Construction and Application of Methods for Fintech Patent Identification and Classification: Postprint

Authors: Xu Lu, Lu Xiaobin, Yang Guancan

Date: 2023-04-01T16:15:55+00:00

Abstract

[Purpose/Significance] Financial technology (fintech) is experiencing rapid development in the information and data era, accompanied by a continuous increase in patent volume. Concurrently, characteristics such as cross-disciplinary convergence and blurred domain boundaries in fintech have heightened the complexity of patent analysis. Consequently, it is essential to construct appropriate identification and classification methodologies to accurately and efficiently process the ever-expanding large-scale datasets. [Method/Process] Initially, based on the connotation and functional attributes of fintech, we systematically delineate the innovation categories it encompasses and clarify the scope and boundaries of fintech patents. Subsequently, we construct a methodological workflow for fintech patent identification and classification by leveraging machine learning algorithms in conjunction with text filtering and manual adjudication. [Results/Conclusions] This study proposes a patent identification and classification workflow grounded in machine learning algorithms, which demonstrates the capability to identify and categorize fintech patents with relatively high accuracy and efficiency. Furthermore, through analysis of the resultant fintech patent classification data, we synthesize a comprehensive overview of the current developmental landscape of fintech.

Full Text

Preamble

FinTech Patent Identification and Classification: Method Construction and Application

Xu Lu, Lu Xiaobin, Yang Guancan
School of Information Resource Management, Renmin University of China, Beijing 100872

Abstract: [Purpose/Significance] FinTech has developed rapidly in the information and data era, with a continuously growing number of patents. Meanwhile, its characteristics of interdisciplinary overlap and blurred boundaries have increased the difficulty of patent analysis. Therefore, it is necessary to construct suitable identification and classification methods to accurately and efficiently process this continuously growing large-volume data. [Method/Process] First, based on the connotation and function of FinTech, this study sorts out its innovation categories and clarifies the scope and boundaries of FinTech patents. Subsequently, using machine learning algorithms combined with text filtering and manual interpretation, we construct a methodological process for FinTech patent identification and classification. [Result/Conclusion] This paper proposes a patent identification and classification process based on machine learning algorithms that can accurately and efficiently identify and classify FinTech patents. By analyzing the resulting classified FinTech patent data, we summarize the current development status of FinTech.

Keywords: FinTech; information economics; patent analysis; machine learning
Classification Number: G250
DOI: 10.13266/j.issn.0252-3116.2020.11.010

With the development and gradual maturation of big data, cloud computing, and artificial intelligence, the application of new technologies in the financial industry has become more widespread. The deep integration of finance and technology has promoted the development of Financial Technology (FinTech), which is considered a profound transformation and innovation in the financial industry in the information age [1-2]. At the same time, the number of FinTech-related patent applications and grants has continued to grow, making it an important channel for market positioning. Therefore, accurately retrieving and analyzing FinTech patents helps identify innovations in the FinTech field, analyze the most promising branches and industry applications, and thus grasp the development trends of FinTech, supporting strategic planning for national industries and micro enterprises.

As an interdisciplinary field between finance and technology, FinTech patent identification and analysis faces two major challenges. On the one hand, the technological scope involved in FinTech is very broad, and its current definition is relatively ambiguous, making it difficult to directly use traditional International Patent Classification (IPC) or keyword-based methods to identify innovation categories and reasonably define the boundaries of FinTech patents. On the other hand, with the rapid development of FinTech in recent years, the number of innovations and patents in related fields has continued to grow, requiring the construction of a suitable retrieval process to quickly and accurately identify different categories of FinTech innovations from massive patent data. To address this, this study starts from the connotation and function of FinTech, sorts out its innovation categories, and clarifies the scope of FinTech patents as the foundation for subsequent identification and analysis. It then innovatively proposes a method process that uses machine learning algorithms combined

with text filtering and manual interpretation for FinTech patent identification and classification. Subsequently, based on the obtained classified patent data, it analyzes the technological layout and development status of FinTech. Additionally, the method proposed in this study also has certain reference value for patent analysis in similar emerging interdisciplinary fields.

