Neural Network-Based Paper-Matching with Relevant Products through Patents

Seonho Hwang seonho07.hwang@gmail.com Samsung Electronics, Sungkyunkwan University

Juneseuk Shin jsshin@skku.edu Sungkyunkwan University

Many researchers put their resources in significant research topics affecting various fields, including Artificial Intelligence (AI) and Big Data analytics, to produce a large number of research publications. Firms, interested in planning future R&D for launching successful products in the future, would like to take advantage of such publications to interpret research activities from a viewpoint of products of interest. That, however, is not easy because research publications contain knowledge-level information but products are of artifact level. So, it is required to link the data of two different levels. We propose an approach in which research publications can be classified according to product fields and examined from the product perspective.

We need 3 sets of data, including Product Fields (PFs), patents and papers. PFs are determined based on analyzers' interests. While patents contain both artifact and knowledge-level information, PFs contain artifact-level information and papers are of knowledge-level. So, we can use patent data to link PFs and papers. Patents can be searched and obtained from patent offices such as USPTO. The detailed methodology is as follows :

1) CPC-based Patent Searching : Cooperative Patent Classification (CPC) is a patent classification system. We used direct CPC (DCPC) and extended CPC (ECPC). DCPCs are found directly from a determined PF and patents associated with DCPCs are searched and defined 'Primary PATents' (PPAT). ECPCs are found from PPAT and patents associated with both DCPCs and ECPCs are defined 'Extended PATents' (EPAT). The union of PPAT and EPAT forms the selected patents for the PF.

2) Training PF classifier : The PF classifier is a Convolutional Neural Network (CNN)-based text classifier. In order to train the classifier, we need two data sets. One is a word vectors obtained from a word embedding and the other is patent data.

3) Classifying papers : Using the PF classifier, we classify the papers according to the PFs.

4) Analyzing distributions : We can analyze the distribution of PFs in the papers using PFs and classified papers.

To illustrate our methodology, we applied the methodology to Google's publications on 'machine intelligence' because this topic has more publications than any other topic at Google. First, we determined 6 PFs among Google's product fields based on the market share, product popularity and growth potential. The 6 PFs include 'Advertisement (Ad)', 'Image', 'Mail', 'Map', 'Search', and 'Video', each of which covers some Google products. For instance, 'Ad' corresponds to AdWords and AdSense, and 'Image' to Google Photo. Next, we searched patents using DCPCs and ECPCs from the 6 PFs. The search was limited to U.S. patents between 2010 and 2017. Total 122,411 patents were found as of 12/15/2017 and total

979 papers were obtained from ‘Research at Google’ (research.google.com/pubs/papers.html) in the machine intelligence area as of 3/14/2018. We focused on the papers that were published since 2010 to have 771 papers in total.

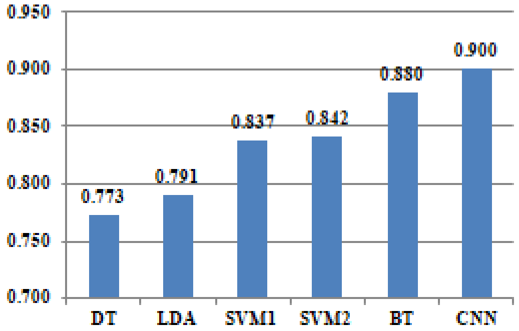

Fig. 1 Classifier performance comparison

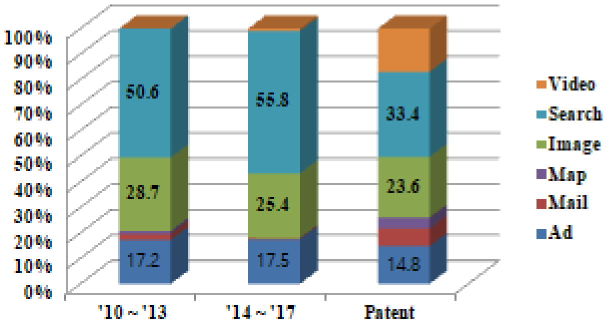

Fig. 2 Distributions of PFs as a percentage

Using word vectors trained from word2vec and randomly selected samples from the 122,411 patents, we trained the CNN-based PF classifier. Before classifying the papers, we evaluated the performance of the classifier to have the average accuracy of 0.900 which was compared with other architectures in Fig. 1. The performances of Decision Tree (**DT**), Linear Discriminant Analysis (**LDA**), Support Vector Machine (**SVM1**), and **SVM2** for 5 classes were reported [1,2]. A three-phase model with Boosted Tree (**BT**) for 2 classes was reported [3]. The measure for all architectures except SVM2 is accuracy and that for SVM2 is f1 score. The figure shows our CNN-based classifier outperforms other architectures. We used this classifier to classify the papers according to the 6 PFs. In order to analyze the distribution of PFs, we selected the papers the predicted probability of which was equal to or greater than 0.90 to have 346 papers in total. Since the number of publications increased rapidly from 2014, we divided the publications into ‘10~’13 and ‘14~’17 depending on the years of publication to compare the PF distributions.

We can figure three PFs are dominant in the order of ‘Search’ > ‘Image’ > ‘Ad’ and other PFs are negligible in terms of number of publications as shown in Fig. 2. Although total number of publications drastically increased from 2014, the distribution of PFs are very similar in both ‘10~’13 and ‘14~’17. The distribution was also compared with that of the 122,411 patents. While the three PFs are significant, other PFs also have a considerable portion in the patents. We can interpret activities in different domains such as research publications or patents from the common product perspective using the methodology. We think the methodology can be extended from one research topic in a firm to multiple topics in multiple firms to construct the research landscape from a product perspective.

References

[1] Wu, J., Chang, P., Tsao, C., Fan, C., 2016. A patent quality analysis and classification system using self-organizing maps with support vector machine. Applied Soft Computing 41, 305-316.

[2] Zhang, X., 2014. Interactive patent classification based on multi-classifier fusion and

active learning. Neurocomputing 127, 200–205.

[3] Al Shamsi, F., Aung, Z., 2016. Automatic patent classification by a three-phase model with document frequency matrix and boosted tree. ICEDSA 2016.