Job forecasting based on the patent information: A word embedding-based approach

Taehyun Ha¹, Mingook Lee¹, Bitnari Yun², Byong-Youl Coh¹
¹Future Technology Analysis Center, Korea Institute of Science and Technology Information, Seoul, Republic of Korea
²Center for R&D Investment and Strategy Research, Korea Institute of Science and Technology Information, Seoul, Republic of Korea

Corresponding author: Byong-Youl Coh (e-mail: cohby@kisti.re.kr).

ABSTRACT The rapid change in technology makes it challenging to forecast the future of jobs. Previous studies have analyzed economics and employment data or employed expert-based methods to forecast the future of jobs, but these approaches were not able to reflect the latest technology trends in an objective way. To overcome the issue, this study matches jobs with patents and forecasts the future of jobs based on changes in the number of patents with time. A word embedding model is trained by patent classification code and job description data and used to find similar patent classification codes of jobs. For an illustration purpose, we identify information technology-related jobs listed in O*NET and discover similar patent classification codes of the jobs. Based on the change in the number of patents, we find promising jobs presenting high technical demands. Several implications of our approach are also discussed.

INDEX TERMS Job attributes, Job forecasting, Patent classification code, Word embedding

I. INTRODUCTION
An advance in technology makes considerable changes in the job market. To occupy a dominant position in business fields, executives need to predict a change of jobs and preemptively address a new business model. For the last decades, several companies such as Google, Microsoft, Amazon, and Facebook have developed information technologies and hired many professional experts to maintain and advance the technologies. Following the significant change, many companies modified their initial business models and tried to respond to the change quickly and adequately. On the other hand, some other companies underestimated the technological advances and failed to address the change sufficiently. Consequently, the failure of the prediction to the technology has led them to lose their status. In other words, this suggests that the proper projection of future change significantly determines the success of business models.

Researchers have considered various approaches to predict a change of jobs. They considered economics and employment data to identify quantitative features of the job market and drew prediction reports for the changes in the market. This way was based on statistical data and thus adequate to predict the near future of the job market. However, this approach was limited in prediction because technological changes in the world were not considered enough. Thus, other researchers have employed expert-based approaches such as Delphi and structured survey methods. These approaches secured experts in the target fields and analyzed their opinions in a systemic way. In this way, researchers tried to render more reliable and practical predictions for the job market. However, because the approach mainly relied on experts, the prediction results could be biased depending on the experts’ opinions, beliefs, and convictions.

Although the economics and employment data does not contain information about technologies, another type of data, patents, provides meaningful information about how technology will change future jobs. Patents contain the contents of technology and the right to use it. Thus, people try to register patents as soon as possible to secure their rights for the technology and occupy a dominant position. Because of this property, patents have been utilized in technology forecasting to identify the latest technology and its future. In other words, if researchers can calculate similarities between jobs and patents, they can track and
forecast the change of a particular job through the related patents. For example, we can argue that the importance of a specific job has increased if the number of patents with a specific patent classification code shows an increasing pattern and the classification code is closely related to the job.

Focusing on this issue, this study reviews existing methods for calculating similarities between jobs and patents and suggests a word embedding-based approach to job forecasting. We collect descriptions of job attributes from the website of occupational information network (O*NET) and descriptions of patent classification codes from the website of cooperative patent classification (CPC). Based on the descriptions, we train a word embedding model and extract embedding vectors for the descriptions. Then, we find the most similar patent classification code for each job and create a matching table.

The results show that the suggested approach can match job information with patent information better than existing methods. Also, for an illustration purpose, we analyze information technology-related jobs and discuss the analysis results in detail. Our approach does not solely rely on people’s subjective viewpoints on jobs but provides a forecast based on the change in the number of patents. This enables people to catch future trends in jobs objectively and helps draw a plausible prediction on jobs.

II. BACKGROUND

A. JOB FORECASTING

Predicting the change in the job market has been an important issue in politics and economics research. In the past, researchers argued that the rapid advance in high technologies had a limited impact on the job market because only a few percentages of the job market were related to the high technology, and the employment growth would be small [1]. However, this belief has changed with the fourth industrial revolution. Technologies of robots and artificial intelligence (AI) are expected to replace simple labor and make workplaces more skill-polarized [2, 3]. Researchers tried to predict the future of the job market based on numerical data (e.g., labor employment [4] and statistics of graduates [5]) or textual data (e.g., job postings [6, 7], job attributes [8], and Google trends [9]). However, these approaches were mainly focused on occupational data and limitedly considered changes in technology.

Several research institutes and firms have utilized Delphi and structured survey methods to forecast the future of the job market, considering the technological aspect. For example, the Millennium Project has conducted Delphi studies for the future of work and released reports based on the findings [e.g., 10]. Similarly, the World Economic Forum has gathered experts’ opinions through structured surveys and announced reports for future jobs [e.g., 11]. These studies focused on a macro perspective on the job market. However, the methods also have been utilized for forecasting specific jobs in detail, such as hotel management [12], marketing [13], and auditing [14] or specific fields in detail such as AI [15]. Researchers tried to forecast future jobs based on experts’ opinions in various fields rather than simply predict future trends.

