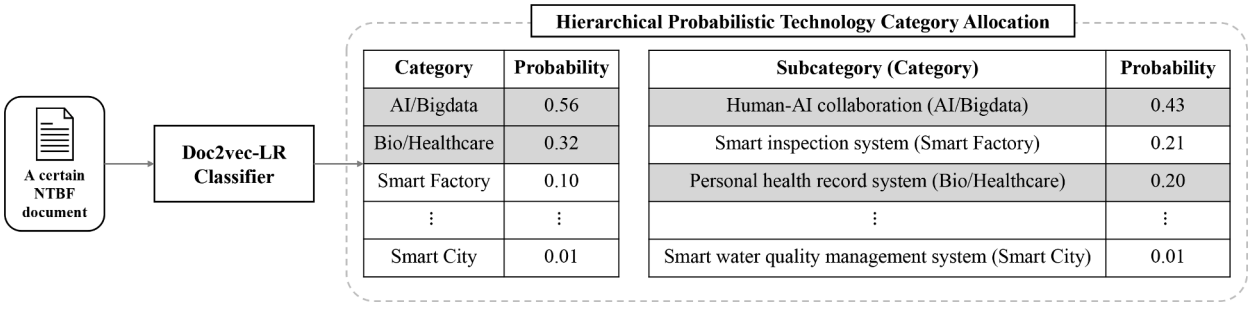

Fig. 3. Process of NTBF's technology categorization

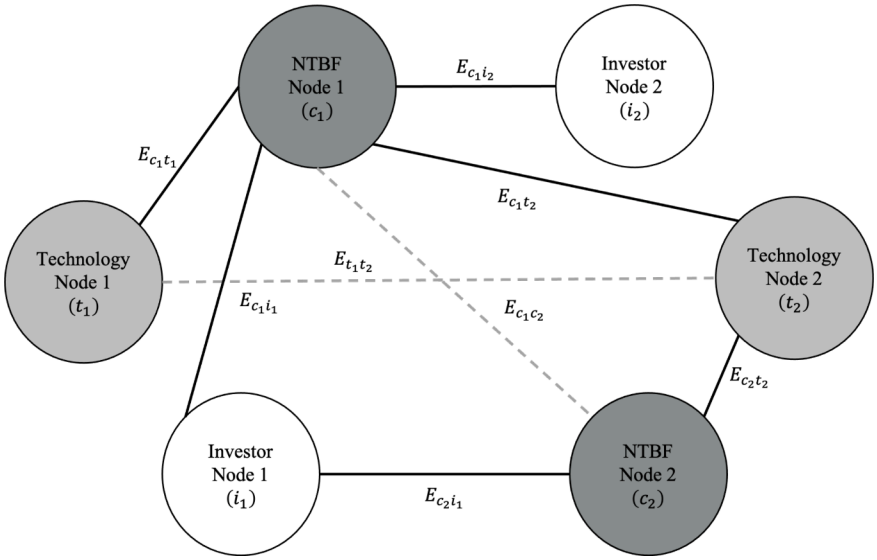

Fig. 4. Example of TOD-KG

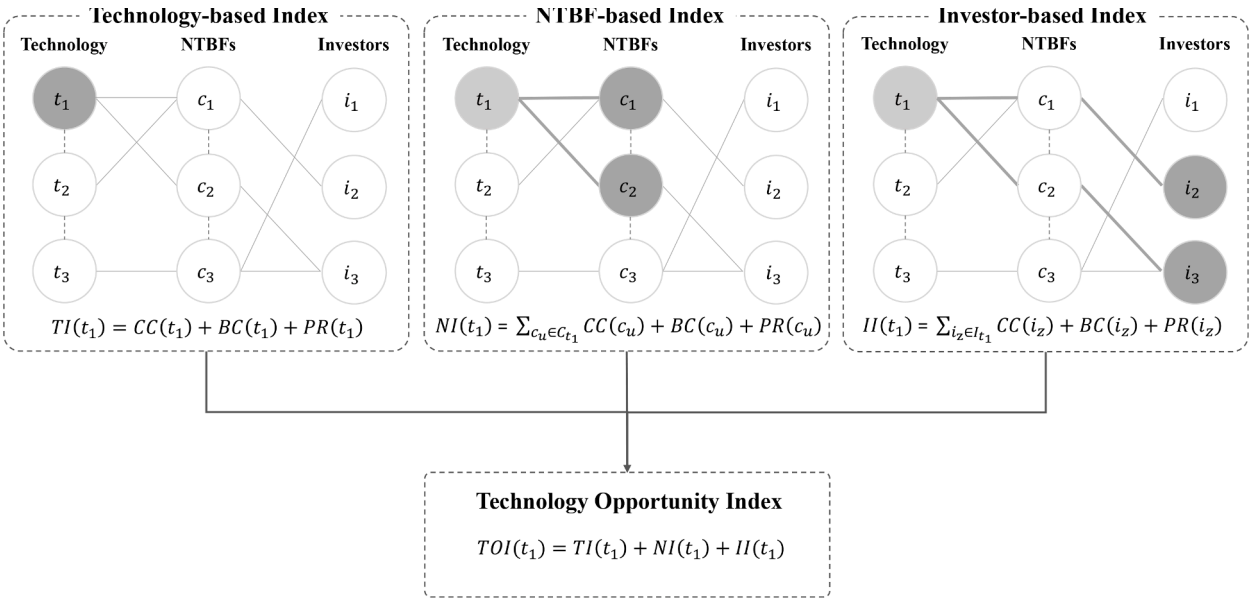

Fig. 5. Example of process for TOD index calculation

when one type of exclusive category (or subcategory) is allocated to each NTBF. We use the Doc2vec-LR model to first assign the k_c categories that had with the highest probability to each NTBF document. Then, k_s subcategories with the highest probability that belong to the k_c previously-assigned upper-level categories are allocated to the NTBF document. k_c and k_s are user-defined parameters for the number of categories and subcategories allocated to the NTBF document, respectively. Even when the subcategories with the highest probability have the same category, they are allocated in order of the probability among the subcategories. Note that the k_s subcategories do not always correspond one-to-one to k_c categories. For instance, as shown in Fig. 3 with $k_c = k_s = 2$, the categories 'AI/Bigdata' and 'Bio/Healthcare' that show the highest probability in the upper-level categories are assigned as technologies of the NTBF. For subcategory allocation, although 'Smart inspection system' has the second-highest probability, we exclude this subcategory from technology candidates of the NTBF because it does not belong to the previously assigned upper-level category (e.g., 'AI/Bigdata' and 'Bio/Healthcare'); instead, we determine 'Personal health record system' that has the next-highest probability, and that is a subcategory of 'Bio/Healthcare', as the NTBF's technology.

4.3. Phase 3. TOD-KG generation

In this study, we aim to evaluate the potential value of technologies by taking the technology, NTBF, and investor information into account for TOD analysis. To approach the problem, one possible way is to first construct a new data structure that can represent the interactions between various subjects comprising the NTBF ecosystem. A KG is a heterogeneous graph to represent different types of entities (i.e., heterogeneous nodes) and their various relationships (i.e., heterogeneous edges) (Ji et al., 2021; Yang et al., 2020). In this phase, by adopting the definition of KG, we newly construct the KG for TOD (i.e., TOD-KG) to effectively integrate the information about heterogeneous objects that have different properties in the NTBF ecosystem as a single form. Our new KG is a fundamental data structure suitable for the purpose of this study, which is used to quantify the value of lime-lighted

Table 3
Qualitative analysis for evaluation of NTBF technology categorization. Unique terms excluding common words of every document in NTBF text data and technology labels assigned to each NTBF are listed. The terms are displayed in Korean-English translated pairs

