Outlier Patent Identification Based on Anomaly Detection-Taking the Field of UAV as an Example

Zhixun Xu*

1Department of Economics and Management, Nanjing University of Science & Technology, Nanjing, China

*Corresponding author: 3479651504@qq.com

Abstract. We identify outlier patents from different perspectives to obtain more comprehensive identification results and assist in outlier innovation. Firstly, we use the BERT model to vectorize patent titles and abstracts to address the issue of polysemy; Then, we identify outlier patents through various anomaly detection algorithms and compare their recognition results; Finally, we obtain outlier patent topics through the BERTopic model. The experiment found that the outlier patent topics in the UAV field mainly focus on automatic charging, path planning, and 3D modeling.

Keywords: outlier patents; anomaly detection; text representation.

1. Introduction

Since the beginning of the 21st century, global technological innovation has entered an unprecedented period of intensive activity, and the transformation of the technology industry is affecting the global economy and industrial structure. Outlier innovation, as an important source of technological innovation[1,2,3], has significant implications for enhancing original innovation capabilities and leading industrial transformation and development[4,5].

There are three main methods for identifying outlier patents: patent citation based[6,7], patent classification based, and patent text based[8,9]. The method based on patent citations has a certain time delay and requires secondary processing to obtain the same cited information. The method based on patent classification relies on IPC numbers and has high computational costs. Furthermore, both methods overlook the more valuable patent text information. The existing methods based on patent text are insufficient in extracting semantic information from patents, and have not obtained deeper semantic information. And this method has a single recognition angle, and the recognition results are not comprehensive enough.

2. Method

This article first uses the BERT model to obtain the semantic representation of patents; Then, outlier patents are identified using three anomaly detection algorithms: LOF[10], KNN[11,12], and iForest[13,14]; Finally, the technical topics of outlier patents are obtained through the BERTopic model.

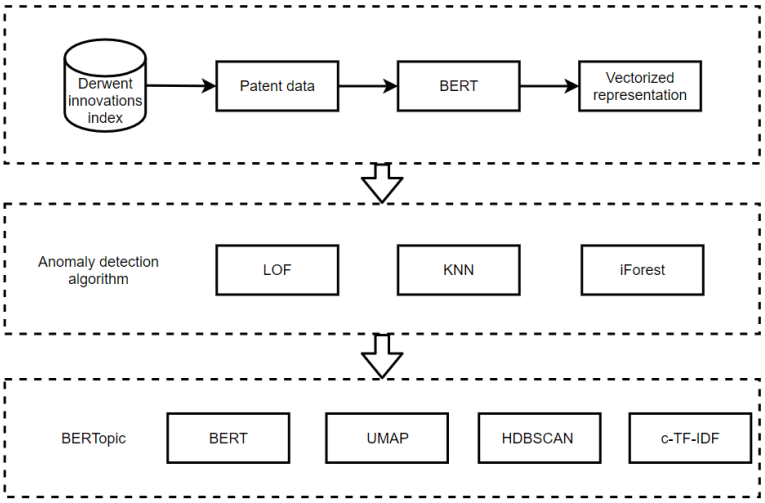

Fig. 1 Technology Roadmap

2.1. Patent semantic vector generation

After obtaining patent data, we preprocess it and input it into the BERT model to obtain patent vectors containing deep semantic information. The BERT model consists of three modules: Embedding module, Transformer module, and Pre tuning module. Traditional language models are typically unidirectional, meaning they can only capture contextual information from front to back and back to front. The BERT model is bidirectional and can capture contextual information before and after words at the same time. The Transformer module of the BERT model is a neural network based on self attention mechanism, which can better capture the dependency relationships between words.

2.2. Identifying Outlier Patents Based on Anomaly Detection Algorithm

This article uses three anomaly detection algorithms, LOF, KNN, and iForest, to identify outlier patents. The LOF algorithm attempts to identify outlier patents from the perspective of local anomalies; The KNN algorithm attempts to identify outlier patents from the perspective of global anomalies. Both of these methods calculate abnormal scores for data points, resulting in low execution efficiency. IForest defines anomalies as outliers that are easily isolated, and this method belongs to ensemble learning algorithms with high processing efficiency. There is a certain complementary relationship among the three anomaly detection algorithms. This article chooses the strategy of taking union sets to fuse the recognition results of the three methods to obtain a more comprehensive set of outlier patents.

2.3. Patent Subject Identification

This article uses the BERTopic model to extract topics from outlier patents. This model is mainly divided into four modules. Firstly, text embedding is used to extract document embedding vectors using pre trained BERT models or any other embedding techniques. An embedding vector can be simply understood as a string of numerical values representing a point in a high-dimensional space. Next is text dimensionality reduction, using UMAP algorithm to reduce the dimensionality of word vectors, mapping them to a low dimensional space while preserving important global and local structural information. Next is text clustering, which uses the HDBSCAN algorithm to cluster similar vectors into the same cluster, forming different themes. Finally, topic representation is used to calculate the importance of topic words in each topic cluster using the c-TF-IDF method, and topic feature words are extracted based on the maximum marginal correlation.