1 Related Research

1.1 FinTech-Related Research

In recent years, academic attention to FinTech has continued to rise, with research perspectives covering specific technological innovations and applications [3-4], challenges and opportunities facing traditional financial industries [5-6], and corresponding risk prevention and market regulation [7-8]. However, few scholars have explored FinTech development from a patent perspective. FinTech patents can create enormous market value [9], and accurately identifying different types of FinTech patents is significant for tracking technological development trends, analyzing innovation impact factors, and supporting national FinTech strategic planning. Among the few studies related to FinTech patents, Zhao Xing searched and analyzed FinTech patents involving facial recognition, big data analysis and prediction, artificial intelligence technology, and digital currency within IPC class G06Q, pointing out that China's patent layout in the FinTech field is relatively weak [10]. The globally influential intellectual property media IPRdaily, in collaboration with the incoPat Innovation Index Research Center, released the "2019 Global FinTech Invention Patent Rankings (Top 100)" [11], which limited IPC numbers to G06Q20, G06Q30, and G06Q40, focusing on major application fields such as finance, payment, shopping, e-commerce, insurance, and taxation, and counted the number of publicly disclosed FinTech invention patent applications by global enterprises in 2019. The above research and practical work on FinTech patents generally limited searches to single or a few IPC numbers, which cannot comprehensively cover all patents for FinTech innovations and makes it difficult to ensure the comprehensiveness and effectiveness of retrieval results.

1.2 Patent Identification-Related Research

Traditional patent identification or retrieval often uses IPC numbers and keyword-based methods. IPC numbers are an internationally recognized tool for managing and retrieving patent literature, recording the "classification number" and "main classification number" for each invention and utility model patent, and are therefore widely used in patent classification and retrieval [12-13]. On the one hand, using IPC numbers, one can achieve rapid retrieval and access to patent information for specific industries by constructing a correspondence between patent classification and industrial classification, including various association models such as expert determination, cross-retrieval, probability calculation, and similarity measurement [14-16]. On the other hand, combining IPC numbers with keywords can further optimize retrieval

methods and improve patent identification efficiency [17-18]. However, scholars and practitioners have also pointed out that with technological development, patent identification and retrieval in many fields such as smartphones and pharmaceuticals face challenges of cross-domain applications and complex technological systems, making traditional IPC-based retrieval methods less effective [19-20]. Similarly, as an emerging field with rapid development, FinTech also has problems of ambiguous concept definition and interdisciplinary overlap. In massive datasets, it is difficult to directly use IPC numbers for accurate and large-scale patent identification and analysis. Therefore, it is necessary to specifically construct a financial vocabulary list, perform text filtering on all G and H class patent data covering essential FinTech technologies to obtain title and abstract information that may be FinTech-related, and conduct word segmentation and TF-IDF feature extraction.

In recent years, data analysis methods using machine learning algorithms have developed rapidly and been widely applied to research problems in science and technology, economics, finance, and other fields [21-22]. Machine learning algorithms can perform batch vectorization processing on various text data, and when combined with manual analysis, can more efficiently analyze various problems in massive datasets [23-24], providing a new approach for FinTech patent identification and classification. Overall, existing FinTech research mostly focuses on the technology itself, discussing specific innovation scenarios and applications, market impacts, and policy regulations, but there is little work on patent data. Meanwhile, FinTech is an emerging interdisciplinary field between finance and technology, with a continuously growing number of patents, making traditional IPC-based retrieval methods unable to fully cover all FinTech fields. Additionally, FinTech lacks a clear conceptual definition, making it difficult to construct an accurate and comprehensive keyword list. Therefore, facing massive patent data, the main content of this study includes three aspects: manually interpreting and sorting out FinTech innovation categories to clarify the scope of FinTech patents as the foundation for subsequent identification and analysis; using machine learning algorithms combined with text filtering and manual interpretation to construct a methodological process that can process large-volume patent data and identify and classify FinTech patents; combining the obtained classified patent data to analyze the type distribution and business applications of FinTech innovations, thereby grasping the current development status of FinTech.