NTBF	Words	Category	Subcategory
C03	'데이터마이닝 (data mining)', '습관 (habits)', '헬스케어 (healthcare)', '위험 (risk)', '생활 (life)', '스타일 (style)', '질병 (disease)', '개선 (improve)', '활용 (utilize)', '발굴 (discovery)', '체중관리 (weight management)', '영양 (nutrition)', '어플리케이션 (mobile application)', '분석 (analysis)', '유전자 (gene)', '추천 (recommended)', '빅데이터 (big data)', '유전정보 (genome)'	'AI/Big data' 'Bio/ Healthcare'	'Genome analysis' 'Real-time healthcare system'
C04	'대출 (loan)', '분석 (analysis)', '투자금 (investment)', '소상공인 (small business owner)', '웹서비스 (web service)', '가치평가 (valuation)', '자금 (fund)', '담보평가 (valuation of mortgage)', '투자 (investment)', '모바일앱 (mobile application)', '빅데이터 (big data)', '플랫폼 (platform)', '담보가치 (value of mortgage)', '건설자금 (construction fund)', '부동산 (real estate)', '온라인 (online)', '정보 (information)', '변화 (change)', '핀테크 (fintech)', '계약 (construction)'	'Fintech' 'AI/ Big data'	'Big data software' 'Fintech big data analysis and application system'
C07	'아이템 (item)', '예약 (reservation)', '자연어처리 (natural language processing)', '매장 (store)', '구매 (purchase)', '맛집 (famous restaurant)', '배달 (delivery)', '메신저 (messenger)', '인공지능 (artificial intelligence)', '거래 (deal)', '중개 (brokerage)', '대화 (conversation)', '커머스 (commerce)', '개인비서 (personal assistant)', '텍스트 (text)', '고객 (client)', '일정관리 (schedule management)', '챗봇 (chatbot)', '문자 메시지 (text message)', '채팅 (chatting)'	'Service platform' 'AI/ Big data'	'Text mining used natural language processing' 'Human-AI collaboration system'
C17	'플랫폼 (platform)', '결제서비스 (payment service)', '데이터 (data)', '편리 (convenient)', '결제시장 (payment market)', '결제 (payment)', '중개 (brokerage)', '여행 (tour)', '중개서비스 (brokerage service)', '확장 (expand)', '핀테크 (fintech)'	'O2O service' 'Fintech'	'O2O service for tourists' 'Fintech big data analysis and application service'
C26	'투자 (investment)', '투자유치 (attract investment)', '안정성 (stability)', '헤지펀드 (hedge fund)', '목표 (aim)', '자산관리 (asset management)', 'AI (artificial intelligence)', '공학 (engineering)', '프로세스 (process)', '자산 (asset)', '알고리즘 (algorithm)', '수익 (profits)', '제도 (system)', '분야 (sector)', '핀테크 (fintech)', '수수료 (fee)', '자동화 (automation)'	'Fintech' 'AI/ Big data'	'Asset management system' 'Fintech big data analysis and application system'
C39	'전환 (switch)', 'sw직군 (software job group)', '머신러닝 (machine learning)', '경험 (experience)', '전문가 (expert)', '이직 (turnover)', '추천 (recommendation)', '컨설팅 (consulting)', '딥러닝 (deep learning)', '인재 (talent)', '인공지능 (artificial intelligence)', '산업 (industry)', '과제 (assignment)', '양성 (training)', '커리어 (career)', '테스트 (test)', '코딩 (coding)', '오픈소스 (opensource)', '러닝플랫폼 (learning platform)', '특화 (specialized)', '공부 (study)', '전략 (strategy)', '커리큘럼 (curriculum)', '개발자 (developer)', '강의 (lecture)'	'Service platform' 'AI/ Big data'	'AI software' 'Edutech'
C44	'플랫폼 (platform)', '용품 (equipment)', '제약회사 (pharmaceutical company)', '합병증 (complications)', '식재료 (food ingredient)', '마케팅 (marketing)', '헬스케어 (healthcare)', '병원 (hospital)', '커머스 (commerce)', '온오프라인 (online and offline)', '질병 (disease)', '식자재회사 (grocery store)', '운동 (excersise)', '완치 (a complete cure)', '건강 (health)', '보건소 (public health center)'	'O2O service' 'Bio/ Healthcare'	'Personalized healthcare monitoring devices and platforms' 'O2O service data analysis system'
C61	'진료 (medical care)', '기록 (record)', '관리 (care)', '내역 (history)', '헬스케어 (healthcare)', '환자 (patient)', '치료 (treatment, care)', 'EMR (Electronic Medical Record)', '조회 (lookup)', '식사 (meal)', '생활 (life)', '질병 (disease)', '운동량 (exercising)', '패턴 (pattern)', '분석 (analysis)', '리포트 (report)', '머신러닝 (machine learning)', '처방전 (prescription)', '건강 (health)', '대학병원 (university hospital)'	'AI/Big data' 'Bio/ Healthcare'	'Data analytics in smart healthcare' 'Big data collection system for smart healthcare'
C72	'근처 (nearby)', '제품 (goods)', '서비스 (service)', '상점 (shop)', '특화 (specialized)', '판매 (sale)', '상점주 (shopkeeper)', '공간공유모델 (space sharing model)', '여행객 (traveler)', '이동경로 (travel route)', '위치기반서비스 (location based service)', '홍보 (promotion)', '여행 (travel)'	'O2O service' 'Service platform'	'O2O service for tourists' 'SNS platform'

technologies by reflecting intrinsic relationships between heterogeneous entities of the graph comprehensively. Fig. 4 presents the structure of the TOD-KG.

The TOD-KG has three types of nodes (i.e., entities): NTBF, NTBF's technology, and NTBF's investor. The node set of TOD-KG, $V = \{C, T, I\}$ where $|V| = N$, consists of NTBF node set $C = \{c_1, \ldots, c_{N_C}\}$, NTBF's

Table 4
Comparison for cosine similarity between text embeddings of TRM and NTBF text data ((A) TRM documents in a subcategory, (B) NTBF documents in a subcategory (C) NTBF documents in other subcategories not belong to (A) and (B))

Subcategory	Average cosine similarity between (A) and (B)	Average cosine similarity between (A) and (C)
Medical data management systems	**0.7814**	0.6140
Video media platform	**0.7271**	0.6058
Big data construction and analysis system for logistics service	**0.7073**	0.6044
Data security and de-identification	**0.7064**	0.6140
Asset management system	**0.7036**	0.6216
Genome analysis	**0.7001**	0.6136
Bio-derived material analysis system	**0.6972**	0.5998
Crowdfunding	**0.6969**	0.6114
Bio-signal measurement and diagnosis device	**0.6921**	0.5901
Data Analytics in Smart Healthcare	**0.6903**	0.6167
Molecular diagnostics	**0.6852**	0.5906
Simple money transfer and payments	**0.6839**	0.6092
E-commerce platform	**0.6837**	0.6134
Fintech big data analysis and application service	**0.6823**	0.6234
O2O service for tourists	**0.6822**	0.6184
SNS platform	**0.6813**	0.6233
Health functional food	**0.6807**	0.5973
Big data collection system for smart healthcare	**0.6771**	0.6224
Human-AI collaboration system	**0.6771**	0.6307
Pet care O2O service	**0.6755**	0.6191
Mobile fintech security	**0.6747**	0.6131
O2O service data analysis system	**0.6653**	0.6206
O2O service platform	**0.6632**	0.6111
Speech recognition system	**0.6614**	0.6147
Visualization tool for data 3D transformation	**0.6606**	0.6073
Mobile application platform	**0.6590**	0.6182
Edutech contents	**0.6585**	0.6132
Big data software	**0.6581**	0.6247
Personal heath record system	**0.6572**	0.6149
Text mining used natural language processing	**0.6551**	0.6255
AI Software	**0.6509**	0.6365
Real-time healthcare system	**0.6496**	0.6054
Personalized healthcare monitoring devices and platforms	**0.6492**	0.6239
Image data-based Artificial Intelligence service	**0.6475**	0.6173
Game platform	**0.6434**	0.6157
Big data analysis and visualization platform	**0.6404**	0.6136
Healthcare design	**0.6375**	0.6163
Sentimental information analysis	**0.6361**	0.6182
Robotic Process Automation	**0.6353**	0.6164
Cognitive science software	**0.6212**	0.6168
Immunochemical diagnosis	**0.6192**	0.5986

technology node set $T = \{t_1, \ldots, t_{N_T}\}$, and investor node set $I = \{i_1, \ldots, i_{N_I}\}$ which are the investors who have invested in the NTBFs. $|C| = N_C$, $|T| = N_T$, and $|I| = N_I$ are the number of nodes in C, T, and I respectively, conditioned with $N = N_C + N_T + N_I$.

The edges in the TOD-KG represent four types of connectivity:

Table 5
Basic statistics for TOD-KG

Component	Statistics
Number of nodes	202
Number of NTBF nodes	77
Number of technology nodes	42
Number of investor nodes	83
Number of edges	593
Average degree	5.875
Average weighted degree	2.497
Average closeness centrality	0.308
Average betweenness centrality	234.631
Average PageRank centrality	0.005

- The edge between NTBF $c_u \in C$ and NTBF's technology $t_v \in T$ (denoted as $E_{c_u t_v}$) has connectivity when t_v is the technology labels allocated to c_u, where $u = 1, \ldots, N_C$ and $v = 1, \ldots, N_T$.
- The edge between NTBF c_u and NTBF's investor i_z (denoted as $E_{c_u i_z}$) has connectivity when i_z, $z = 1, \ldots, N_I$, is the investor who has invested in c_u. To determine $E_{c_u i_z}$, the frequency or amount of investment is not considered here.
- The edge between two certain technologies t_v and $t_{v'} \in T \backslash \{t_v\}$ (denoted as $E_{t_v t_{v'}}$) has connectivity when the cosine similarity between their document embeddings $\cos(d_{t_v}, d_{t_{v'}}) > \gamma$ where $d_{t_v} \in \mathbb{R}^\delta$ and $d_{t_{v'}} \in \mathbb{R}^\delta$ are the document embeddings of t_v and $t_{v'}$ extracted from Doc2vec.
- The edge connectivity between two NTBFs c_u and $c_{u'} \in C \backslash \{c_u\}$ (denoted as $E_{c_u c_{u'}}$) has connectivity when the cosine similarity between their document embeddings $\cos(d_{c_u}, d_{c_{u'}}) > \xi$ where d_{c_u} and $d_{c_{u'}}$ are the document embeddings of c_u and $c_{u'}$. γ and ξ are user-defined parameters for the similarity threshold of edge connection.

Subsequently, $E_{c_u t_v}$ and $E_{c_u i_z}$ are defined as direct interconnections between nodes, whereas $E_{t_v t_{v'}}$ and $E_{c_u c_{u'}}$ represent the coherence between the document embeddings of nodes.

4.4. Phase 4. Emerging technology discovery

TOD analysis urgently requires a precise and overall index to quantitatively evaluate technology opportunities, so we introduce a TOD index that is calculated from the TOD-KG by measuring the network centrality to represent the influential power of the entities of TOD-KG. The assumption of the TOD index is that a high centrality of a technology can represent a high potential value of technology opportunities in the NTBF-related technological ecosystem. Specifically, in this phase, we generate a technology opportunity index (TOI) (Fig. 5) by aggregating three different TOD indices, i.e., a technology index (TI), an NTBF index (NI), and an investor index (II), in terms of three perspectives, technology, NTBF, and investor, respectively. The TOI can represent the characteristics of main subjects and their relationship in the NTBF ecosystem. The processes for computing the TOD indices are explanations below.