On the other hand, some studies suggested a hybrid approach combining experts’ opinions and job-related data. For instance, Bakhshi, et al. [16] gathered experts’ opinions for jobs and trained a machine learning model based on the opinions, job description data from O*NET, and employment data from the bureau of labor statistics (United States) and the office for national statistics (UK). Then, they predicted future demands of jobs utilizing the model and anticipated new occupations. However, similar to Delphi and focused group interview techniques, the expert-based approach is limited because the experts’ views on future jobs can be largely different even though they share a similar background. For example, Gruetzemacher, et al. [17] surveyed attendees of AI conferences and found that their forecasts on labor-displacing AI scenarios were significantly different depending on what conferences they attended. This suggests that the forecast of future jobs should be based on quantitative data that objectively shows the change of technology.

B. PATENT ANALYSIS

State-of-the-art technologies are often disclosed to the public through academic articles or patent documents. Because of the different characteristics of the sources, it is hard to determine which one is more useful to capture the trend of technology. In general, academic articles are published after a peer-review, and this process is helpful when researchers need to verify and improve their idea. However, the academic article limitedly protects researchers’ right to the idea because it does not have legal force. In contrast, patent documents clarify applicants’ right to the idea and legally protect it. For this issue, examiners of patents mainly focus on the distinction and exclusiveness of the idea rather than evaluating the idea itself. In the absence of the peer-review process, patents may not promptly reflect the advance of technologies compared to academic articles. However, patents are often preferred in business fields because they have legal force to claim their exclusive use. In addition, because there is a periodic cost to renew a patent, the patent is likely to contain practical technologies compared to the academic article.

Focusing on the characteristics of patents, researchers have analyzed patents to forecast future technologies. Kim and Bae [18] clustered patents based on the patent classification code and assessed the clustering groups based on the top 10 classification codes. Promising patents in the groups were identified by the forward citation, triadic patents, and independent claims. Cho, et al. [19] identified keywords related to high-rise building construction and secured 2875
patents in the United States, Europe, Korea, China, and Japan. Through analyzing the growth of numbers of patents in each technology field, they forecasted promising technologies and diagnosed the levels of technologies in each country. Kyebambe, et al. [20] considered that new classes added to the existing patent classification category indicate emerging technologies and utilized the classes to develop a supervised labeling method to forecast emerging technologies. Zhang, et al. [21] suggested that a topic modeling method can capture a technology roadmap. They divided patents into three groups based on their registration years and constructed topic models with the divided patent data. Then they calculated similarities of topics in different models and suggested what major topics were considered in the field of blockchain and how they were evolved with time.

Although patents were frequently utilized to forecast future technologies in specific fields, only a few researchers used patents to forecast future jobs. Dechezleprêtre, et al. [22] analyzed the impacts of a particular technology on the labor market based on changes in the number of patents and wages. However, they focused on retrospective analysis with a macroscopic perspective and did not provide certain links between jobs and patents. Webb [23] employed a dependency parsing algorithm to extract verb-noun pairs from task descriptions of jobs and titles of patents and calculated similarity scores between jobs and patents. The study suggested a patent-based approach to investigate changes in the job market. However, the accuracy of matching jobs and patents could be improved further by utilizing a more sophisticated natural language processing model.

Recent studies showed some text embedding models such as Word2Vec [24], Doc2Vec [25], FastText [26], and BERT [27] were able to learn a large size of texts and embed them as numerical vectors. This suggests that a text embedding model trained by descriptions of patents and jobs can transform the descriptions as embedding vectors, and similarities between jobs and patents can be calculated based on the vectors. Analyzing changes in patents and linking patents with similar jobs, we can forecast how jobs would be changed following changes in technology.

III. METHOD
As reviewed, job forecasting has been relied on occupational

![FIGURE 1. Overall process of the job forecasting framework](image_url)
data or experts’ knowledge but not on data reflecting technological changes. Previous studies showed that patents could be a reliable source to catch technological changes in specific fields. However, patents were not used for job forecasting because there was no method to bridge the job and patent information. To solve the problem, we need to examine several text embedding models for matching patents with jobs and find an effective one. Once the model is found, we can forecast a particular job’s future based on the change in the number of patents related to the job. Developing this analogical process will contribute to drawing a more plausible job forecasting.

Concerning the issue, this study suggests a framework to investigate job futures based on patent classification codes. To match jobs with similar patent classification codes, we trained a word embedding model with descriptions of job attributes and patent classification codes and extracted embedding vectors from the descriptions. The extracted vectors were used to calculate similarities of descriptions, and jobs were matched with similar patent classification codes. After the matching process, job futures were forecasted based on the change of patent classification codes with time. FIGURE 1 shows the overall process of the job forecasting framework.