3. Empirical research

3.1. Experimental data acquisition and processing

The experimental data is sourced from the Derwent database, with a search period from January 1, 2020 to December 31, 2020. The search equation is TI=(((unmanned OR automatic OR autonomous OR remotely piloted OR nonhuman) AND (aircraft OR "aerial vehicle" OR airship* OR drone OR plane OR aerocraft* OR airplane OR aerobat* OR aerostat*)) OR "UAV")， A total of 13787 patent data were obtained. After obtaining patent data, we perform processes such as removing invalid values, morphological reduction, stem extraction, and removing stop words.

3.2. Outlier Patent Identification

This experiment adopted three anomaly detection algorithms, LOF, KNN, and iForest, to identify outlier patents, and compared them from the perspectives of visualizing the distribution of outlier patents and identifying the number of patents.

3.2.1 Outlier distribution

As shown in the figure below, we compared the outlier patent distribution results of LOF, KNN, iForest, and fusion methods. The outlier patents identified based on the KNN algorithm are distributed near sparse points in the global lower part, indicating that they focus on global outliers; The outlier patents identified based on the iForest algorithm are distributed near the sparse point in the lower left corner of the world, indicating that it focuses on global outliers; The outlier patents based on the LOF algorithm are distributed in the dense point area in the upper right, indicating that they focus on local outliers. The recognition results of the fusion of the three methods identified more outlier patents near both dense and sparse points, which validates the effectiveness of the multi anomaly detection algorithm and enables a more comprehensive identification of outlier patents.

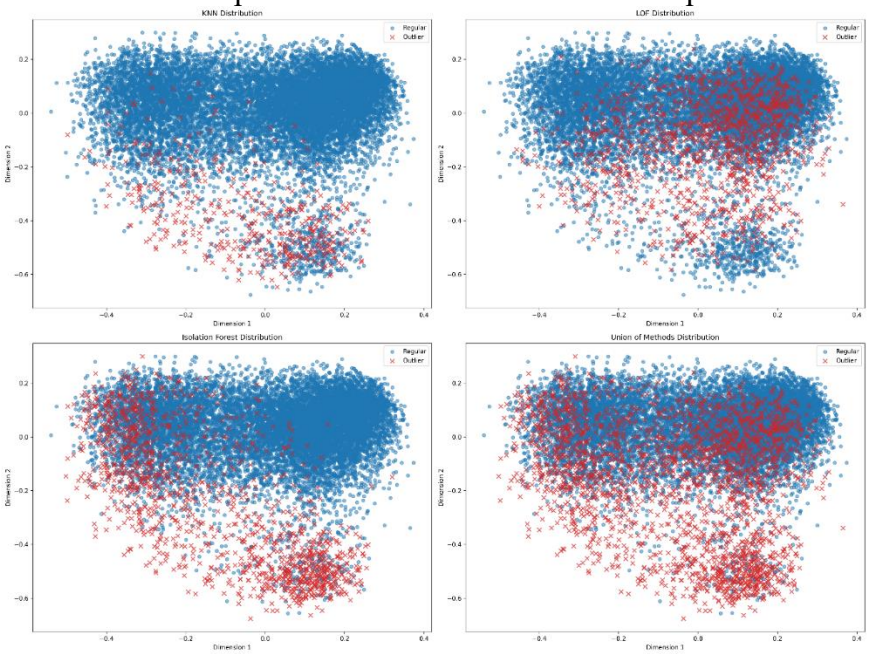

Fig. 2 Outlier distribution map

3.2.2 Comparison of Identification Quantity

As shown in the table below, the number of outlier patents identified based on the LOF algorithm is 1378; The number of outlier patents identified based on KNN algorithm is 457; The number of outlier patents identified based on the iForest algorithm is 1378.

Table 1. Number of outlier patents identified

Method	Identification Quantity
LOF	1378
KNN	457
iForest	1378

On this basis, we conducted intersection analysis on the recognition results of the three methods, and the specific results are shown in the table below. Intersection analysis was conducted on outlier patents identified by LOF algorithm and KNN algorithm, and it was found that there were 295 shared patents, accounting for 21.4% of the number identified by LOF algorithm and 64.5% of the number identified by KNN algorithm. An intersection analysis was conducted on the outlier patents identified by the LOF algorithm and the iForest algorithm, and it was found that there were 232 shared patents, accounting for 16.8% of the number identified by the LOF algorithm and 50.7% of the number identified by the iForest algorithm. An intersection analysis was conducted on the outlier patents identified by KNN algorithm and iForest algorithm, and it was found that there were 371 shared patents, accounting for 81.1% of the number identified by KNN algorithm and 26.9% of the number identified by iForest algorithm. And the intersection identified by the three methods is a total of 181, accounting for 13.1% of the recognition quantity of LOF algorithm, 39.6% of the recognition quantity of KNN algorithm, and 13.1% of the recognition quantity of iForest algorithm. This to some extent proves the complementarity of the three anomaly detection algorithms, as they identify different outlier patents from different perspectives. Adopting the union method can obtain a more comprehensive set of outlier patents.