First, the TI of the technology node, $TI(t_v)$, is the connection centrality of a technology among nodes connected to the target technology node. To calculate $TI(t_v)$, we use three well-known centrality measures of a network: closeness centrality (CC), betweenness centrality (BC), and page rank (PR). CC of a node is defined as the inverse of the average shortest path distance to all other nodes; $CC(t_v) = \frac{N-1}{\sum_{j \in V \backslash \{t_v\}} d(t_v, j)}$, where $d(t_v, j)$ is the length of the shortest path from node t_v to node $j \in V \backslash \{t_v\}$. BC refers to the frequency at which a node is on the shortest path that connects two other nodes except itself, i.e., $BC(t_v) = \sum_{j \in V, k \in V \backslash \{j\}} \frac{\sigma_{j,k}(t_v)}{\sigma_{j,k}}$ where $\sigma_{j,k}$ is the number of shortest paths between nodes j and k and $\sigma_{j,k}(t_v)$ is $\sigma_{j,k}$ that passes through node t_v. PR ranks the importance of a

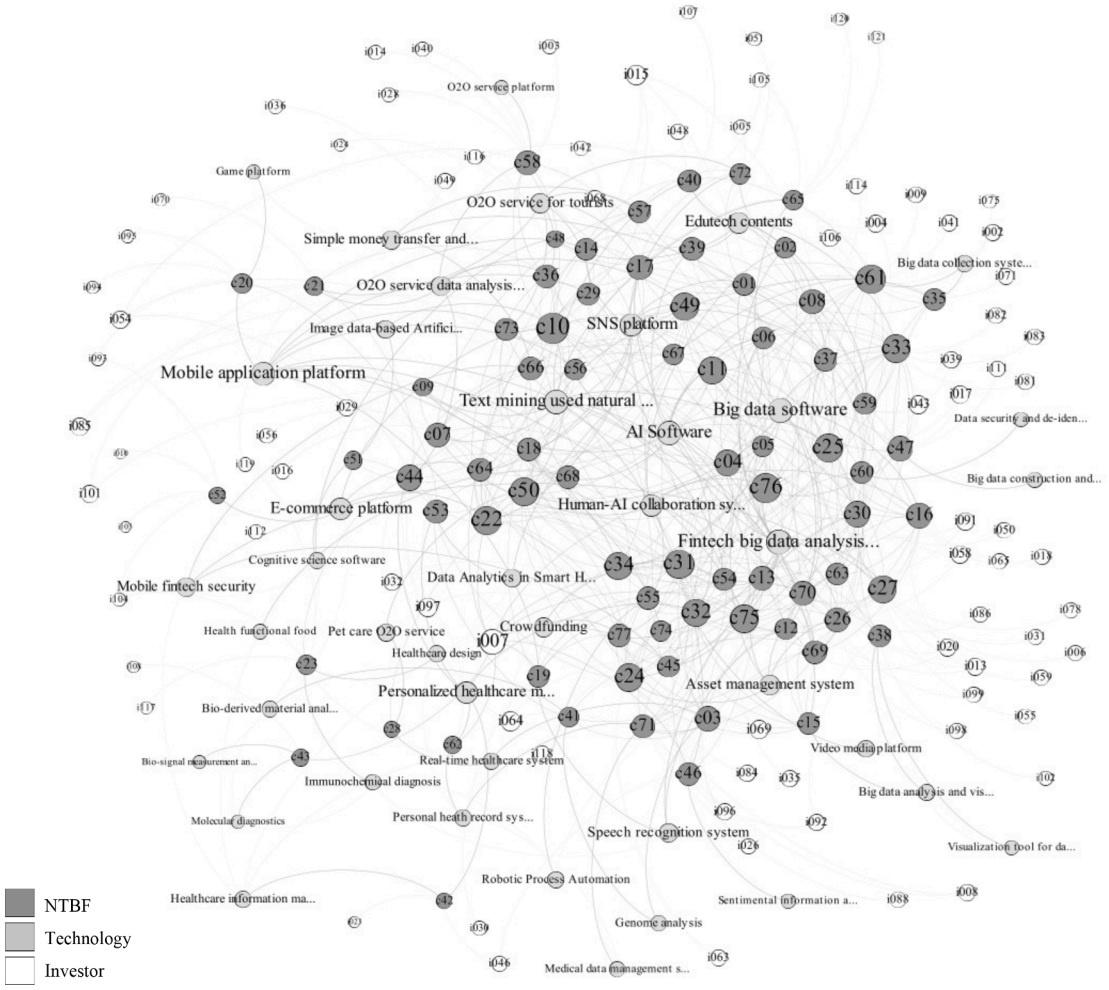

Fig. 6. Example of the entire TOD-KG. Dark gray nodes: NTBFs; white nodes: investors; light gray nodes: technologies. Node size represents the average value of three centrality measures. The size of each technology node does not indicate the TOI index.

node in relation to the other nodes. Initially, PR values of all nodes are set to $\frac{1}{N}$, and at each iteration, $PR(t_v)$ is iteratively calculated as $PR(t_v) = \sum_{j \in P_j} \frac{PR(j)}{L(t_v,j)}$, where $P_j = \{j \in V \backslash \{t_v\} | E_{jt_v} = 1\}$ is the set containing all nodes that connect to node t_v and $L(t_v, j)$ is the number of neighbors of node j.

Here, each network centrality measure has its own characteristics that can be interpreted in terms of the TOD. $CC(t_v)$ captures the closeness of technology t_v to other nodes (i.e., NTBFs and investors) in the TOD-KG. $BC(t_v)$ represents the significance of technology t_v by calculating how frequently t_v appears on all the shortest paths in the TOD-KG. $PR(t_v)$ means the connection intensity of node t_v with the other influential nodes that have a large number of neighbors and the high PR value; high $PR(t_v)$ indicates that node t_v is highly connected to the other influential nodes.

We aggregate $CC(t_v)$, $BC(t_v)$, and $PR(t_v)$ with equal weights to calculate $TI(t_v)$:

$$TI(t_v) = CC(t_v) + BC(t_v) + PR(t_v) = \frac{N-1}{\sum_{j \in V \backslash \{t_v\}} d(t_v, j)} + \sum_{j \in V, k \in V \backslash \{j\}} \frac{\sigma_{j,k}(t_v)}{\sigma_{j,k}} + \sum_{j \in P_j} \frac{PR(j)}{L(t_v, j)}. \quad (1)$$

In a similar way as for TI, the NI and II of t_v can be calculated using CC, BC, and PR. $NI(t)$ aggregates the centrality measures around the NTBFs allocated to t_v:

$$NI(t_v) = \sum_{c_u \in C_{t_v}} CC(c_u) + BC(c_u) + PR(c_u), \quad (2)$$

where $C_{t_v} = \{c_u \in C | E_{c_u t_v} = 1\}$. $II(t_v)$ aggregates the centrality measures of the investors for the technology connected via NTBFs. It is formulated by

$$II(t_v) = \sum_{i_z \in I_{t_v}} CC(i_z) + BC(i_z) + PR(i_z), \quad (3)$$

where $I_{t_v} = \{i_z \in I | E_{c_u i_z} = 1 \wedge E_{c_u t_v} = 1 \text{ for } \forall c_u\}$. $NI(t_v)$ and $II(t_v)$ mean respectively the intrinsic influence of NTBFs and investors related to t_v in the TOD-KG. Finally, we generate the TOI by aggregating the TI, NI, and II:

$$TOI(t_v) = TI(t_v) + NI(t_v) + II(t_v). \quad (4)$$

$TOI(t_v)$ represents the integrated influence of the TOD-KG nodes in different centrality measures; i.e., $TOI(t_v)$ is an overall TOD index to incorporate the TI, NI, and II that can measure the connection centrality of technology node t_v in various aspects. Weighting of the summations may be appropriate, but we do not consider this possibility here. Individually, TI, NI, and II can be also useful tools for in-depth analysis of promising technologies.

Table 6
Emerging technologies identified using the TOI. Top-5 and bottom-5 ranked technologies are marked in bold text.

Technology	TOI
Fintech big data analysis and application service	**1.101**
Big data software	**1.090**
Human-AI collaboration system	**1.089**
Text mining used natural language processing	**1.073**
AI Software	**1.073**
Mobile application platform	1.045
Asset management system	1.034
Edu-tech contents	1.016
O2O service for tourists	1.016
E-commerce platform	1.000
Mobile fintech security	0.988
Data Analytics in Smart Healthcare	0.984
O2O service data analysis system	0.978
Simple money transfer and payment	0.971
Personalized healthcare monitoring devices and platforms	0.867
Image data-based Artificial Intelligence service	0.844
Speech recognition system	0.837
Genome analysis	0.799
SNS platform	0.765
Big data construction and analysis system for logistics service	0.749
Healthcare design	0.716
Visualization tool for data 3D transformation	0.695
Crowdfunding	0.541
Healthcare information managing service	0.529
Molecular diagnostics	0.529
Video media platform	0.528
Personal health record system	0.502
Bio-derived material analysis system	0.501
Big data collection system for smart healthcare	0.495
O2O service platform	0.492
Data security and de-identification	0.492
Real-time healthcare system	0.485
Medical data management systems	0.483
Pet care O2O service	0.476
Robotic Process Automation	0.465
Sentimental information analysis	0.461
Big data analysis and visualization platform	0.452
Cognitive science software	**0.451**
Immunochemical diagnosis	**0.445**
Game platform	**0.435**
Health functional food	**0.428**
Bio-signal measurement and diagnosis device	**0.237**

5. Empirical results and analysis

5.1. Technology classification results

We first implemented Doc2vec to vectorize the TRM documents. For training, we set the parameters to $\delta = 300$, $\lambda = 10$, and $\tau = 3$, and then we extracted the embeddings of documents in the training set. Afterward, we used them as input to the LR classifier to train it to allocate both technology category and subcategory to each document. To validate the Doc2vec-LR model, we additionally considered three baseline models with different combinations of text embedding techniques such as Doc2vec and three other word embedding models, Word2vec, Glove (Pennington et al., 2014), and FastText (Joulin et al., 2016), and three classifiers such as LR, multi-layer perceptron (MLP) and RandomForest (RF). The combinations of text embedding models and classifiers are denoted as Doc2vec-MLP, Doc2vec-RF, Word2vec-LR, Word2vec-MLP, Word2vec-RF, Glove-LR, Glove-MLP, Glove-RF, FastText-LR, FastText-MLP and FastText-RF, respectively. For the Word2vec-based models, the document embeddings were generated by aggregating the embeddings of words for each document (Huang et al., 2018), as extracted using word embedding models. We compared the accuracy of Doc2vec-LR with the baseline models for technology category and subcategory classification. To reduce the sampling error in model construction, we conducted 10-fold cross validation (CV) for each classifier, and we optimized the models to maximize the F1-score for technology classification. All experiments were implemented in Python 3.7 with an i7-8700 CPU and 16 GB of RAM with a GENSIM3 module.