A. TRAINING A WORD EMBEDDING MODEL
To calculate similarities between jobs and patents, we secured data for descriptions of job attributes from the website of O*NET (www.onetonline.org) and patent classification codes from the website of CPC (www.cooperativepatentclassification.org). The O*NET data contains information about 1016 jobs, including job description, knowledge, skill, ability, task, technology skill, tool, and work activity, and has been used for forecasting future jobs [e.g., 16, 28]. A patent document is composed of multiple components. Previous studies have focused on the title, abstract, claims, and classification codes of patents to capture the contents of patents [e.g., 29]. The job attributes...
and patent components can be used to train a text embedding model. Similarities for every pair of job attributes and patent components can be calculated based on the embedding vectors. However, this approach would not be practical because the calculation cost is prohibitive. Thus, we only considered descriptions of patent classification codes rather than all components in every single patent. In this way, we trained a text embedding model with descriptions of job attributes and patent classification codes and extracted embedding vectors for the descriptions.

In general, document embedding models are expected to show better performance than word embedding models because they can reflect contextual information in texts more effectively. However, because the length of texts and amount of the data used in this study are short and small, we considered word and document embedding models together and examined their performances. Among various models, Word2Vec, Doc2Vec, FastText, and BERT are the most popular models used in previous studies and we trained the model through unsupervised learning with the descriptions of job attributes and patent classification codes. Because BERT cannot be trained by an unsupervised learning method (i.e., fine-tuning), we employed the pre-trained BERT-base for comparison. All parameters of the models were set to the default values.

Before the training process, we preprocessed the description texts to facilitate the model learning the descriptions effectively. In detail, we tokenized each text using the WordNet tokenizer and made tokenized words. Then, we lemmatized the texts using the WordNet lemmatizer and removed stopwords using NLTK (Natural Language Toolkit) packages in Python. The lemmatization was conducted considering words’ part-of-speech and the part-of-speech was identified by NLTK part-of-speech tagger.

B. VALIDATING THE TRAINED MODEL

Because the model was trained by an unsupervised learning method, we additionally devised classification tasks to
measure the model’s performance. O*NET divided 1,011 jobs into 31 classes and named the classes as career clusters (Although the total number of jobs was 1,016, O*NET only clustered 1,011 of them). Similarly, the CPC system has a hierarchical architecture, and classification codes can be categorized into 136 classes. We extracted embedding vectors for the descriptions of job attributes and made job representation vectors for each job. Then, we clustered the job representation vectors into 31 groups and compared them with the career clusters. Likewise, we extracted vectors for patent classification codes, clustered them into 136 groups, and compared them with the classes in the CPC system. To examine a proper form of the job representation vector, we made four different types of job representation vectors extracted from (1) job title, (2) job description, (3) job title and description, and (4) job title and attributes, and examined which one showed the best performance in the validation process. K-means algorithm and normalized mutual information score were used for clustering and comparing the clustered groups, respectively. FIGURE 2 and FIGURE 3 show the validation process in detail.

C. MATCHING AND FORECASTING JOBS
We calculated cosine similarities between the job representation vectors and vectors for patent classification codes and found similar patent classification codes for each job. Different results can be gained in the calculation process depending on how deep patent classification codes are considered for comparison (e.g., full codes or a certain level of codes). Thus, we examined similarities between jobs and patent classification codes with different levels of patent classification codes. Based on the result of similarity calculation, jobs were matched with similar patent classification codes, and their futures were forecasted based on the change of matched codes with time.

IV. RESULT
A. MODEL TRAIN AND VALIDATION
We gathered 231,097 descriptions of job attributes for 1,016 jobs (job description: 1,016; knowledge: 28,809; skill: 30,555; ability 45,396; task: 17,851; technology skill: 30,139; tool: 41,538; work activity: 35,793). All jobs were described in terms of job title and description, but not all job attributes were assigned to all jobs (i.e., some jobs were only described in terms of job title and description). In the case of patent classification codes, we were able to secure 260,289 descriptions of CPC codes. However, because too short CPC codes covered a broad range of technology, we only considered 260,144 CPC codes longer than four digits (i.e., subclass). A FastText model was constructed using Gensim (radimrehurek.com/gensim/models/fasttext.html), a Python library package for natural language processing, and trained by our job and patent description dataset. The number of iterations (epoch) was set to 1000, and other parameters were set to default values.

Descriptions of job attributes for each job were vectorized by the trained embedding model and used to construct four different types of job representation vector (job title, job description, job title and description, and job title and attributes). We applied the K-means algorithm for each type of job representation vector and clustered the vectors into 31 groups. The clusters were compared with 31 career clusters developed by O*NET and evaluated based on the normalized mutual information.

We examined four different types of models and their performances are presented in TABLE 1. As presented in the table, FastText model showed the best performance for the validation tasks. Normalized mutual information scores for job representation vectors made by (1) job title, (2) job description, (3) job title and description, and (4) job title and attributes were presented by 0.313, 0.415, 0.424, and 0.264, and we confirmed that the job title and description was the most effective attribute representing the job information. Descriptions of CPC codes were also vectorized by the trained model (CPC description vector) and clustered into 136 groups. The clusters were compared with 136 classes of the CPC system, and normalized mutual information score was presented by 0.445. Approximately, normalized mutual information score closed to 0.4 indicates 0.84 true positive rates [30], and we confirmed that the trained embedding model adequately represented job and patent information. FIGURE 4 shows the results of job and patent vector clustering.

