Research on the Determination Method of Prior Art Documents for Patent Invalidity: Postprint

Authors: Guo Shiqi, Yun Qiang, Chen Liang, Zhou Jie

Date: 2023-04-01T16:02:45+00:00

Abstract

[Purpose/Significance] Prior art documents are critical for determining patent grant or invalidation. Addressing the limitations of traditional information retrieval methods and the scarcity of machine learning-based research on prior art retrieval, this study constructs a patent relevance determination model by incorporating prior art document information. [Method/Process] Experiments are conducted using target patents and prior art documents from patent invalidation decisions as the dataset, extracting features including text similarity, co-occurring vocabulary, and co-word quantity. The GBDT model transforms the prior art retrieval problem into a classification task for determining relevance. [Results/Conclusion] Results demonstrate that different fields contribute variably to classification performance, with the specification field achieving accuracy, recall, and F1-score of 79%, 48%, and 59%, respectively. The multi-feature integrated classification significantly outperforms single text similarity-based approaches. Finally, misclassification cases are analyzed to identify future research directions.

Full Text

Research on Methods for Determining Reference Documents in Patent Invalidation

Guo Shiqi[1],[2], **Yun Qiang**[2], **Chen Liang**[2], **Zhou Jie**[2] [1] Institute of Medical Information/Medical Library, Chinese Academy of Medical Sciences, Beijing 100020 [2] Institute of Scientific and Technical Information of China, Beijing 100038

Abstract

[**Purpose/Significance**] Reference documents are crucial for determining whether a patent can be granted or invalidated. Addressing the limitations

of traditional information retrieval methods and the scarcity of research employing machine learning for reference document retrieval, this study constructs a patent relevance determination model by incorporating reference document information. [**Method/Process**] Using target patents and reference documents from patent invalidation judgments as the experimental dataset, we extract features including text similarity, co-occurring vocabulary, and co-word quantity information. The Gradient Boosting Decision Tree (GBDT) model transforms the reference document retrieval problem into a classification problem that determines relevance. [**Result/Conclusion**] Results demonstrate that different fields contribute differently to classification effectiveness. The description field achieves accuracy, recall, and F1-score of 79%, 48%, and 59% respectively. Moreover, the classification performance with multi-feature integration significantly outperforms that of single text similarity. Finally, we analyze misclassification cases to identify directions for future research.

Keywords: patent invalidity; reference document; feature selection; machine learning

Introduction

As economic globalization accelerates and technological innovation's driving role in national economic development strengthens, countries increasingly emphasize intellectual property protection for technological innovation, resulting in explosive growth in patent filings. In China alone, patent applications processed by the National Intellectual Property Administration in 2018 increased nearly 300-fold compared to 1985. In contrast, patent examination remains primarily based on "retrieval systems + manual interpretation," which is costly, inefficient, and subject to examiners' professional backgrounds and technical expertise. This not only leads to massive backlogs of pending applications but also creates vulnerabilities that can result in erroneous grants, posing significant risks for technology holders' subsequent market activities. In this demanding environment with increasingly stringent requirements for patent examination quality and efficiency, these contradictions become more pronounced.

Among the various factors affecting patent examination quality and efficiency, reference document determination represents a critical bottleneck. Reference documents are relevant files used to assess whether an invention or utility model possesses novelty and inventiveness. The ability to identify such documents has long been a key indicator of patent examiners' and practitioners' competence. According to interviews with winners of the 3rd Patent Retrieval Competition, even top domestic patent examiners find it extremely difficult to obtain an effective reference document within four hours. The challenge lies not only in patents' rich and highly specialized technical and legal content but also in the fact that such content is frequently obscured through textual transformations, hypernym-hyponym replacements, conventional technical means substitutions, and implicit disclosures for commercial or technical considerations, rendering conventional patent retrieval systems based on inverted indexes and text simi-

larity calculations inadequate.

However, the third wave of artificial intelligence has brought statistical machine learning techniques that offer possibilities for automating reference document determination. This constitutes our research theme: moving beyond the traditional information retrieval perspective on reference document determination, we employ a supervised learning approach to identify and determine reference documents. Specifically, we first obtain ground-truth data of target patents and their corresponding reference documents to create a training dataset, extract multiple features from the association information between target patents and candidate documents, and finally identify actual reference documents hidden among candidates through classification. Compared with information retrieval methods, this approach not only incorporates more effective features beyond text similarity but also enables error analysis to identify current limitations in data processing, feature engineering, and model construction, thereby providing clear directions for future optimization.