2 Process and Method for FinTech Patent Identification and Classification

2.1 Process Framework

This study proposes a FinTech patent identification and classification process framework that uses machine learning combined with text filtering and manual interpretation. First, a financial vocabulary list is constructed to perform text

filtering on all patent data covering essential FinTech technologies in G and H classes, obtaining title and abstract information that may be FinTech-related, and conducting word segmentation and TF-IDF feature extraction. Second, based on the connotation and function of FinTech, financial technology innovation categories are manually sorted out to define the scope of FinTech patents; further, a sample set is reasonably constructed through two rounds of manual interpretation combined with the K-means algorithm. Third, sample data is randomly extracted and split, with machine learning algorithm parameters tuned on the training set; then tested on the test set, with the optimal model determined by comprehensively considering each algorithm's accuracy, precision, recall, and F1 value; and finally, patent identification and classification are performed on the full dataset. This constitutes a core process framework of data collection → sample selection → model construction, as shown in Figure 1 [Figure 1: see original paper]. In Figure 1, blue solid rectangles represent the main machine learning algorithms, orange solid rectangles represent manual intervention and judgment based on financial expertise, and green solid rectangles represent the text filtering process. Therefore, on the basis of text-filtered patent data, the entire processing process is always driven by machine learning algorithms and interacts with manual judgment to comprehensively identify and classify FinTech patents.

2.2 Dataset Construction and Text Preprocessing

2.2.1 Data Collection The data for this study comes from the Lens patent database, jointly developed by Cambia and Queensland University of Technology, which provides open global patent information and academic literature data with an update cycle of 3-4 weeks, covering full text and images of U.S. patent applications since 2001 and patent grants since 1976. The United States leads global FinTech development, so this study uses U.S. patent grant data as the research object. This study focuses on FinTech patent grants in the past five years, with the search time limited from January 1, 2014, to December 31, 2018, retrieving a total of 1,528,774 U.S. granted patents.

To preliminarily understand FinTech patents, this study used FinTech essential technology-related terms such as "big data," "cloud computing," and "AI" to search in the dataset, with results showing that International Patent Classification numbers are concentrated in G (Physics) and H (Electricity) classes. Therefore, this study further limited IPC classification numbers to G and H classes, which can cover all patents for electronic computing technologies essential to FinTech. This process deleted 1,328,623 data entries, leaving 200,151 records.

Finally, this study constructed a financial vocabulary list to further narrow the dataset through word list filtering. Drawing on the financial vocabulary list constructed by M. Chen et al. [9], which combines the C.R. Harvey financial vocabulary list and words and phrases from the Oxford Dictionary of Finance, we extracted words that can clearly reflect association with financial services

(such as bourse, chargeback, futures, security, bank) and phrases related to financial services (such as health insurance, mutual fund). On this basis, this study added some new words recently identified as FinTech terms (such as digital currency, smart contract), ultimately forming a financial vocabulary list of 478 words closely related to financial services (not reported here due to space limitations, available upon request). Using this financial vocabulary list, we further filtered out patents that did not contain any financial vocabulary in their titles and abstracts, totaling 162,995 data entries. The final dataset consisted of 37,156 patent records, covering complete detailed data including title, abstract, and claims for each patent. The dataset construction process is shown in Table 1 .

2.2.2 Text Preprocessing The text preprocessing stage uses the text mining module of KNIME. First, word segmentation is performed on the raw text. Using the OpenNLPEnglish Tokenization tokenizer, segmentation is performed on words, punctuation marks, and numbers, preserving punctuation; then punctuation, pure numbers, words containing abnormal characters, words with length less than 3, and words in the common stop word list such as be, the, that are removed. The Stanford tagger annotation tool is used to filter nouns and noun phrases; finally, the Stanford lemmatizer tool is used to further integrate related vocabulary, reducing the impact of tense and mood on word items in English context, ultimately achieving text processing of FinTech-related patents and transforming patent text language into word groups composed of important terms.