B. MATCHING AND FORECASTING JOBS: A CASE OF INFORMATION TECHNOLOGY-RELATED JOBS
We calculated cosine similarities between job representation vectors and CPC description vectors and made a matching table for 1,016 jobs and 260,144 CPC codes. To forecast job futures, we traced changes in technology during the last five years based on the number of patents for each CPC code. However, considering all CPC codes is inappropriate to identify changes in technology because new CPC codes are

| Model  | Job title | Job description | Job title + description | Job title + all attributes | Patent information |
|--------|-----------|-----------------|------------------------|---------------------------|-------------------|
| Word2Vec | 0.257     | 0.374           | 0.347                  | 0.297                     | 0.430             |
| Doc2Vec | 0.297     | 0.415           | 0.382                  | 0.319                     | 0.276             |
| FastText | 0.313     | 0.415           | **0.424**              | 0.264                     | **0.445**         |
| BERT-base | 0.182     | 0.196           | 0.211                  | 0.145                     | 0.146             |
added in the regular update of the CPC system, and these codes do not appear before the updates. For example, A01K3/001 code was added in August 2021, and this code did not appear before that time. If someone matches specific jobs with CPC codes and the matched codes include A01K3/001 code, forecasting the future of jobs would be inaccurate because the number of patents with A01K3/001 code would be zero before August 2021. Thus, we only considered valid CPC codes during the last five years and matched them with jobs. TABLE 2 shows a matching result for jobs in the career cluster of information technology developed by O*NET.

The table shows that G06 class CPC codes (computing, calculating, and counting) appeared most frequently. Some jobs presented CPC codes sharing the same CPC section (e.g., G06, Y04, or H04), but others did not. This implies that a single CPC code limitedly described multiple features of a job. Thus, we described a job with multiple CPC codes and utilized them to identify changes in technology related to the job. TABLE 3 shows the top 5 most similar full CPC codes for information technology-related jobs. Changes in technology related to the job were identified based on the

### Table II

| Job                              | Top 1 subclass CPC | Top 1 full CPC   |
|----------------------------------|-------------------|-----------------|
| Business Intelligence Analysts   | G06Q (0.924)      | Y10S707/936 (0.946) |
| Computer Occupations, All Other  | G05B (0.867)      | B42D15/0066 (0.894) |
| Computer Systems Engineers/Architects | Y04S (0.968)  | Y04S04/20 (0.978) |
| Computer User Support Specialists | G06Q (0.956)     | H04M2201/32 (0.978) |
| Data Warehousing Specialists     | G06Q (0.933)      | G06F21/6245 (0.944) |
| Document Management Specialists  | G05B (0.950)      | G06F21/805 (0.968) |
| Geographic Information Systems Technologists and Technicians | G06Q (0.945) | G06F19/36 (0.953) |
| Information Technology Project Managers | G06Q (0.930) | G06Q100063118 (0.946) |
| Project Management Specialists   | G06Q (0.928)      | G06F090/20 (0.936) |
| Search Marketing Strategists     | G06Q (0.943)      | G06F090/20 (0.952) |
| Software Quality Assurance Analysts and Testers | G06N (0.946) | G06F8/75 (0.963) |
| Video Game Designers             | B44F (0.939)      | A63F2009/2486 (0.962) |
| Web Administrators               | G06Q (0.926)      | G06F8/76 (0.942) |
| Web and Digital Interface Designers | G06N (0.933) | H04L41/142 (0.944) |
| Computer Network Architects      | G06N (0.929)      | H04L41/14 (0.967) |
| Computer Network Support Specialists | H04W (0.909) | H04L12/40143 (0.969) |
| Database Administrators          | G06N (0.917)      | Y10S707/936 (0.963) |
| Database Architects              | G06N (0.943)      | Y10S707/936 (0.959) |
| Information Security Analysts    | Y04S (0.936)      | G06F21/577 (0.960) |
| Network and Computer Systems Administrators | G06N (0.928) | H04L63/0876 (0.965) |
| Telecommunications Engineering Specialists | G06Q (0.937) | H04W4/20 (0.970) |
| Computer Programmers             | G01M (0.947)      | G06K15/4005 (0.966) |
| Computer Systems Analysts        | G06Q (0.956)      | G06P90/20 (0.963) |
| Health Informatics Specialists    | G16H (0.922)      | A61B2505/01 (0.927) |
| Software Developers              | G06Q (0.945)      | G06F8/20 (0.963) |
| Web Developers                   | G16Z (0.932)      | H04L41/0253 (0.944) |

Numbers in parentheses indicate cosine similarity; Top 1 subclass CPC and full CPC represent the most similar subclass and full CPC code, respectively.
number of international patents published in the last five years (2016 to 2020). For each information technology-related job, we searched patents for the matched top five CPC codes using LexisNexis PatentSight and summed numbers of patents for each year. TABLE 3 and FIGURE 5 show the numbers of patents during the last five years.