Table 7
Comparison of top-10 ranked emerging technologies identified using the four TOD indices

Technology	TOI	Technology	TI
Fintech big data analysis and application service	1.101	Fintech big data analysis and application service	0.408
Big data software	1.090	Big data software	0.408
Human-AI collaboration system	1.089	Text mining used natural language processing	0.395
Text mining using natural language processing	1.073	Mobile application platform	0.393
AI Software	1.073	AI Software	0.392
Mobile application platform	1.045	SNS platform	0.374
Asset management system	1.034	E-commerce platform	0.364
Edu-tech contents	1.016	Human-AI collaboration system	0.364
O2O service for tourists	1.016	Personalized healthcare monitoring devices and platforms	0.358
E-commerce platform	1.000	Edutech contents	0.356
Technology	**NI**	**Technology**	**II**
O2O service for tourists	0.410	Image data-based AI service	0.344
Fintech big data analysis and application service	0.408	Human-AI collaboration system	0.339
Asset management system	0.403	Mobile fintech security	0.338
E-commerce platform	0.399	Personalized healthcare monitoring devices and platforms	0.318
Big data software	0.397	Speech recognition system	0.309
Edutech contents	0.394	Simple money transfer and payments	0.304
SNS platform	0.390	Genome analysis	0.296
Data Analytics in Smart Healthcare	0.388	AI Software	0.296
Text mining using natural language processing	0.388	Text mining using natural language processing	0.290
Human-AI collaboration system	0.386	Big data construction and analysis system for logistics service	0.290

Table 8
Evaluation results for TOI corresponding to average IFS and IAS between R_T and R_B

		TOI	IFS		IAS	
# of technologies in each group	Group	Mean	Mean	*p*-value	Mean	*p*-value
5	R_T	1.08	6.46	0.01*	251.87	0.01*
	R_B	0.39	2.80		48.60	
10	R_T	1.05	5.77	<0.01**	201.89	<0.01**
	R_B	0.43	2.70		38.05	
15	R_T	1.02	5.28	<0.01**	159.47	<0.01**
	R_B	0.45	2.53		59.33	
20	R_T	0.96	4.82	<0.01**	146.44	<0.01**
	R_B	0.47	2.27		47.47	

Notes: ** $p < 0.01$. * $p < 0.05$.; p is denoted by p-value for t-test. Unit of IAS is 100 million KRW.

Technology-classification accuracy was calculated for category and subcategory (Table 2). The classification accuracy of Doc2vec-LR for category is superior to that for subcategory, but it shows remarkable accuracy even for > 200 subcategory labels. These results indicate that the Doc2vec training process reflecting the characteristics of both category and subcategory can be an effective way for improving the accuracy of technology classification.

5.2. NTBF technology categorization results

We applied the Doc2vec-LR to categorize the adequate technologies of 118 NTBFs. For hierarchical technology category allocation, the user-

defined parameters k_c and k_s were set to 2. Although the validation for the accuracy of NTBF technology categorization is challenging since there are no true labels (i.e., targets) for technology fields of NTBFs, we verified the NTBF technology categorization in terms of qualitative and quantitative analyses. First, we performed the qualitative analysis (Table 3) to directly compare the contents in the NTBF corpus with its categorized technology labels by listing unique words of each NTBF document after excluding the common words in it.

From the results in Table 3, the pre-trained Doc2vec-LR model could accurately categorize the technology of NTBFs. To be specific, in the case of 'AI/Big data' and 'Bio/Healthcare' categories, words such as "healthcare," "disease," "nutrition," "patient," and "big data" are prominent in the NTBF corpus, whereas the words such as "loan," "investment," "fund," "fintech," and "payment" are dominant in the NTBF corpus of 'Fintech' category. Additionally, for the technology subcategory, the NTBF of which the corpus contains "natural language processing," "messenger," and "chatbot" was classified into the 'Text mining used natural language processing' subcategory, but another NTBF corpus containing words such as "learning platform," "study," and "lecture" was classified into the 'Edutech' subcategory.

Next, for the quantitative evaluation, we compared the average cosine similarity of TRM and NTBF documents assigned to the same category with that of TRM and NTBF documents assigned to different categories of each other (see Table 4). As shown in Table 4, the similarities between documents of TRM and NTBFs with the same category are mostly higher than those with different categories. This result means that the text embeddings of TRM and NTBF documents from the same category, extracted from the Doc2vec-LR model, can have a more similar distribution than the cases with different categories. To sum up, both qualitative and quantitative evaluation results verify the accuracy of NTBF categorization using the proposed Doc2vec-LR model, and demonstrate the applicability of our model to technical text data that differ from TRM data.

5.3. TOD-KG generation results

We conducted TOD-KG generation using the technology and document embeddings extracted from Doc2vec and the relationship of NTBF and technology labels from NTBF technology categorization. To generate the TOD-KG, we focused on elements such as NTBFs and technology subcategories in five technology categories: 'AI/Bigdata,' 'Service platform,' 'Offline-to-Online (O2O) service,' 'Fintech,' and 'Bio/Healthcare,' which are the top-5 ranked technology categories and account for about 70 % of the total number of technology categories allocated to the NTBFs (Table B1). Note that this process can be easily extended to cover the other technology categories for TOD-KG generation.

Furthermore, we only considered technology subcategories as the technology nodes of TOD-KG. We excluded technology categories for two reasons. First, if they were used as the nodes of TOD-KG together, we might represent redundant technology information in the TOD-KG generation; i.e., technology subcategories intrinsically contain category information because they are connected with the unique upper-level category, as a consequence of the hierarchical relationship. Second, such a relationship between category and subcategory could give strong connection centrality to every technology category node, so abnormally large TOD indexes would be assigned to them.

Finally, 77 NTBFs, 42 technology subcategories, and 83 investors who invested in the NTBFs involved in such five categories were connected in the TOD-KG. We set the user-defined parameters γ and ξ to 0.9 for the edge connection of TOD-KG. The basic statistics for the TOD-KG are presented in Table 5, and the entire TOD-KG as an actual example is displayed in Fig. 6. In Fig. 6, we can identify the core subcategory technology nodes such as 'Big data software,' 'AI software,' and 'Fintech big data analysis and application service' intuitively, which are densely coherent to NTBF and investor nodes. The sub-TOD-KG for each of five technology categories was also visualized (see Appendix C); the whole graph of Fig. 6 is divided into five sub-graphs in terms of each technology category.

5.4. Emerging technology discovery results

Given the TOD-KG, we performed the emerging technology discovery by calculating the TOI, TI, NI, and II for 42 technology subcategories allocated to 77 NTBFs (see Table 6 and Table D1). Most of the top-5 ranked emerging technologies sorted by the TOI are related to AI and big data or to convergence between AI and other fields such as finance. On the contrary, most of the bottom-5 ranked technologies are involved in games, health food, and chemistry.

The top-10 ranked emerging technologies were listed by each of the four indices, TOI, TI, NI, and II (Table 7). TI and TOI show top-tier scores to similar emerging technologies, but NI and II give top ranks to technologies that differ from those identified by TI and TOI. Especially, NI (the NTBF perspective) identifies 'O2O service for tourists' and 'Asset management system' as promising, whereas II (the investor perspective) identifies 'Image data-based AI service,' 'Speech recognition system,' and 'Genome analysis', which do not have high scores in the other TOD indices. These results demonstrate that the various perspectives can identify different emerging technologies.

6. Discussion

6.1. Evaluation of the TOI in real investment decision

This study discovered the emerging technology opportunities using the TOI proposed by reflecting the NTBF ecosystem. It is necessary to economically and quantitatively evaluate those, but it is not an easy task due to the intrinsic properties of the NTBF ecosystem, such as the lack of a firm's financial accounting information and unclear cost information. Here, for the evaluation, we instead consider economic valuation from the real investment decisions of venture capitalists who invested in the NTBFs. These decisions could be reliable evidence of the perceived value of emerging technologies presented by the NTBF ecosystem, because the investors' main considerations include the value and novelty of an NTBF's technology and the future growth of the NTBF (Han and Hwangbo, 2020; Heo, 2020).

To be specific, using the number and amount of investments in each NTBF, we computed two measures to validate the TOI: (1) investment frequency score: $IFS(t_v) = \frac{1}{N_{C_{t_v}}} \sum_{c_u \in C_{t_v}} IF_{c_u}$ and (2) investment amount score: $IAS(t_v) = \frac{1}{N_{C_{t_v}}} \sum_{c_u \in C_{t_v}} IA_{c_u}$, where IF_{c_u} and IA_{c_u} are total investment frequency and amount of c_u respectively, and $N_{C_{t_v}}$ is the number of NTBFs belong to each t_v. Then, we sorted the technologies by TOI values in descending order and then classified them into a top-ranked group R_T and a bottom-ranked group R_B by their ranks according to the predefined number of technologies included in each group. We finally evaluate the TOI in terms of its ability to distinguish emerging technologies in R_T and R_B by examining the difference of the averaged IFS and IAS between two groups, respectively.