Subsequently, the text is vectorized, with the number of feature words set to 3,000 dimensions. The bag-of-words model is a commonly used model for converting sentences into vector representation. This model does not consider the order of words in sentences, only considering the frequency of words from the word list appearing in the text. This study uses Term Frequency-Inverse Document Frequency (TF-IDF) instead of simple word frequency to construct the bag-of-words model. The TF-IDF method is a commonly used text feature vectorization method for evaluating the importance of words to a document in a corpus. Its advantage is that it considers the impact of high-frequency words while using inverse document frequency for weighted processing: the more times a word appears in a document, the higher its importance, but if it appears in more documents, its importance in the classification process is lower, thus reflecting the word's ability to distinguish the current document from others. The specific calculation formulas are as follows:

$$TF(w) = \frac{\text{Number of times word w appears in the document}}{\text{Total number of words in the document}} \tag{1}$$

$$IDF(w) = \log\left(\frac{\text{Total number of documents in the corpus}}{\text{Number of documents containing word w}}\right) \tag{2}$$

$$TF\text{-}IDF(w) = TF(w) \times IDF(w) \tag{3}$$

2.3 FinTech Patent Sample Set Construction

2.3.1 Clarifying the Scope and Categories of FinTech Patents As a new concept, FinTech does not have a completely consistent and clear definition. As shown in Table 2 , different institutions and organizations have provided definitions of FinTech from their respective perspectives, but all emphasize the important role of emerging technologies in the FinTech field. Meanwhile, domestic scholars have also elaborated on the connotation of FinTech from multiple dimensions. For example, Han Mei pointed out that FinTech is a convergence industry of finance and information technology, with the connotation of using data and technology as drivers to improve the overall operational efficiency of the financial industry and reduce industry operating costs [25]. Ye Chunqing believes that FinTech is the further development of Internet finance, enhancing service efficiency and customer experience through the integration of technology and finance [26].

Regarding FinTech itself and its included technology types, although domestic and international organizations and scholars have different understandings, they generally agree that the core of FinTech is the integration of finance and technology, and the application of new technologies in the financial field to achieve efficiency improvements and cost savings in the financial industry. Under this basic connotation, the continuous integration process of frontier technology and finance has given rise to various technological innovations in the FinTech field. The FinTech patents analyzed in this study also refer to frontier technological innovations that can be applied to various businesses and processes in the financial field.

Currently, there are two main classification methods for FinTech innovations: one is classification according to the application direction of FinTech in the financial industry; the other is classification based on the underlying technology of FinTech. For example, the Financial Stability Board (FSB) in 2017 classified according to industry application perspectives into five economic functions: payment management, deposit/loan and capital raising, investment management, market facilities, and insurance, and further subdivided categories such as IoT, electronic trading, cloud computing, big data, robo-advisory, distributed ledger, electronic identity authentication, and mobile payment by combining underlying technologies.

Based on a thorough understanding of FinTech's connotation and underlying technologies and their applications in the financial field, and through reading patent literature in the dataset, this study proposes dividing FinTech patents into six categories of technological innovation: encryption security, mobile payment, data analysis, IoT network, intelligent trading, and online lending, thereby clarifying the content and boundaries of FinTech patents and providing

a foundation for constructing accurate patent identification and classification methods. Specific classifications and application examples are shown in Table 3 .

2.3.2 Constructing Sample Set by Combining Machine Learning Algorithms and Manual Interpretation First, manual reading and classification were used to construct an initial sample. Using a list of FinTech innovation enterprises [27] and highly relevant FinTech terms (such as mobile transaction, mobile payment, internet security, internet of things, blockchain), we conducted preliminary searches in the dataset to obtain 765 patent data entries. Table 3 summarizes the six specific categories of FinTech, while the dataset also includes patent data that are not FinTech, recorded as: Category 7, "financially related but non-FinTech patents," including some pure financial business methods and design patents that, although related to finance, do not match the connotation of "frontier technology applied to financial business and processes"; and Category 0, "non-financially related patents," including purely technical innovation patents without specific feasible financial application scenarios. Therefore, the 0-7 classification can completely cover all categories in the dataset, where categories 1-6 are FinTech patents. By manually reading patent titles and abstracts, we annotated 765 patent data entries, with initial classification results shown in the "Initial Sample" column of Table 4 .