1 Related Research

1.1 Fundamental Concepts of Reference Documents Reference documents include both patent and non-patent literature, with the invention or utility model under examination referred to as the target patent. Patent retrieval can be categorized based on different stages of the patent lifecycle and retrieval purposes, including state-of-the-art searches, invalidity searches, infringement searches, and validity searches. Invalidity searches are initiated when a validity challenger questions a patent right, aiming to retrieve evidence of erroneous grants due to examination oversights or concealment of prior art, thereby reexamining the novelty of the invention. Finding invalidity evidence is critical to the success of such challenges.

Reference documents can be classified into categories X, Y, A, R, P, and E based on their relationship with claims in search reports, as shown in . Categories X and Y are most closely related to the target patent.

** Types and Meanings of Reference Documents**

Type	Meaning
X	Documents that alone affect the novelty or inventiveness of a claim
Y	Documents that, in combination with other Y-category documents in the search report, affect the inventiveness of a claim
A	Background art documents that reflect some technical features of the claim or related prior art

Type	Meaning
R	Documents filed by any entity with the patent office on the filing date that belong to the same invention
P	Intermediate documents whose publication date falls between the application's filing date and the claimed priority date
E	Conflicting application documents that alone affect the novelty of a claim

1.2 Reference Document Retrieval

1.2.1 Traditional Reference Document Retrieval Examiners or validity challengers typically conduct searches using technical keywords or combined patent classification symbols in massive databases. Although combination approaches improve retrieval efficiency to some extent, issues such as ambiguous terms and synonyms persist. Even experienced examiners employing advanced techniques like inventor/applicant information mining, reference tracing, and frontier non-patent literature tracking for efficient reference document acquisition face cumbersome processes of repeatedly constructing search queries, reading relevant documents, and determining their suitability as reference documents.

Research by K. Rajshekhar et al. demonstrated that target patents and at least 20% of highly relevant patents show no obvious similarity in technical terminology, while state-of-the-art semantic retrieval technologies can retrieve at most 10% of highly relevant prior art. Long Jin proposed obtaining reference documents from patent citations based on their similarity with target patents' technical content, offering theoretical guidance but lacking a concrete retrieval methodology. Existing approaches remain heavily dependent on retrieval experience.

1.2.2 Machine Learning Applications in Reference Document Retrieval Traditional information retrieval models rank relevance between queries and documents using manually fitted formulas based on term frequency, inverse document frequency, and document length. As more factors influence relevance, learning-to-rank has emerged as a popular domain that combines outputs from multiple ranking models to train new models with automatically learned parameters. While applying machine learning to search result ranking has become a hot research topic, no relevant applications or studies exist in the patent invalidity retrieval domain.

Domestic research primarily approaches from similar patent identification. Zhang Jie et al. proposed using subject-predicate-object structures from claim text for similar patent identification. Liu Yuqin et al. constructed a Chinese

patent invalidity retrieval model combining independent claim structural features with two-step retrieval. Traditional methods relying on text similarity or retrieval systems like cosine similarity and ElasticSearch produce limited results due to string-level comparison. Analyses based on claim text uniqueness explore patent text segmentation models, named entity recognition, and text classification to further advance patent identification and retrieval.

International research began earlier, predominantly employing machine learning methods. These include optimizing term frequency-based patent feature mining using weighted maximum confidence methods, automatic patent topic classification based on metadata and citation information, and collaborative training methods for annotating functional sentences in abstracts to improve retrieval precision. F. Kreuchauf et al. proposed patent retrieval strategies based on part-of-speech, citation, or combined methods using title, abstract, and IPC information from a small core patent dataset in service robotics. W. Ho et al. from UC Berkeley developed a program to predict PTAB patent invalidation request probability based on text similarity. L. Ryan et al. from Stanford University used metadata features such as the ratio of granted patents to company applications and examiner grant rates with convolutional neural networks for patent grant prediction, demonstrating superior performance over text-only models.

Machine learning assistance indeed outperforms traditional retrieval methods. However, existing research based on patent text, structured information, and bibliographic data requires improvements in term relevance judgment, semantic representation, and expert experience integration to enhance reference document retrieval effectiveness. This study focuses on reference document acquisition in invalidity searches, using machine learning to identify reference documents from invalid evidence databases.

2 Research Design for Automatic Reference Document Identification

2.1 Overall Framework This research transforms the reference document retrieval problem into a binary classification problem of determining relevance between reference documents and target patents in machine learning. By incorporating richer features beyond text similarity, we achieve reference document identification. The overall framework, shown in [Figure 1: see original paper], follows the process of "dataset construction → data preprocessing → feature selection and extraction → label information extraction → model testing" to explore the feasibility of using machine learning to solve manual retrieval problems. Detailed steps are provided in .