The TOI validation used a statistical t-test to compare the average IFS and IAS between R_T and R_B with various numbers of technologies selected in each group (Table 8). Our proposed TOI shows statistically significantly higher average IFS and IAS in R_T compared to those in R_B at all the number of technologies. Furthermore, when the number of technologies in R_T increases, R_T includes even technologies that are assigned low potential by the TOI. Because of this phenomenon, both IFS and IAS averages gradually decrease. This result is noteworthy for two reasons. First, in a real investment decision, a consistent trend between two investment measures and TOI is hard to achieve because of the existence of many countercases that have high frequency but a small amount of investment, or vice versa. Second, this implies that the proposed TOI can be used as a reliable indicator to guide practical decision-making for several stakeholders in the NTBF ecosystem.

6.2. Further analysis on TOD results

This section aims to provide an interpretation of emerging technologies that were identified using our proposed TOD index across various industry and academic fields. The TOI identified that 'Fintech big data analysis and application service' is a top-tier emerging technologies (Table 6). This technology is a convergence of fintech and big data analysis fields. After the global financial crisis in 2008, the fintech industry expanded rapidly by linking internet-conducted fintech technology and its network, to AI and big data technologies (Kim et al., 2020; Lee and Shin, 2018). In particular, due to the spread of smartphones and tablet personal computers, the fintech market has been expanded with the increasing demand from mobile simple-payment and E-commerce technologies (Ju et al., 2016). This phenomenon has enabled fintech companies to amass and store a vast amount of financial data collected from such devices, so the companies could differentiate their services from other competitors by providing their customers with personalized financial services and data-driven solutions (Lee and Shin, 2018). Furthermore, recently, AI technologies have continued to converge with various industries such as finance, bio/healthcare, manufacturing, and O2O services, especially in terms of data analysis and process automation (Kim et al., 2018). Especially, 'Human-AI collaboration' technology is an intelligent interface technology to guide people's decision-making by using various technologies such as natural-language processing, computer vision, and speech recognition. Thus, this technology is widely used in various industries such as 'Edutech,' 'Marketing platforms,' and 'Customer consultation automation through chatbots,' and can create new added values by the fusion of AI technologies.

We also analyze the results of NI, TI, and II to show different perspectives for the TOD (Table 7). Subsequently, we can consider the NI as the NTBF's level of market entry. For example, 'O2O service for tourists' has the highest NI; this result can be interpreted to mean that many NTBFs have interest in technology related to 'O2O services for tourists,' and it also has high connectivity with other technologies. An O2O service is a promising business model that connects online and offline commerces on a mobile app, and can help customers in the sharing economy in many ways by integrating diverse services such as transportation, accommodation, and food delivery (Du and Tang, 2014; Ryu et al., 2020). In fact, the potential of 'O2O service for tourists' can be considered an institutional peculiarity, especially in South Korea, which has allowed NTBFs to provide accommodation-sharing services for both domestic and foreign tourists since 2019. With the policy, many NTBFs were founded using novel technologies and services with the inter-industry connection of the O2O industry; especially, NTBFs' market entry into this technology sector has been more active than other O2O services. For another instance, we can adopt the evidence to identify 'Fintech big data analysis and application service' as a promising fintech technology in terms of both the NI and TI, from the following two aspects. One is the technical aspect of combining fintech technologies such as peer-to-peer network loans, simple payment, remittance systems, and AI technologies to create a suitable environment for personalized fintech services to users (Suh and Kim, 2019). With increase in users' needs for efficient and convenient fintech services that use AI, many NTBFs have been developing and patenting related technologies to meet these needs and thereby increase market share. The other aspect is a market-entry barrier for new companies due to the fintech technology regulations, which are related to information protection on electronic financial transactions in the banking, insurance, and financial investment sectors (Choi and Kim, 2019). Since 2019, policymakers worldwide have created finance-regulatory sandboxes that ease these regulations to foster innovation in the financial sector so the potential of this technology is expected to increase for NTBFs continuously.

From the investor perspective, 'Genome analysis' and 'Personalized healthcare monitoring devices and platforms' received a large amount of attention from the investors; it means that the investors foresee a great potential for such technology fields. In particular, in South Korea, the healthcare industry is a field that has a great investment ripple effect, and government and public investments are also actively progressing (Jang and Kim, 2021). More specifically, South Korea has sufficient digital healthcare infrastructure, such that the number of public medical big data is approximately six trillion, and the penetration rate of electronic medical records approaches nearly 92 %. Thus, decision-makers in the investment industry have eagerly invested in healthcare-related technologies with the expectation of potential market demand, growth potential, and future value-added creation in the healthcare industry.

However, use of AI technologies in the bio/healthcare industry raises possible ethical and legal concerns, such as accountability, transparency, and privacy preservation for private and sensitive data collection (Davenport and Kalakota, 2019; Lee et al., 2018), so the barriers to market entry for these NTBFs were high. Although the legal and regulatory issues for data were partially resolved with the passage of three data-related bills in 2020, the bio/healthcare industry has a very high entrance level because these complex questions remain, compared to other investment prospects (Kim, 2020). Consequently, the future value of the Bio/healthcare industry has increased as a result of a convergence of AI and healthcare technology, but remains an industry in which technology opportunities coexist with risks.

7. Conclusions

In this paper, we presented a new systematic framework to identify potential technology opportunities. We proposed a text mining technique that uses Doc2vec and LR models to develop a technology-classification model called Doc2vec-LR, then it to categorize each NTBF accurately into technology fields. Next, we constructed the TOD-KG consisting of three factors, technology, NTBF, and NTBF's investor. We then used the TOD-KG to measure the potential value of technology opportunities with network measurement indicators. We analyzed the top-scored promising technologies as identified by TOD indices, and provided evidence and interpretation of technologies with the actual situations of the practice industry. This study proposed a functional technology intelligence system for providing the meaningful TOD indices, and our results will be a basis for the TOD to represent the information of NTBF and investment.

Although our study makes several contributions, some limitations and future work remain. First, we performed TOD analysis using technical documents and NTBF information in Korean; because of this focus, the usability of our approach may be debatable. However, our framework considered only keywords (i.e., nouns) that are a general form of part-of-speech, and that exist in all languages. Thus, our model can be easily applied to technical text data in other languages after text-preprocessing to extract the nouns. In addition, we used data that had already been accumulated from the past for TOD analysis, so the analysis should be repeated on the latest data to identify whether the changes cause the difference in the predicted evolution of future technology. However, most of these models or systems need to be updated periodically to obtain the latest analysis results when new data is accumulated. Regarding this, our framework can be easily implemented again by pre-processing in the form of input data to the framework even if data is updated.

CRediT authorship contribution statement

MyoungHoon Lee: Conceptualization, Methodology, Software, Writing – original draft. **Suhyeon Kim:** Conceptualization, Methodology, Software, Writing – original draft. **Hangyeol Kim:** Data curation, Investigation. **Junghye Lee:** Conceptualization, Methodology, Supervision, Writing – review & editing.

Acknowledgement

Funding: This work was supported by the National Research Foundation of Korea (NRF) Grant funded by the Korea Government (MSIT) under Grant 2020R1C1C1011063 and by Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (No. 2020-0-01336, Artificial Intelligence graduate school support (UNIST)). This work was also supported

by Korea Environment Industry & Technology Institute (KEITI) through (Digital Infrastructure Building Project for Monitoring, Surveying and Evaluating the Environmental Health), funded by Korea Ministry of Environment (MOE) (No. 2021003330001).

Appendix A

Table A1
A few examples of TRM data

Tokenized document	Category	Subcategory
'의료시스템 (medical system)', '유전체정보 (genome information)', '검사 (test)', '개인유전자 (personal gene)', '축적 (accumulation)', '개인별 (individual)', '맞춤의학 (personalized medicine)', 'ngs (ngs)', '체질 (constitution)', '빅데이터 (big data)', '임상 (clinical)', '소비자 (consumer)', '유전자 (gene)', '데이터 (data)', '건강정보 (health information)', '가능해짐 (made possible)', '생명정보 (life information)', '가치사슬 (value chain)', '임상정보 (clinical information)', '웨어러블 (wearables)', '생명정보학 (bioinformatics)'	Bio/ Heathcare	Genome analysis
자문 (advice)', '요인 (factor)', '오류 (error)', '지불 (payment)', '위험 (danger)', '주문 (order)', '트렌드 (trend)', '경쟁 (compete)', '은행 (bank)', '수익구조 (revenue structure)', '수익성 (profitability)', '대면 (interview)', '해킹 (hacking)', '금융상품 (financial products)', '자산 (asset)', '시스템적 (systematic)', '관리체계 (management system)', '금융소비자 (financial consumer)', '자본시장 (capital market)', '운영 (operation)', '분석 (analyze)', '온라인 (online)', '저성장 (low growth)', '장기적 (long term)', '수익률 (yield)', '금융기관 (financial institution)', '운용 (operation)', '자산관리서비스 (wealth management service)'	Fintech	Asset management system
'통계적 (statistical)', '인식기술 (recognition technology)', '엔터테인먼트 (entertainment)', '응용서비스 (application service)', '추론 (inference)', '패턴인식 (pattern recognition)', '한국어 (korean)', '단어 (word)', '음소 (phoneme)', '입력 (input)', '발음 (pronunciation)', '음향 (acoustic)', '질의응답 (q & a)', '이해 (understand)', '음성언어 (spoken language)', '음향모델 (acoustic model)', '음성데이터 (voice data)', '신경망 (neural network)', '데이터 (data)', '분류 (classification)', '자동번역 (automatic translation)', '대화처리 (conversation processing)', '텍스트 (text)', '음성합성 (speech synthesis)', '인공지능 (a.i)', '인식대상 (recognition target)', '형태소 (morpheme)', '분석 (analyze)'	AI/Bigdata	Speech recognition system
'o2o (o2o)', '스타일 (style)', '매장 (store)', '근접 (almost)', '체크아웃 (check out)', '렌터카 (car rental)', '부여 (grant)', '예상 (prediction)', '추구 (pursuit)', '거래 (deal)', '화상처리 (image processing)', '센서 (sensor)', '음식 (food)', '포장 (packaging)', '매칭 (matching)', '판매현황 (sales status)', '중개 (mediation)', '오프라인 (offline)', '식당 (restaurant)', '숙박 (lodgment)', '빅데이터 (big data)', '분석공유 (analysis sharing)', '실시간 (real time)', '쿠폰 (coupon)', '유통 (circulation)', '수집저장 (collection and storage)', '객실 (guest room)', '동일 (same)', '대금 (price)', '정확 (exact)', '온라인 (online)', '숙박업소 (accommodation)'	O2O service	O2O service for tourists