Subsequently, the K-means method was used to expand the sample set. The independent and identically distributed nature of labeled samples and unlabeled datasets is a basic premise for supervised learning, meaning that annotated samples should be randomly extracted and independent of each other. To ensure the randomness of annotated sample patents while ensuring sufficient representative quantity for each category, this study further expanded the training sample set based on the 765 initially classified samples. The K-means clustering algorithm is an iterative clustering analysis algorithm. By running K-means with K=1 on each category of the initial classified samples, we obtained cluster centers for each FinTech patent category. We then iteratively calculated the cosine distance between all patents in the complete dataset and each cluster center. If a patent had the shortest cosine distance to a cluster center of a certain category in the initial samples, it was assigned to that category. Using the initial classified samples, this study applied the K-means method to cluster the entire dataset, selecting the 100 patents closest to the cluster center in each category and randomly extracting 100 patents as supplements. This ensures both the selection of basic features for each category and increased randomness of extraction, providing better representation of the whole, ultimately obtaining 1,600 patent data entries across 8 categories.

Finally, manual reading and classification were performed again to determine the final sample. The 1,600 samples were manually read and annotated again to form the final sample for machine learning-based patent identification and classification, with final results shown in the "Final Sample" column of Table 4 .

Compared with the initial classification, the final sample has richer annotations and better representation of the overall dataset.

2.4 FinTech Patent Identification and Classification

2.4.1 Parameter Tuning on Training Set This stage first performs sample splitting, randomly extracting 80% of samples as training data for learning training and parameter tuning, with the remaining 20% used as test set data for validation, using the annotated sample set to find the optimal machine learning model. Commonly used machine learning classification algorithms include: Support Vector Machine (SVM), Gradient Boosting Decision Tree (GBDT), Random Forest (RF), Decision Tree (DT), and K-Nearest Neighbor (KNN), each with wide applications in different scenarios and classification problems.

This study uses accuracy rate, precision rate, recall rate, and F1 value as comprehensive evaluation metrics for machine learning effectiveness. Accuracy rate is the ratio of all correctly classified patent documents to the total number of patent documents, which can intuitively measure the identification and classification effectiveness of machine learning algorithms. Precision rate is the ratio of accurately classified patent documents to all documents predicted as that class. Recall rate is the ratio of accurately classified documents to actual documents. The F1 value is a comprehensive evaluation metric that considers both precision and recall. Since this study is a multi-classification problem, it is necessary to calculate overall comprehensive evaluation metrics based on each binary classification evaluation metric. This study uses the "macro_" calculation method. Specific calculation formulas for binary and multi-classification evaluation metrics are shown in Table 5 .

Therefore, this study selected all five machine learning algorithms mentioned above, using cross-validation methods and maximizing the comprehensive F1 value metric for parameter tuning on the training set. The final parameter selections were: SVM algorithm with cost=0.6, loss=0.1, nu=0.5; Gradient Boosting algorithm with treedepth=10, learningrate=0.05; Random Forest algorithm with treedepth=10; Decision Tree algorithm using Gini index and Minimum Description Length (MDL) principle; KNN with k=6. The accuracy, precision, recall, and F1 values achieved by each machine learning method on the training set are shown in Table 6 .

2.4.2 Determining Optimal Model on Test Set Based on the parameter values obtained from the training set, we tested the effectiveness of each machine learning method through cross-validation, calculating various evaluation metrics using the same methods. Table 7 shows the accuracy, precision, recall, and F1 values of each machine learning algorithm on the test set data, with Random Forest showing the best comprehensive evaluation across all metrics (bolded), including 75.28% accuracy, 60.27% recall, and 64.39% F1 value.

Random Forest is an ensemble learning model that uses decision trees as base

classifiers. It contains multiple decision trees trained through ensemble learning techniques. When inputting samples for classification, the final classification result is determined by voting on the output results of individual decision trees. Random Forest overcomes the overfitting problem of decision trees, has good tolerance for noise and outliers, and exhibits good scalability and parallelism for high-dimensional data classification problems. Additionally, Random Forest is a data-driven non-parametric classification method widely applied in practical fields such as bioinformatics, business management, text classification, and economic finance. Based on the comprehensive data, the Random Forest method achieves the best results for patent identification and classification in this study. Therefore, this study ultimately uses the Random Forest model for FinTech patent identification and classification.