As shown in TABLE 4 and FIGURE 5, information security analysts showed the highest number of patents in 2020 (14039), followed by computer network architects (5224), data warehousing specialists (4789), network and computer systems administrators (4701), computer network support specialists (2963). Information security analysts presented the steepest rise in the five years (annual average growth of 3560.4), followed by computer network architects (5224), data warehousing specialists (4789), network and computer systems administrators (4701), computer network support specialists (2963). In contrast, database administrators presented the lowest number of patents in


tabular data

| Job                                           | Top 1 code          | Top 2 code          | Top 3 code          | Top 4 code          | Top 5 code          |
|-----------------------------------------------|---------------------|---------------------|---------------------|---------------------|---------------------|
| Business intelligence analysts               | Y105707/936 (0.946) | Y105707/935 (0.944) | Y092237/20 (0.943)  | Y092237/30 (0.942)  | Y1053707/934 (0.940) |
| Computer occupations, all other              | B42D115/0066 (0.894) | G05B2219/321 (0.976) | H04Q2213/05 (0.988) | H04Q2213/55 (0.988) | G05B2219/374 (0.988) |
| Computer systems engineers/architects       | Y04540/20 (0.978)   | Y04540/00 (0.976)   | Y04540/128 (0.976)  | Y04540/00 (0.976)   | Y04540/00 (0.976)   |
| Computer user support specialists            | H04M2201/32 (0.978) | H04M2201/42 (0.973) | H04M2201/00 (0.973) | H04M2201/38 (0.972) | H04M2201/54 (0.972) |
| Data warehousing specialists                 | G06F21/6245 (0.944) | G06F21/6227 (0.944) | G06F21/6236 (0.943) | G06F21/6204 (0.943) | G06F21/6204 (0.943) |
| Document management specialists              | G06F21/805 (0.968)  | G06F21/80 (0.968)   | G06F21/79 (0.967)   | G06F21/79 (0.967)   | G06F21/79 (0.967)   |
| Geographic information systems technicians   | G06F19/36 (0.953)   | G06F90/205 (0.950)  | G06F90/205 (0.950)  | G06F90/205 (0.950)  | G06F90/205 (0.950)  |
| Information technology project managers      | G06Q10/6311 (0.946) | G06Q10/6311 (0.944) | G06Q10/6311 (0.943) | G06Q10/6311 (0.943) | G06Q10/6311 (0.943) |
| Project management specialists               | G06Q90/20 (0.936)   | G06Q90/20 (0.935)   | G06Q90/20 (0.934)   | G06Q90/20 (0.934)   | G06Q90/20 (0.934)   |
| Search marketing strategists                 | G06F8/57 (0.952)    | G06F8/10 (0.963)    | G06F8/10 (0.963)    | G06F8/10 (0.963)    | G06F8/10 (0.963)    |
| Software quality assurance analysts and testers| G06F8/76 (0.942)   | H04L1/42/09 (0.944) | H04L1/42/09 (0.940) | H04L1/42/09 (0.940) | H04L1/42/09 (0.940) |
| Web administrators                           | H04L1/42/14 (0.944) | H04L1/42/14 (0.942) | H04L1/42/14 (0.942) | H04L1/42/14 (0.942) | H04L1/42/14 (0.942) |
| Web and digital interface designers          | H04L1/42/14 (0.967) | H04L1/42/14 (0.967) | H04L1/42/14 (0.967) | H04L1/42/14 (0.967) | H04L1/42/14 (0.967) |
| Computer network architects                  | H04L1/42/14 (0.967) | H04L1/42/14 (0.967) | H04L1/42/14 (0.967) | H04L1/42/14 (0.967) | H04L1/42/14 (0.967) |
| Computer network support specialists          | H04L1/42/14 (0.967) | H04L1/42/14 (0.967) | H04L1/42/14 (0.967) | H04L1/42/14 (0.967) | H04L1/42/14 (0.967) |
| Database administrators                      | Y105707/936 (0.963) | Y105707/947 (0.960) | Y105707/947 (0.958) | Y105707/947 (0.956) | Y105707/947 (0.956) |
| Database architects                          | Y105707/936 (0.963) | Y105707/947 (0.958) | Y105707/947 (0.956) | Y105707/947 (0.954) | Y105707/947 (0.954) |
| Information security analysts                | G06F21/577/0.960    | G06F21/577/0.960    | G06F21/577/0.960    | G06F21/577/0.960    | G06F21/577/0.960    |
| Network and computer systems administrators  | H04L1/42/0876 (0.966) | H04L1/42/0876 (0.967) | H04L1/42/0876 (0.967) | H04L1/42/0876 (0.967) | H04L1/42/0876 (0.967) |
| Telecommunications engineering specialists    | H04M4/20 (0.970)    | H04M4/20 (0.969)    | H04M4/20 (0.969)    | H04M4/20 (0.969)    | H04M4/20 (0.969)    |
| Computer programmers                         | G06K15/4005 (0.966) | G06K15/4005 (0.965) | G06K15/4005 (0.965) | G06K15/4005 (0.965) | G06K15/4005 (0.965) |
| Computer systems analysts                    | G06Q90/20 (0.963)   | G06Q90/20 (0.963)   | G06Q90/20 (0.963)   | G06Q90/20 (0.963)   | G06Q90/20 (0.963)   |
| Health informatics specialists               | A61B2505/05 (0.927) | A61B2505/05 (0.919) | A61B2505/05 (0.917) | A61B2505/05 (0.916) | A61B2505/05 (0.916) |
| Software developers                          | G06F8/20 (0.963)    | G06F8/20 (0.962)    | G06F8/20 (0.962)    | G06F8/20 (0.962)    | G06F8/20 (0.962)    |
| Web developers                               | H04L1/42/066 (0.944) | H04L1/42/066 (0.943) | H04L1/42/066 (0.943) | H04L1/42/066 (0.943) | H04L1/42/066 (0.943) |