Table A2
Variable description of NTBF data

Variable name	Description	Example
Company name	Anonymized NTBF name	C61
Tokenized document	A document after pre-processing NTBF's IR text and related news.	'진료 (medical care)', '기록 (record)', '관리 (care)', '내역 (history)', '헬스케어 (healthcare)', '환자 (patient)', '치료 (treatment, care)', 'EMR (Electronic Medical Record)', '식사 (meal)', '생활 (life)', '질병 (disease)', '운동량 (exercising)', '패턴 (pattern)', '분석 (analysis)', '리포트 (report)'
The year of Establishment	The year of establishment	2017

Table A3
Variable description of investment data

Variable name	Description	Example
Company name	Anonymized NTBF name	C61
Investments frequency	The total number of investment attraction	3
Investment amount	The total amount invested by an investor (100 million KRW)	20.9
Investors	List of Anonymized investors investing in each NTBF	I002, I009, I041, I071

Appendix B

Table B1
Ratios of technology categories allocated to NTBFs

Technology category	Ratio of categorized NTBFs (%)
AI/Big data	**23.42**
Service platform	**18.47**
Offline-to-Online service	**8.56**
Fintech	**8.11**
Bio/Healthcare	**5.86**
Digital contents design	5.41
Smart city	4.5
Blockchain	3.6
Smart factory	3.6
IoT	3.6
Display	2.7
AR/VR	2.25
Smart media devices	1.8
Security	1.8
3D printing	1.35
Cosmetics	1.35
Embedded SW	1.35
Robotics	0.9
LED/Shining	0.9
Energy	0.45