[Figure 1: see original paper] Technical Route of Patent Relevance Determination Model

** Basic Research Steps**

2.2 Data Preparation Given that invention patents offer broader protection scope, better stability, and stronger legal protection, this study focuses

on patents that underwent invalidity examination between 1990-2018, totaling 4,246 target patent samples. Each sample includes basic information such as invalidation decision number, requester, and patentee, as well as legal basis, decision points, and full text.

Patent invalidation judgments consist of legal provisions, decision points, and full text. The full text contains four parts: (1) target patent basic information; (2) reasons for invalidation request, relevant articles, and evidence attachments; (3) detailed evidence recognition results; and (4) case decision determining the patent's legal status (maintained valid, partially valid, or wholly invalid) based on evidence and patent law.

We used regular expressions to extract invalid evidence patent numbers from evidence texts provided by requesters in invalidation judgments, obtaining 21,718 numbers. These were batch-retrieved from the Wanxiangyun database to download bibliographic information as invalid evidence samples. Finally, we randomly selected 60 patents from the target patent sample pool and extracted matching invalid evidence as reference document datasets based on patent numbers, yielding 299 documents. The title, abstract, claims, and description fields constitute the most important research data. Patent fields and their meanings are shown in .

** Patent Fields and Their Connotations and Functions**

Patent Document Field	Connotation and Main Function
Title	Concisely and accurately indicates the subject matter and type of protection sought
Abstract	States the patent title and technical field, clearly reflecting the technical problem to be solved
Claims	Based on the description, clearly and concisely defines the scope of protection sought. Records technical features of the invention and serves as the basis for patent examination
Description	Clearly and completely describes the invention so that technical personnel in the field can understand and implement it. Includes technical field, background art, invention content, brief description of drawings, and detailed embodiments

Note: All descriptions in are from the Patent Examination Guidelines 2010.

2.3 Feature Extraction This step selects valuable key information and eliminates noise to enable the classifier to learn the most important text information, thereby improving performance. This study primarily uses three features: text similarity, co-occurring vocabulary, and co-word quantity. The extraction process is shown in [Figure 2: see original paper].

[Figure 2: see original paper] Feature Extraction Process

1. **Text Similarity**: We first trained TF-IDF models to obtain document vectors for each patent field (title, abstract, claims, description, and combined document), then trained LDA models to map documents to topic spaces, and finally calculated text similarity between corresponding fields.
2. **Co-occurring Vocabulary**: While text similarity is important, relevance results based solely on it are suboptimal. If target patents and reference documents share co-occurring vocabulary, they may be relevant, with more shared vocabulary indicating higher relevance probability. We extracted the vocabulary intersection between corresponding fields as co-occurring vocabulary. Rather than using unfiltered vocabulary, we employed information gain to select the top 600 vocabulary items (optimal after comparing top 50, 100, 250, 300, and 600) as a dictionary, then vectorized these terms as features. Information gain measures how much feature X reduces uncertainty about category Y, filtering vocabulary noise while reducing storage and computational burden.
3. **Co-occurring Vocabulary Quantity**: The quantity of co-occurring vocabulary between target patents and reference documents ranges from [0, ∞]. To reduce variance interference with the model's ability to learn other features, we applied MinMaxScaler from sklearn.preprocessing to normalize the quantity to [0, 1].

Finally, text similarity, co-occurring vocabulary, and normalized co-occurring vocabulary quantity features were merged into a feature matrix for this study.

2.4 Label Generation This critical step extracts text features and label vectors. Using Google's Python machine learning library scikit-learn, we split data into training and test sets at a 7:3 ratio. If a reference document was indeed submitted by the invalidation requester as evidence to prove a target patent invalid—meaning a matching relationship exists between the reference document and target patent—we set the matching label to 1; otherwise, 0. These labels serve as the gold standard for evaluating subsequent model classification performance.

2.5 Model Testing and Evaluation Various machine learning algorithms have developed rapidly and play important roles across domains, including logistic regression, hidden Markov models, and conditional random fields. This study applies the Gradient Boosting Decision Tree (GBDT) model, proposed

by J.H. Friedman in 2001, to relevance determination. GBDT is a boosting algorithm that involves calculating candidate split points, creating decision trees, finding split nodes, and computing predictions for merged leaf nodes. The process initializes predictions, then iteratively adds classification trees, obtaining predictions and residuals from new leaf nodes, learning from residuals to generate new trees until residuals between actual values and final predictions become sufficiently small. Due to its excellent performance, GBDT is widely used in data competitions and engineering applications.

We primarily use the F1-score—the harmonic mean of precision and recall—to evaluate model classification effectiveness.

3 Empirical Study

3.1 Experimental Overview This study conducted experiments using the Gensim framework with Python 2.7.13 on a Windows 7 64-bit operating system with a 16-core Intel processor and 64GB RAM. The overall technical route is shown in [Figure 3: see original paper].