Numbers in parentheses indicate cosine similarity; Top 1 subclass CPC and full CPC represent the most similar subclass and full CPC code, respectively.
TABLE IV
NUMBERS OF JOB-RELATED PATENTS DURING 2016-2020

| Job                                                                 | Top 1 code | Top 2 code | Top 3 code | Top 4 code | Top 5 code |
|----------------------------------------------------------------------|------------|------------|------------|------------|------------|
| Business intelligence analysts                                      | 61         | 64         | 74         | 68         | 77         |
| Computer occupations, all other                                      | 3          | 2          | 8          | 6          | 7          |
| Computer systems engineers/architects                               | 253        | 309        | 389        | 428        | 490        |
| Computer user support specialists                                   | 70         | 154        | 308        | 149        | 151        |
| Data warehousing specialists                                         | 1030       | 1492       | 1982       | 3099       | 4789       |
| Document management specialists                                      | 208        | 253        | 315        | 463        | 557        |
| Geographic information systems technologists and technicians         | 359        | 435        | 17         | 113        | 46         |
| Information technology project managers                             | 340        | 960        | 1377       | 1998       | 2942       |
| Project management specialists                                       | 179        | 360        | 432        | 612        | 878        |
| Search marketing strategists                                         | 223        | 304        | 411        | 511        | 680        |
| Software quality assurance analysts and testers                      | 214        | 318        | 400        | 564        | 827        |
| Video game designers                                                | 22         | 25         | 72         | 99         | 73         |
| Web administrators                                                   | 138        | 149        | 209        | 294        | 447        |
| Web and digital interface designers                                 | 605        | 926        | 644        | 818        | 935        |
| Computer network architects                                          | 1608       | 2848       | 4261       | 4829       | 5224       |
| Computer network support specialists                                 | 1074       | 1808       | 2335       | 2510       | 2963       |
| Database administrators                                              | 1          | 0          | 0          | 0          | 0          |
| Database architects                                                  | 21         | 15         | 21         | 20         | 45         |
| Information security analysts                                        | 3763       | 5530       | 7186       | 10155      | 14039      |
| Network and computer systems administrators                          | 1644       | 2715       | 3893       | 4242       | 4701       |
| Telecommunications engineering specialists                            | 447        | 418        | 435        | 490        | 514        |
| Computer programmers                                                 | 47         | 51         | 26         | 26         | 14         |
| Computer systems analysts                                            | 274        | 511        | 603        | 908        | 1195       |
| Health informatics specialists                                       | 331        | 472        | 214        | 288        | 208        |
| Software developers                                                  | 432        | 817        | 1249       | 1352       | 1722       |
| Web developers                                                       | 204        | 287        | 355        | 367        | 390        |

FIGURE 5. Changes in the number of patents related to jobs

2020 (0), followed by computer occupations, all other (7), computer programmers (14), database architects (45), and geographic information systems technologists and technicians (46). Database administrators presented the lowest growth in the five years (annual average growth of 0.2), followed by computer occupations, all other (2.0), computer programmers (12.2), database architects (13.2), and video game designers (19.0).
Based on the results, we can say that people’s interest in information security analysts has rapidly grown for the last five years, and technical demand for information security analysts is expected to rise. People’s interests in computer network architects, data warehousing specialists, network and computer systems administrators, and computer network support specialists also have increased and technical demands for the jobs would be high. However, database administrators, minor computer occupations, computer programmers, database architects, and video game designers did not reveal notable increases in the number of patents. Compared to other jobs, these jobs were not receiving much attention, and technical demands for the jobs are expected to maintain the current status.

V. Discussion

We developed a word embedding model utilizing the O*NET and CPC data to match jobs with patents and forecast jobs from a technological perspective. To develop the model, we employed a FastText model and trained it with descriptions of job attributes and patent classification codes. Because the model was trained by an unsupervised learning method, we devised a validation method with existing classification systems (job clusters developed by O*NET and classes of the CPC system). The validation result showed that embedding vectors extracted by the trained model were consistent with the existing classification systems in terms of the normalized mutual information score, and we confirmed that the model properly reflected the job and patent information.