Appendix C

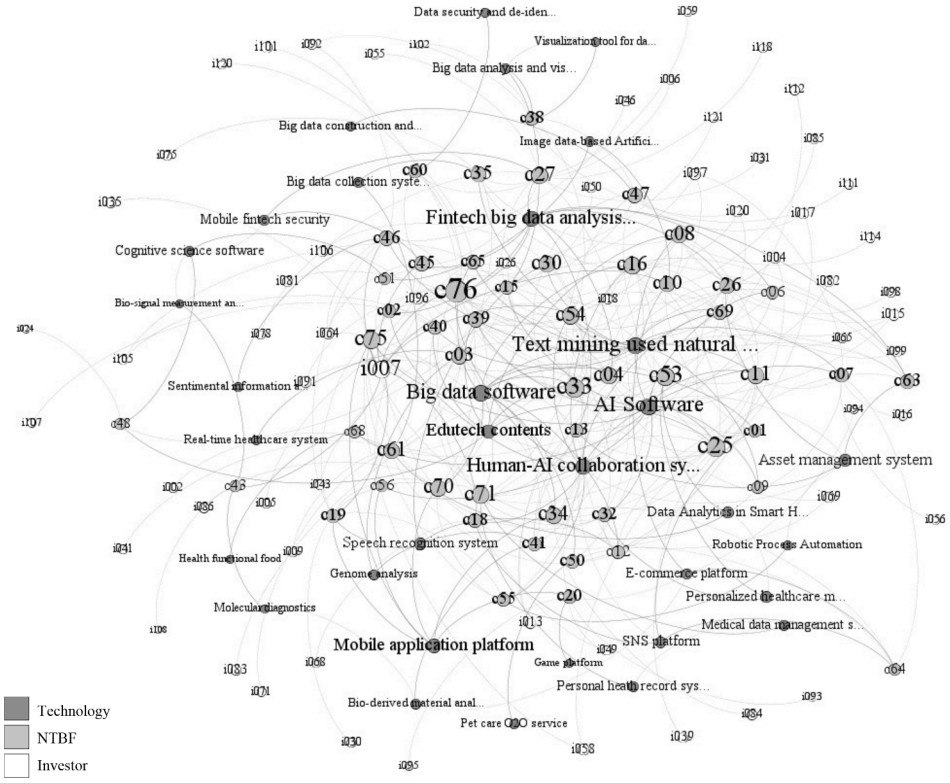

Fig. C1. TOD-KG for AI/Big Data category

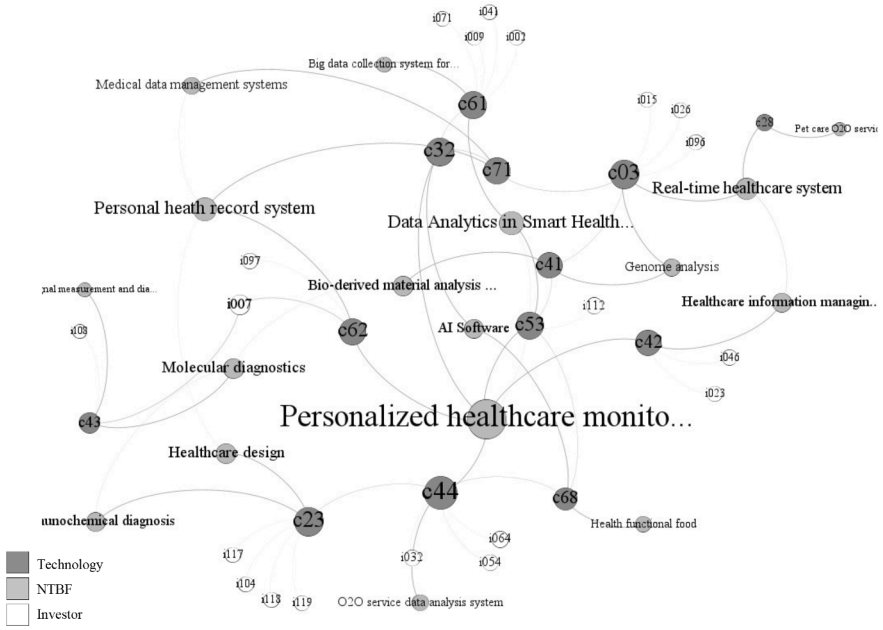

Fig. C2. TOD-KG for Bio/Healthcare category

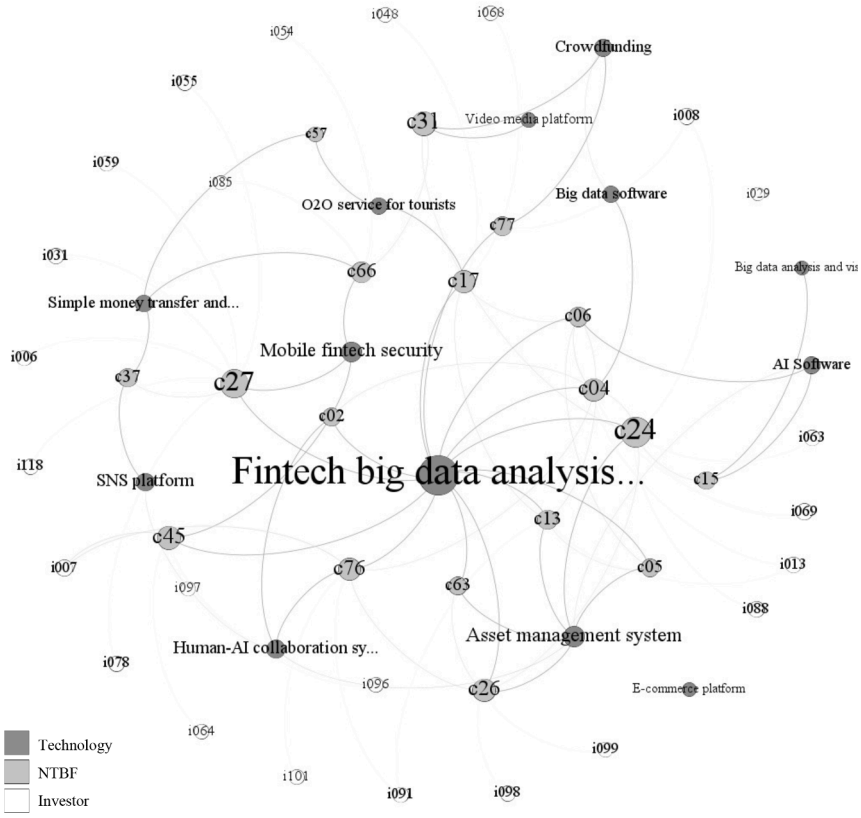

Fig. C3. TOD-KG for Fintech category

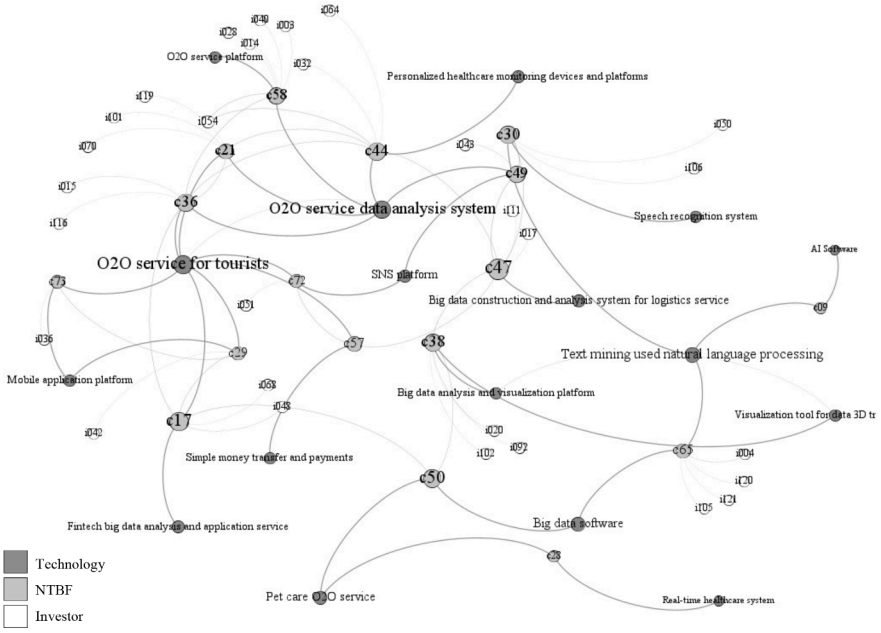

Fig. C4. TOD-KG for Offline-to-Online Service category

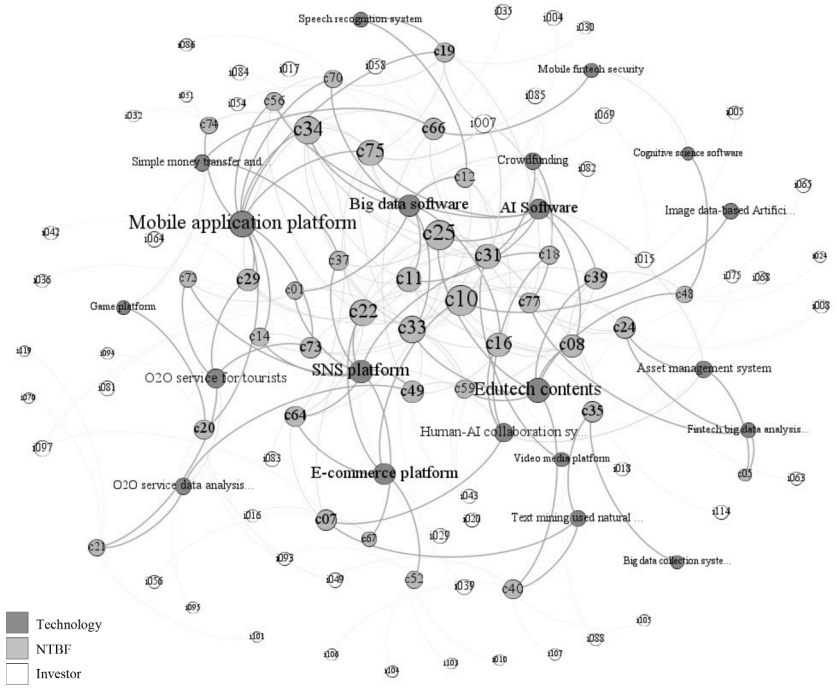

Fig. C5. TOD-KG for Service Platform category

Appendix D

Table D1
Scores of TI, NI, and II for 42 technologies

Technology	TI	NI	II
Big data software	0.408	0.397	0.285
Fintech big data analysis and application service	0.408	0.408	0.285
Text mining used natural language processing	0.395	0.388	0.290
Mobile application platform	0.393	0.373	0.278
AI Software	0.392	0.385	0.296
SNS platform	0.374	0.390	0.000
E-commerce platform	0.364	0.399	0.237
Human-AI collaboration system	0.364	0.386	0.339
Personalized healthcare monitoring devices and platforms	0.358	0.190	0.318
Edutech contents	0.356	0.394	0.266
Asset management system	0.343	0.403	0.287
O2O service data analysis system	0.336	0.368	0.273
O2O service for tourists	0.334	0.410	0.272
Mobile fintech security	0.330	0.319	0.338
Crowdfunding	0.327	0.213	0.000
Simple money transfer and payments	0.324	0.342	0.304
Speech recognition system	0.320	0.208	0.309
Data Analytics in Smart Healthcare	0.315	0.388	0.282
Image data-based Artificial Intelligence service	0.308	0.192	0.344
Pet care O2O service	0.302	0.173	0.000
Big data collection system for smart healthcare	0.294	0.201	0.000
Video media platform	0.293	0.235	0.000
Personal health record system	0.292	0.210	0.000
Healthcare information managing service	0.290	0.000	0.239
Healthcare design	0.289	0.169	0.258
Real-time healthcare system	0.289	0.196	0.000
Cognitive science software	0.285	0.166	0.000
Bio-derived material analysis system	0.283	0.217	0.000
Robotic Process Automation	0.283	0.181	0.000
Big data analysis and visualization platform	0.281	0.171	0.000
Genome analysis	0.281	0.222	0.296
Immunochemical diagnosis	0.276	0.169	0.000
Big data construction and analysis system for logistics service	0.275	0.184	0.290
Medical data management systems	0.273	0.210	0.000
Health functional food	0.262	0.166	0.000
Data security and de-identification	0.261	0.231	0.000
Game platform	0.261	0.174	0.000
Sentimental information analysis	0.259	0.202	0.000
O2O service platform	0.258	0.233	0.000
Visualization tool for data 3D transformation	0.257	0.161	0.278
Molecular diagnostics	0.248	0.000	0.281
Bio-signal measurement and diagnosis device	0.237	0.000	0.000

References

Bengisu, M., Nekhili, R., 2006. Forecasting emerging technologies with the aid of science and technology databases. Technol. Forecast. Soc. Change 73 (7), 835–844. https://doi.org/10.1016/j.techfore.2005.09.001.

Cho, C., Yoon, B., Coh, B.Y., Lee, S., 2016. An empirical analysis on purposes, drivers and activities of technology opportunity discovery: The case of Korean SMEs in the manufacturing sector. R. D. Manag. 46 (1), 13–35. https://doi.org/10.1111/radm.12107.

Choi, J., Jeong, B., Yoon, J., 2019. Technology opportunity discovery under the dynamic change of focus technology fields: Application of sequential pattern mining to patent classifications. Technol. Forecast. Soc. Change 148. https://doi.org/10.1016/j.techfore.2019.119737.

Choi, J., Kim, K., 2019. Case Study of Global Convergence-based Fintech Innovations and Domestic Financial Regulation: Focusing on Start-up chosen by Forbes. Korea Bus. Rev. 23 (3), 69–97.

Choi, S., Park, H., Kang, D., Lee, J.Y., Kim, K., 2012. An SAO-based text mining approach to building a technology tree for technology planning. Expert Syst. Appl. 39 (13), 11443–11455. https://doi.org/10.1016/j.eswa.2012.04.014.

Colombo, M.G., D'Adda, D., Pirelli, L.H., 2016. The participation of new technology-based firms in EU-funded r&d partnerships: The role of venture capital. Research Policy 45 (2), 361–375. https://doi.org/10.1016/j.respol.2015.10.011.

Curran, C.S., Leker, J., 2011. Patent indicators for monitoring convergence - examples from NFF and ICT. Technol. Forecast. Soc. Change 78 (2), 256–273. https://doi.org/10.1016/j.techfore.2010.06.021.

Davenport, T., Kalakota, R., 2019. The potential for artificial intelligence in healthcare. Future Hosp. J. 6 (2), 94–98. https://doi.org/10.7861/futurehosp.6-2-94.

Du, Y., Tang, Y., 2014. Study on the Development of O2O E-commerce Platform of China from the Perspective of Offline Service Quality. Int. J. Bus. Soc. 5 (4).

Dushnitsky, G., Lenox, M.J., 2006. When does corporate venture capital investment create firm value? Journal of Business Venturing 21 (6), 753–772. https://doi.org/10.1016/j.jbusvent.2005.04.012.

Han, J.-H., Hwangbo, Y., 2020. Determinants of Accelerators' Investment. Asia-Pacific Journal of Business Venturing and Entrepreneurship 15 (1), 31–44.

Han, X., Zhu, D., Wang, X., Li, J., Qiao, Y., 2019. Technology Opportunity Analysis: Combining SAO Networks and Link Prediction. IEEE Trans. Eng. Manage. 1–11.

Heo, J.-y., 2020. A Study on the Importance and Priorities of the Investment Determinants of Startup Accelerators. Asia-Pacific Journal of Business Venturing and Entrepreneurship 15 (6), 27–42.

Huang, Y., Lee, J., Wang, S., Sun, J., Liu, H., Jiang, X., et al., 2018. Privacy-preserving predictive modeling:Harmonization of contextual embeddings from different sources. JMIR medical informatics 6, e9455.

Jang, P.-H., Kim, Y.-H., 2021. A Study on the Trend of Employment Effect and Employment Policy in the Digital Bio-healthcare Industry. J. Converg. Inf. Technol. 11 (1), 175–182.

Ji, S., Pan, S., Cambria, E., Marttinen, P., Philip, S.Y., 2021. A survey on knowledge graphs: Representation, acquisition, and applications. IEEE Transactions on Neural Networks and Learning Systems.

Joulin, A., Grave, E., Bojanowski, P., Douze, M., Jégou, H., Mikolov, T., 2016. Fasttext.zip: Compressing text classification models. arXiv preprint arXiv:1612.03651.

Ju, K.J., Lee, M.H., Yang, H.J., Ryu, D.J., 2016. Fintech, Financial Industry, and Policy Implications. Korean Journal of Financial Studies 45 (1), 145–170.