Table 2. Number of intersections between different methods

Method	Identification Quantity
LOF-KNN	295
LOF-iForest	232
KNN-iForest	371
LOF-KNN-iForest	181

3.3. Topic Identification

This article inputs outlier patents into the BERTopic model to obtain their topics, topic word distributions, and patent distributions. As shown in the figure below, the outlier patents in 2020 are roughly clustered into 12 themes.

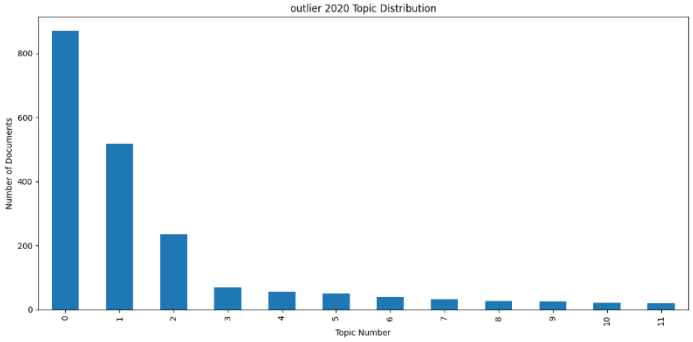

Fig. 3 Outlier patent topic distribution

By analyzing the keywords it contains, we can obtain its theme. The main themes of outlier patent technologies in 2020 are focused on "drone spraying", "drone conveying system", "path planning", "automatic inspection", "automatic charging", "real-time monitoring", "de icing", "3D modeling", "fire extinguishing", "feeding device", "inertial navigation", and "feature point extraction".

4. Summary

This article uses the BERT model to obtain deep semantic information of patents and vectorize it; Integrating three anomaly detection algorithms for outlier patent recognition to obtain a more comprehensive set of results; Using BERTopic model to obtain technical topics of outlier patents. The effectiveness of this method has been verified through experiments on patent data in the field of drones, but further optimization of the anomaly detection algorithm is needed in the future.

References

[1] Lee J, Park S, Lee J. Technology Opportunity Analysis Based on Machine Learning[J]. Axioms, 2022, 11(12).

[2] Song K, Kim K, Lee S. Identifying promising technologies using patents: A retrospective feature analysis and a prospective needs analysis on outlier patents [J]. Technological Forecasting & Social Change, 2018, 128: 118-132.

[3] Kneeland M K, Schilling M A, Aharonson B S. Exploring Uncharted Territory: Knowledge Search Processes in the Origination of Outlier Innovation [J]. Organization Science, 2020, 31(3): 535-795+C2.

[4] Aharonson B S, Schilling M A. Mapping the technological landscape: Measuring technology distance, technological footprints, and technology evolution [J]. Research Policy, 2016, 45(1): 81-96.

[5] Michael G, Balachandran N L, Matthias W, et al. Using Outliers for Theory Building [J]. Organizational Research Methods, 2020, 24(1): 109442811989887-109442811989887.

[6] Zhou Y, Dong F, Liu Y, et al. A deep learning framework to early identify emerging technologies in large-scale outlier patents: an empirical study of CNC machine tool [J]. Scientometrics, 2021, 126(2): 1-26.

[7] Rodriguez A, Tosyali A, Kim B, et al. Patent Clustering and Outlier Ranking Methodologies for Attributed Patent Citation Networks for Technology Opportunity Discovery [J]. IEEE Transactions on Engineering Management, 2016, 63(4): 426-437.

[8] Yoon J, Kim K. Detecting signals of new technological opportunities using semantic patent analysis and outlier detection[J]. Scientometrics, 2011, 90(2): 445-461.

[9] Wang J, Chen Y-J. A novelty detection patent mining approach for analyzing technological opportunities[J]. Advanced Engineering Informatics, 2019, 42.

[10] Tomohiro M, Libo Z, Taisuke O, et al. Development of machine-learning based anomaly detection system for manufacturing:- Autoencoder-LOF model [J]. Proceedings of International Conference on Leading Edge Manufacturing in 21st century : LEM21, 2021, 2021.10(0): 148-010.

[11] Venkataramanaiah B, Kamala J. Retraction Note: ECG signal processing and KNN classifier-based abnormality detection by VH-doctor for remote cardiac healthcare monitoring [J]. Soft Computing, 2024, 28(suppl 2): 1-1.

[12] Esan D, Owolawi P A, Tu C. Anomalous Detection in Noisy Image Frames using Cooperative Median Filtering and KNN [J]. IAENG International Journal of Computer Science, 2022, 49.0(1.0).

[13] Liu L, Min X. A Study of an Engine Anomaly Detection Model IForest-ADOA [J]. Journal of Physics: Conference Series, 2022, 2171(1).