2.4.3 Patent Identification and Classification on Full Dataset Using the Random Forest machine learning method to identify and classify the dataset, we counted the annual number of granted patents in the six FinTech categories—encryption security, mobile payment, data analysis, IoT network, intelligent trading, and online lending—in the United States from 2014 to 2018, as shown in Table 8 .

The results show that in 2014, the number of U.S. FinTech grants reached a small peak of 417, decreased to only 198 in 2015, but then showed a year-by-year stable growth trend. The concept of FinTech was formally proposed in 2011 [28], and the integration of finance and technology gradually moved from the Internet finance model to the new FinTech model. Before this, relying on mature Internet technology, various related patents emerged explosively, particularly in mobile payment, intelligent trading, and online lending categories, which showed faster technological layout with more granted patents in 2014. Subsequently, FinTech entered a stage of in-depth development, with gradual deepening integration with underlying technologies such as blockchain, cloud computing, and artificial intelligence. Related technology R&D and layout proceeded steadily, so after 2015, various FinTech patents gradually showed a stable growth trend, and overall FinTech development entered a mature period of stable growth.

Meanwhile, the obtained FinTech patent data involved 618 IPC numbers. We counted the top five IPC numbers by quantity in each patent category, as shown in Table 9 . The results show that, on the one hand, the IPC numbers involved in all FinTech patents are extensive; on the other hand, the IPC number distribution also varies significantly across different FinTech patent categories. Therefore, this also demonstrates that applying traditional retrieval methods makes it difficult to comprehensively and accurately analyze FinTech patents. Additionally, the results of this study can also provide reference and guidance for future research using IPC numbers to retrieve and analyze FinTech patents and for constructing associations between IPC numbers and related industries.

3 Analysis of FinTech Patent Data

This section analyzes and summarizes the current development status of FinTech from the perspectives of category distribution and business application based on the obtained classified patent data.

3.1 Type Distribution Analysis

Statistics were compiled for FinTech patents according to the six categories of encryption security, mobile payment, data analysis, IoT network, intelligent trading, and online lending. As shown in Figure 2 [Figure 2: see original paper], mobile payment FinTech patents have the largest quantity with 682, followed by encryption security and intelligent trading with 226 and 214 respectively. Relatively speaking, online lending FinTech patents are the least numerous with only 58. From the development trends of various FinTech patents shown in Figure 3 [Figure 3: see original paper], the proportion of mobile payment FinTech patents has remained at a high level, with an average proportion of 47% and showing a year-by-year expansion trend. Additionally, the proportions of encryption security, data analysis, and IoT network are relatively stable, with average proportions of 16%, 11%, and 8% respectively. In contrast, intelligent trading patents have shown a clear decreasing trend in recent years, with an average proportion of 15%, while online lending patents have the smallest overall proportion, averaging only 4%, but showing a trend of first decreasing then slightly increasing.

In summary, mobile payment FinTech patent technology has an absolute advantage in both absolute numbers and development trends, maintaining stable growth. Intelligent trading patent technology has a moderate overall quantity but is gradually shrinking, while online lending patent technology has relatively small absolute numbers and proportion but shows a growth trend in recent years, indicating certain development potential. In the future, with the maturation and application of 5G technology and facial recognition technology, the mobile payment field will remain a key potential area for FinTech layout. Additionally, the maturation of blockchain technology will also bring new opportunities for patent layout in encryption security and online lending.

3.2 Business Application Analysis

FinTech patent application business scenarios can be divided into seven major categories: banking business, capital market business, insurance business, payment business, lending business, and non-specific business. By reading titles and abstracts and combining IPC numbers and other information, we conducted statistics and analysis on the application business scenarios of the obtained FinTech patents. From the perspective of the total number of FinTech patents applied in different business areas (see Figure 4 [Figure 4: see original paper]), non-specific business has the most patents with 770, accounting for 45.8%, followed by payment business with 460 (27.3%), and then banking business with

290 (17.2%). Relatively speaking, FinTech patents are less applied in insurance business and lending business, with 31 and 32 respectively, accounting for only about 2% of the total.