Kang, H.D., Nanda, V.K., Park, H.D., 2021. Technology spillovers and capital gains in corporate venture capital investments: evidence from the biopharmaceutical industry. Venture Capital 23 (2), 129–155.

Kim, K., Park, K., Lee, S., 2019. Investigating technology opportunities: the use of SAOx analysis. Scientometrics 118 (1), 45–70. https://doi.org/10.1007/s11192-018-2962-9.
Kim, M., Kim, N.S., Pyo, C.S., 2018. 4 th Industrial Revolution Driven by AI Service Platform. The Journal of Korean Institute of Communications and Information Sciences 43 (10), 1763–1769.
Kim, Y.-K., 2020. Consideration of legal policy for revitalization of healthcare. Legal Theory & Practice Review 8 (4), 175–220.
Kim, S., Park, H., Lee, J., 2020. Word2vec-based latent semantic analysis (w2v-lsa) for topic modeling: A study onblockchain technology trend analysis. Expert Systems with Applications 152, 113401.
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. https://doi.org/10.1016/0048-7333(93)00762-I.
Le, Q., Mikolov, T., 2014. Distributed representations of sentences and documents. 31st International Conference on Machine Learning, ICML 2014 4, 2931–2939.
Lee, C., Jeon, D., Ahn, J.M., Kwon, O., 2020. Navigating a product landscape for technology opportunity analysis: A word2vec approach using an integrated patent-product database. Technovation 96-97. https://doi.org/10.1016/j.technovation.2020.102140.
Lee, C., Lee, G., 2019. Technology opportunity analysis based on recombinant search: patent landscape analysis for idea generation. Scientometrics 121 (2), 603–632. https://doi.org/10.1007/s11192-019-03224-7.
Lee, I., Shin, Y.J., 2018. Fintech: Ecosystem, business models, investment decisions, and challenges. Bus. Horiz. 61 (1), 35–46. https://doi.org/10.1016/j.bushor.2017.09.003.
Lee, J., Kim, C., Shin, J., 2017. Technology opportunity discovery to R&D planning: Key technological performance analysis. Technol. Forecast. Soc. Change 119, 53–63. https://doi.org/10.1016/j.techfore.2017.03.011.
Lee, S., Park, G., Yoon, B., Park, J., 2010. Open innovation in SMEs-An intermediated network model. Res. Policy 39 (2), 290–300. https://doi.org/10.1016/j.respol.2009.12.009.
Lee, J., Sun, J., Wang, F., Wang, S., Jun, C.H., Jiang, X., 2018. Privacy-preserving patient similarity learning in afederated environment: development and analysis. JMIR medical informatics 6, e7744.
Lee, S., Yoon, B., Park, Y., 2009. An approach to discovering new technology opportunities: Keyword-based patent map approach. Technovation 29 (6-7), 481–497. https://doi.org/10.1016/j.technovation.2008.10.006.
Lee, Y., Kim, S.Y., Song, I., Park, Y., Shin, J., 2014. Technology opportunity identification customized to the technological capability of SMEs through two-stage patent analysis. Scientometrics 100 (1), 227–244. https://doi.org/10.1007/s11192-013-1216-0.
Li, W., 1992. Random texts exhibit zipf`s-law-like word frequency distribution. IEEE Transactions on Information Theory 38 (6), 1842–1845. https://doi.org/10.1109/18.165464.
Mikolov, T., Chen, K., Corrado, G., Dean, J., 2013. Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781.
Moghaddam, A.G., Kashkoueyeh, M.S., Talezadeh, M., Aala, M., Ebrahimpour, M., Tehranypour, M., 2015. The Impact of Capital Structure on Corporate Performance. Int. J. Acad. Res. Bus. Soc. Sci. 5 (3) https://doi.org/10.6007/ijarbss/v5-i3/1535.
Nieto, M., Quevedo, P., 2005. Absorptive capacity, technological opportunity, knowledge spillovers, and innovative effort. Technovation 25, 1141–1157. https://doi.org/10.1016/j.technovation.2004.05.001.
Olsson, O., 2005. Technological opportunity and growth. J. Econ. Growth 10 (1), 35–57. https://doi.org/10.1007/s10887-005-1112-4.
Park, Y., Yoon, J., 2017. Application technology opportunity discovery from technology portfolios: Use of patent classification and collaborative filtering. Technol. Forecast. Soc. Change 118, 170–183. https://doi.org/10.1016/j.techfore.2017.02.018.
Pennington, J., Socher, R., Manning, C.D., 2014. GloVe: Global vectors for word representation. 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP 2014, Proceedings of the Conference 1532–1543.
Porter, A.L., Detampel, M.J., 1995. Technology opportunities analysis. Technol. Forecast. Soc. Change 49 (3), 237–255. https://doi.org/10.1016/0040-1625(95)00022-3.
Powers, D.M.W., 1998. Applications and explanations of zipf`s law. New methods in language processing and computational natural language learning.
Roh, T., Jeong, Y., Jang, H., Yoon, B., 2019. Technology opportunity discovery by structuring user needs based on natural language processing and machine learning. PLoS ONE 14 (10). https://doi.org/10.1371/journal.pone.0223404.
Rocketpunch, Rocketpunch by Alicon, 2019, https://www.rocketpunch.com/.
Ryu, D.-H., Lim, C., Kim, K.-J., 2020. Development of a service blueprint for the online-to-offline integration in service. J. Retail. Consum. Serv. 54, 101944. https://doi.org/10.1016/j.jretconser.2019.101944.
Saura, J.R., Palos-Sanchez, P., Grilo, A., 2019. Detecting indicators for startup business success: Sentiment analysis using text data mining. Sustainability (Switzerland) 11 (3). https://doi.org/10.3390/su11030917.
Smeroadmap, SME Roadmap by Korea Technology and Information Promotion Agency for SMEs, 2019, http://smroadmap.smtech.go.kr/.
Suh, M.S., Kim, D.H., 2019. A Study on the Changing Direction of FinTech Service Model based on Big Data. The e-Business Studies 20 (2), 195–213.
Tegarden, L.F., Lamb, W.B., Hatfield, D.E., Ji, F.X., 2012. Bringing emerging technologies to market: Does academic research promote commercial exploration and exploitation? IEEE Transactions on Engineering Management 59 (4), 598–608. https://doi.org/10.1109/TEM.2011.2170690.
Von Wartburg, I., Teichert, T., Rost, K., 2005. Inventive progress measured by multi-stage patent citation analysis. Res. Policy 34 (10), 1591–1607. https://doi.org/10.1016/j.respol.2005.08.001.
Wang, J., Chen, Y.J., 2019. A novelty detection patent mining approach for analyzing technological opportunities. Adv. Eng. Inform. 42 https://doi.org/10.1016/j.aei.2019.100941.
Xin, L., Jiwu, W., Lucheng, H., Jiang, L., Jian, L., 2010. Empirical research on the technology opportunities analysis based on morphology analysis and conjoint analysis. Foresight 12 (2), 66–76. https://doi.org/10.1108/14636681011035753.
Yang, C., Xiao, Y., Zhang, Y., Sun, Y., Han, J., 2020. Heterogeneous network representation learning: A unified framework with survey and benchmark. IEEE Transactions on Knowledge and Data Engineering.
Yoon, B., Park, I., youl 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. https://doi.org/10.1016/j.techfore.2013.10.013.
Yoon, B., Park, Y., 2004. A text-mining-based patent network: Analytical tool for high-technology trend. J. High Technol. Manag. Res. 15 (1), 37–50. https://doi.org/10.1016/j.hitech.2003.09.003.
Yoon, B., Park, Y., 2005. A systematic approach for identifying technology opportunities: Keyword-based morphology analysis. Technol. Forecast. Soc. Change 72 (2), 145–160. https://doi.org/10.1016/j.techfore.2004.08.011.
Yoon, J., Park, H., Seo, W., Lee, J.M., youl Coh, B., Kim, J., 2015. Technology opportunity discovery (TOD) from existing technologies and products: A function-based TOD framework. Technol. Forecast. Soc. Change 100. https://doi.org/10.1016/j.techfore.2015.04.012.

MyoungHoon Lee is a Ph.D candidate at Department of Industrial Engineering, UNIST, Ulsan, Republic of Korea. He received the B.A degree at the Department of Public Administration and Economics, KyungHee University, Seoul and the M.S. degree at Department of Business Analytics, Graduate School of Interdisciplinary Management, UNIST, Republic of Korea. His research interests include machine learning, deep learning, natural language processing, text mining, recommendation system, and federated learning.

Suhyeon Kim is a Ph.D. candidate at Department of Industrial Engineering, UNIST, Ulsan, Republic of Korea. She received the B.S. degree at the Department of Statistics, Pusan National University, Pusan and the M.S. degree at Department of Business Analytics, Graduate School of Interdisciplinary Management, UNIST, Republic of Korea. Her research interests include machine learning, deep learning, text mining, graph representation learning, bio-informatics, and federated learning.

Hangyeol Kim is a researcher at SK C&C, Seoul, Republic of Korea. She received the B.S. degree at the Department of Statistics, Inha University, Incheon and the M.S. degree at Department of Business Analytics, Graduate School of Interdisciplinary Management, UNIST, Republic of Korea. Her research interests include machine learning, deep learning, amd text mining.

Junghye Lee is an associate professor at Department of Industrial Engineering & Graduate School of Artificial Intelligence, UNIST, Ulsan, Republic of Korea. She received the B.S. and Ph.D. degrees at the Department of Industrial and Management Engineering, Pohang University of Science and Technology, Pohang, Republic of Korea. She worked as a post-doctoral researcher with the Biomedical Informatics Department, UCSD, USA. Her main interests include machine learning and deep learning with applications, especially deep representation learning-based predictive modeling and secure and privacy-preserving federated learning.