We examined four different types of job representation vectors and found that combining job title with description was the most effective attribute in presenting the job information. Job representation vectors summing all job attributes showed the worst performance among the four types of job representation vectors. We guess this was because not all job attributes were distributed to all jobs, and thus job information was not adequately reflected in the job representation vector. For instance, the O*NET data described hydrologic technicians only in terms of the job description, tasks, and detailed work activities, whereas business intelligence analysts were described in terms of the job description, knowledge, skills, abilities, tasks, technology skills, tools, and work activities. The unbalanced description of job attributes made the unbalanced reflection in job representation vectors and low performance in the validation process.

To specify the jobs in terms of patents, we identified valid CPC codes during the last five years (2016-2020) and matched them to the jobs. In the process, some jobs were matched with CPC codes that were less representative than other CPC codes. For example, health informatics specialists were matched with CPC codes belonging to A61B2505 group (evaluating, monitoring, or diagnosing in the context of a particular type of medical care). However, the latest CPC system has G16H subclass, which directly represents healthcare informatics. CPC codes belonging to G16H subclass can be regarded as closer to the job than CPC codes belonging to A61B2505 group. However, G16H subclass was not considered in this study because the codes belonging to the subclass were added in 2018. This shows that a period for analysis not only affects the number of patents at each time but also affects CPC codes for the analysis and yields a different result of forecasting jobs.

For an illustration purpose, we identified CPC codes related to information technology jobs and tracked the numbers of patents for the identified CPC codes for the last five years. Changes in the number of patents showed that people’s interests in information security analysts, computer network architects, data warehousing specialists, network and computer systems administrators, and computer network support specialists have highly increased, and technical demands for the jobs would be high. In contrast, we could not find significant increases in technical demands for database administrators, minor computer occupations, computer programmers, database architects, and video game designers. However, before accepting the forecast result, we need to care that there could be bias selection of patent classification codes in applying for a patent. For example, patent attorneys could rarely use some CPC codes because of implicit rules or customs. In this case, a job forecast would be invalid, although CPC codes are appropriately matched with jobs. Thus, it is recommended to accept a job forecast result comparing it with other data and analysis reports.

We considered that the increasing number of patents for a specific technology indicates the increasing demand for the technology, and the demand for jobs related to the technology will increase accordingly. However, someone can interpret it differently. As the number of patents for the technology increases, the supply of the technology increases, consequently making decreased demands for jobs related to the technology. It would be hard to confirm which one is the correct interpretation based on the pattern of the number of patents only. Thus, as mentioned above, researchers should consider other indicators such as statistics of employment, earnings, and graduates together to reach a reasonable decision.

In addition, although we tried to match various jobs with patents, some jobs were hard to describe in terms of patents and not adequately matched with patent classification codes. For example, actors and dancers were matched with F21V33/0056 (structural combinations of lighting devices with other articles, not otherwise provided for - personal or domestic articles - audio or video equipment, e.g., televisions, telephones, cameras or computers; remote control devices therefor - audio equipment, e.g., music instruments, radios or speakers) and H04N7/155 (television systems - systems for two way working - conference systems - involving storage of or access to video conference sessions), respectively. This suggests that not all jobs are proper to forecast in terms of
patents, and researchers should care about the appropriateness of forecasting based on the matching result of jobs and CPC codes.

VI. Conclusion

The present study suggested a patent-based approach to forecasting the future of jobs. To identify a relationship between patents and jobs, we secured CPC code description and job description data and trained a word embedding model with the data. Validation results confirmed that the trained model properly reflected the context in CPC code and job descriptions. For a demonstration purpose, information technology-related jobs were matched with CPC codes, and the future of jobs was forecasted based on the number of patents for CPC codes. The forecast results showed that our approach can be practical but also limited in some cases. We discussed several implications based on the results.

Future studies can consider the different job and patent data sources to improve job forecasting performance (e.g., online job posts and titles and abstracts of patents). Also, because our analytical approach mainly relies on patent data, securing reliable and up-to-date patent data would be critical. For example, Google patent search, which shows patent information in real-time, can be considered for building a data pipeline for our approach. A data pipeline with real-time data would enable an agile system for job forecasting. The results of the studies will confirm the validity and effectiveness of our approach.

REFERENCES

[1] R. W. Rumberger and H. M. Levin, "Forecasting the impact of new technologies on the future job market," Technological Forecasting and Social Change, vol. 27, no. 4, pp. 399-417, 1985.

[2] S. Chuang and C. M. Graham, "Embracing the sobering reality of technological influences on jobs, employment and human resource development: A systematic literature review," European Journal of Training and Development, 2018.

[3] G. Graetz and G. Michaels, "Robots at work," Review of Economics and Statistics, vol. 100, no. 5, pp. 753-768, 2018.

[4] A. Akaev, T. Devezas, Y. Ichkitidze, and A. Sarygulov, "Forecasting the labor intensity and labor income share for G7 countries in the digital age," Technological Forecasting and Social Change, vol. 167, p. 120675, 2021.