From the perspective of specific types of FinTech patents applied in different business areas, Figure 5 [Figure 5: see original paper] shows the top two patent categories by application number in each business domain. The results show that in banking business, payment business, and non-specific business, mobile payment and encryption security FinTech patents are more widely applied; in capital market business and insurance business, intelligent trading and data analysis FinTech patents are more applied; while in lending business, mobile payment and online lending FinTech patents rank higher.

In summary, the main application industries for current FinTech patents include traditional financial industry banking business and payment business that has emerged relying on Internet finance, with relatively deep integration of technology and corresponding financial business. At the same time, this study finds that mobile payment and encryption security FinTech patents have the most extensive application scenarios, providing transaction payment convenience and information security guarantees in banking business, payment business, and other non-specific businesses. In comparison, intelligent trading is mainly applied in capital market business and insurance business, providing intelligent stock and insurance recommendations; while online lending patents, such as "P2P" and "crowdfunding" related technologies, promote the development of new lending businesses.

Conclusion

This study sorted out innovation categories based on the connotation and function of FinTech, thereby clarifying the scope of FinTech patents. Subsequently, using machine learning algorithms combined with text filtering and manual interpretation, we constructed a model for FinTech patent identification and classification, achieving automated patent retrieval. Meanwhile, based on the obtained classified data, this study also preliminarily explored the current development status of FinTech patents. When facing massive patent data with interdisciplinary overlap and complex technological systems, traditional patent retrieval methods have problems such as complex search terms and inaccurate IPC number retrieval. The patent retrieval process proposed in this paper has certain reference significance for patent analysis in such emerging interdisciplinary fields. Additionally, and equally importantly, this study can also provide reliable patent data and acquisition methods for future in-depth exploration of important issues such as FinTech enterprise innovation and industrial development influencing factors. Therefore, this study has both theoretical and practical significance.

In future research work, we can further optimize text processing methods to improve the accuracy of machine learning classification models, expand the scope

of datasets for global patent data analysis, and make more detailed considerations for inter-country differences. Additionally, subsequent research can further correlate FinTech patent grant data with company micro-data, industry meso-data, and economic macro-data to deeply explore how to promote the sustained and stable development of FinTech innovation from corporate governance and institutional construction perspectives.

References

[1] Mou Naimi. Seizing FinTech opportunities to improve the level of serving the real economy[J]. Banker, 2018, 35(6): 29.

[2] Xu Yang, Liu Shuwen, Teng Fei, et al. Application prospects of intelligence research in FinTech practice[J]. Library and Information Service, 2017, 61(16): 107-112.

[3] Taylor M B. The evolution of bitcoin hardware[J]. Computer, 2017, 50(9): 58-66.

[4] Moon W Y, Kim S D. Adaptive fraud detection framework for FinTech based on machine learning[J]. Advanced science letters, 2017, 23(10): 10167-10171.

[5] Japparova I, Rupeika-Apoga R. Banking business models of the digital future: the case of Latvia[J]. European research studies, 2017, 20(3): 864-878.

[6] Lee I, Shin Y J. FinTech: ecosystem, business models, investment decisions, and challenges[J]. Business horizons, 2018, 61(1): 35-46.

[7] Barberis J, Arner D W. FinTech in China: from shadow banking to P2P lending[C]//TASCHE P, ALETTER D, PELIZZON L, et al. Banking beyond banks and money. Switzerland: Springer, 2016: 69-96.

[8] Ng A W, Tang W. Regulatory risks and strategic controls in the global financial centre of China[C]//CHOI J J, POWERS M R, ZHANG X T. The political economy of Chinese finance. Bradford: Emerald Group Publishing Limited, 2016: 243-270.

[9] Chen M A, Wu Q, Yang B. How valuable is FinTech innovation?[J]. The review of financial studies, 2019, 32(5): 2062-2106.

[10] Zhao Xing. Analysis of FinTech patent competition situation[N]. China Intellectual Property News, 2018-07-25(5).

[11] IPRdaily Chinese website and incoPat Innovation Index Research Center. 2019 Global FinTech Invention Patent Rankings (Top 100)[EB/OL]. [2020-03-14]. https://www.bankbuy.net/news/detail/art_{id}/556.html.