[5] I. Kashirina, Y. Bondarenko, and T. Azarnova, "Analysis and Forecasting of the Market of Educational Services of the Region," in 2021 1st International Conference on Technology Enhanced Learning in Higher Education (TEL), 2021, pp. 30-34: IEEE.

[6] N. Dawson, M.-A. Rizoiu, B. Johnstone, and M.-A. Williams, "Adaptively selecting occupations to detect skill shortages from online jobs," in 2019 IEEE International Conference on Big Data (Big Data), 2019, pp. 1637-1643: IEEE.

[7] S. Das, S. Steffen, W. Clarke, P. Reddy, E. Brynjolfsson, and M. Fleming, "Learning occupational task-shares dynamics for the future of work," in Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society, 2020, pp. 36-42.

[8] C. B. Frey and M. A. Osborne, "The future of employment: How susceptible are jobs to computerisation?" Technological forecasting and social change, vol. 114, pp. 254-280, 2020.

[9] D. Borup and E. C. M. Schütte, "In search of a job: Forecasting employment growth using google trends," Journal of Business & Economic Statistics, pp. 1-15, 2020.

[10] C. Daheim and O. Wintermann, "2050: The Future of Work," Gütersloh, Germany 2016.

[11] World Economic Forum, "The Future of Jobs Report 2020," World Economic Forum Switzerland 2020.

[12] S. Xu, J. Stienmetz, and M. Ashton, "How will service robots redefine leadership in hotel management? A Delphi approach," International Journal of Contemporary Hospitality Management, 2020.

[13] P.-L. Yeoh, "A critical assessment of skills and knowledge for entry-level marketing jobs: A Delphi study," Marketing Education Review, vol. 29, no. 4, pp. 242-265, 2019.

[14] V. Tiberius and S. Hirth, "Impacts of digitization on auditing: A Delphi study for Germany," Journal of International Accounting, Auditing and Taxation, vol. 37, p. 100288, 2019.

[15] R. Gruetzemacher, F. E. Dorner, N. Bernaola-Alvarez, C. Giattino, and D. Manheim, "Forecasting AI progress: A research agenda," Technological Forecasting and Social Change, vol. 170, p. 120909, 2021.

[16] H. Bakhshi, J. M. Downing, M. A. Osborne, and P. Schneider, The future of skills: Employment in 2050, Pearson, 2017.

[17] R. Gruetzemacher, D. Paradise, and K. B. Lee, "Forecasting extreme labor displacement: A survey of AI practitioners," Technological Forecasting and Social Change, vol. 161, p. 120323, 2020.

[18] G. Kim and J. Bae, "A novel approach to forecast promising technology through patent analysis," Technological Forecasting and Social Change, vol. 117, pp. 228-237, 2017.

[19] H. P. Cho, H. Lim, D. Lee, H. Cho, and K.-I. Kang, "Patent analysis for forecasting promising technology in high-rise building construction," Technological Forecasting and Social Change, vol. 128, pp. 144-153, 2018.

[20] M. N. Kyebambe, G. Cheng, Y. Huang, C. He, and Z. Zhang, "Forecasting emerging technologies: A supervised learning approach through patent analysis," Technological Forecasting and Social Change, vol. 125, pp. 236-244, 2017.

[21] H. Zhang, T. Daim, and Y. P. Zhang, "Integrating patent analysis into technology roadmapping: A latent dirichlet allocation based technology assessment and roadmapping in the field of Blockchain," Technological Forecasting and Social Change, vol. 167, p. 120729, 2021.

[22] A. Dechelezeprêtre, D. Hemous, M. Olsen, and C. Zanella, "Automating labor: Evidence from firm-level patent data," in CEP Discussion Papers, Centre for Economic Performance, LSE, London, UK, 2020, vol. 1679.

[23] M. Webb, "The impact of artificial intelligence on the labor market," Available at SSRN 3482150, 2019.

[24] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," arXiv preprint arXiv:1301.3781, 2013.

[25] Q. Le and T. Mikolov, "Distributed representations of sentences and documents," in International Conference on Machine Learning, 2014, pp. 1188-1196: PMLR.

[26] P. Bojanowski, E. Grave, A. Joulin, and T. Mikolov, "Enriching word vectors with subword information," Transactions of the Association for Computational Linguistics, vol. 5, pp. 135-146, 2017.

[27] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," arXiv preprint arXiv:1810.04805, 2018.

[28] M. R. Frank et al., "Toward understanding the impact of artificial intelligence on labor," Proceedings of the National Academy of Sciences, vol. 116, no. 14, pp. 6531-6539, 2019.

[29] J.-S. Lee and J. Hsiang, "Patent classification by fine-tuning BERT language model," World Patent Information, vol. 61, p. 101965, 2020.

[30] N. N. Kachouie and M. Shutaywi, "Weighted Mutual Information for Aggregated Kernel Clustering," Entropy, vol. 22, no. 3, p. 351, 2020.