[12] Song Qun, Wu Guangyin. Preliminary exploration of patent data organization based on IPC[J]. Digital Library Forum, 2013, 9(4): 67-70.

[13] Wen Fangfang. Application of patent classification number coupling analysis in identifying potential cooperative relationships among enterprises[J]. Modern Information, 2018, 38(7): 142-147.

[14] Peng Maoxiang, Xu Yong. Research on constructing and applying the correspondence between patent classification and industrial classification[J]. Science Management Research, 2017, 35(5): 30-33.

[15] Tian Chuang, Zhao Yajuan. A patent-industry category mapping model based on similarity: taking International Patent Classification and National Economic Industry Classification as examples[J]. Library and Information Service, 2016, 60(20): 123-131.

[16] Tian Chuang, Zhao Yajuan. Research progress on patent-industry mapping[J]. Library and Information Service, 2016, 60(1): 135-141.

[17] Luo Xiaoning, Zheng Naizhang. Construction of patent search formulas in sci-tech novelty search[J]. Sci-Tech Information Development & Economy, 2014, 24(18): 112-114, 119.

[18] Zhang Chen. New patent search strategy: combining keywords with classification numbers[J]. Sci-Tech Information Development & Economy, 2014, 24(13): 112-113.

[19] He Xijia, Li Wen. Search techniques for smartphone patent applications[J]. Management & Technology of SME, 2019, 12(5): 99-101.

[20] Xie Jingjing, Hong Lijuan. Discussion on rapid search techniques for pharmaceutical patent applications[J]. China Invention & Patent, 2018, 15(12): 125-128.

[21] Hoberg G, Phillips G. Text-based network industries and endogenous product differentiation[J]. Journal of political economy, 2016, 124(5): 1423-1465.

[22] Wager S, Athey S. Estimation and inference of heterogeneous treatment effects using random forests[J]. Journal of the American Statistical Association, 2018, 113(2): 1228-1242.

[23] Zhou Yuan, Liu Yufei, Xue Lan. A machine learning-based method for emerging technology identification: taking robotics as an example[J]. Journal of the China Society for Scientific and Technical Information, 2018, 37(9): 939-955.

[24] Venugopalan A S, Rai V. Topic-based classification and pattern identification in patents[J]. Technological forecasting and social change, 2015, 94(1): 236-250.

[25] Han Mei. Analysis of FinTech development status and financial innovation[J]. Economic Research Guide, 2016, 12(23): 88-90.

[26] Ye Chunqing. "FinTech" and Internet finance[J]. FinTech Time, 2016, 25(8): 88.

[27] Top 100 companies in FinTech[EB/OL]. [2019-12-21]. https://www.americanbanker.com/news/top-100-companies-in-fintech-ab107192.

[28] Xie Zongxiao. Promoting "best practices" for financial network security in the FinTech era[J]. China Information Security, 2017, 8(7): 63-66.

Author Contributions: Xu Lu: Designed research plan, FinTech classification, analyzed data, wrote paper; Lu Xiaobin: Proposed research ideas and framework, revised paper; Yang Guancan: Constructed model, processed data.

Identify and Classify FinTech Patent Xu Lu, Lu Xiaobin, Yang Guancan School of Information Resource Management, Renmin University of China, Beijing 100872

Abstract: [Purpose/significance] FinTech has developed rapidly in the information and data era, and the number of patents has continued to increase. At the same time, its field crossover and blurred borders characteristics have also increased the difficulty of patent analysis. Therefore, it is necessary to construct suitable identification and classification methods to accurately and efficiently process the continuously growing large volume of data. [Method/process] Firstly, based on the connotation and function of FinTech, this paper sorts out the innovation categories it contains and clarifies the scope and boundaries of FinTech patents. Then, it uses machine learning algorithms combined with text filtering and manual interpretation to construct a method process for FinTech patent identification and classification. [Result/conclusion] This paper proposes a patent identification and classification process based on machine learning algorithms that can more accurately and efficiently identify and classify FinTech patents. By analyzing the obtained FinTech patent classification data, it summarizes the current development status of FinTech.

Keywords: FinTech; information economics; patent analysis; machine learning

Note: Figure translations are in progress. See original paper for figures.