[Figure 3: see original paper] Research Process of Relevance Determination Model

3.2 Data Acquisition Using the Wanxiangyun database as our data source, we randomly selected 60 patents from the target patent sample pool and manually collected basic information including application number, title, abstract, claims, and description to construct the target patent dataset. We then extracted matching invalid evidence patent numbers and downloaded their bibliographic information to construct the reference document dataset.

After processing, the target patent dataset contained 60 patents and the reference document dataset contained 299 patents.

3.3 Experimental Results After multiple parameter adjustments, the optimal performance was achieved with the hyperparameter n_{estimators}=20 (maximum iterations of weak learners). Results were evaluated using metrics A (accuracy), P (precision), R (recall), and F1, with specific meanings provided in .

** GBDT Model Evaluation Metrics and Their Meanings**

Metric	Meaning
Accuracy (A)	Proportion of correctly predicted patent pairs among all patent pairs
Precision (P)	Proportion of correctly predicted relevant pairs among all predicted relevant pairs

Metric	Meaning
Recall (R)	Proportion of correctly predicted relevant pairs among all actually relevant pairs
F1-score	Harmonic mean of precision and recall, commonly used for comprehensive evaluation

Main experimental results are shown in , where Title, Abstract, Claims, Description, and All represent patent title, abstract, claims, description, and combined document respectively. The control group used only text similarity from the combined document field.

** GBDT Classification Results Statistics**

Field	A	P	R	F1
Title	0.9863	0.5682	0.3125	0.4032
Abstract	0.9864	0.5493	0.4875	0.5166
Claims	0.9874	0.7308	0.2375	0.3585
Description	0.9903	0.7917	0.4750	0.5937
All (Experimental)	0.9894	0.7018	0.5000	0.5839
All (Control)	0.9851	0.5000	0.0625	0.1111

The experimental group using combined document fields with co-occurring vocabulary and quantity features showed nearly 5 times better F1 performance than the control group using only text similarity.

3.4 Experimental Discussion This section discusses results from three perspectives—fields, features, and errors—to identify directions for future research.

3.4.1 Field Evaluation As shown in [Figure 4: see original paper], classification effectiveness (measured by F1) when using individual fields in the GBDT model ranks as: Description (0.5937) > Combined Document (0.5839) > Abstract (0.5166) > Claims (0.3585) > Title (0.4032). The description field performs best, while the combined document field shows similar but slightly degraded performance compared to description alone, indicating that increased text volume is not a sufficient condition for improved classification. The description field carries the most detailed patent information, including technical field, background, invention content, and embodiments, reflecting richer information about relevance and uniqueness between texts.

Beyond the description field, the abstract field also demonstrates strong classification capability despite its smaller size. The abstract is a concentrated

summary of patent content (limited to 300 characters) encompassing main background and technical information. Claims, as the legal basis for patent protection scope, use numerous specialized and uncommon terms, focusing not only on technical details but also on differences in novelty and inventiveness from related patents, containing unique technical aspects.

3.4.2 Feature Evaluation This experiment employed text similarity, normalized co-occurring vocabulary quantity, and 600 co-occurring vocabulary items as features. Feature weight ranking, shown in [Figure 5: see original paper], reveals that the highest-weight features are highly specialized co-occurring vocabulary terms.

Among 64 features with weights greater than 0 (ranging from 0 to 0.072), text similarity and normalized co-occurring vocabulary quantity ranked 2nd and 23rd respectively. This indicates that not all features contribute to model classification, with highly specialized co-occurring vocabulary contributing most significantly. Multiple experiments consistently show that text similarity and most co-occurring vocabulary features have substantially greater weights than normalized co-occurring vocabulary quantity.

3.4.3 Error Analysis Error analysis examines misclassified development set samples to inspire new research directions. The test set contained 52 erroneous patent pairs: 42 false negatives (FN) where actual reference documents were judged irrelevant, and 10 false positives (FP) where irrelevant documents were judged as reference documents.

From text similarity perspective, FN and FP similarity ranges were [0.096, 0.995] and [0.041, 0.563] respectively. From co-occurring vocabulary perspective, the most frequent co-occurring terms in misclassified patents were “concrete” (0.023437) and “water” (0.000588). Error analysis results are summarized in .

** Error Analysis**

Error Cause	FN Count	FN %	FP Count	FP %
Low text similarity but high relevance upon manual reading	14	33.33%	5	50.00%
Domain experts determine different fields	37	88.10%	6	60.00%
More than 5 obvious segmentation errors	7	16.67%	2	20.00%

Error Cause	FN Count	FN %	FP Count	FP %
Patent text fluency issues (foreign patents only)	25	59.52%	5	50.00%
Common values, units, or abbreviations	14	33.33%	4	40.00%

Based on these results, improvements can be attempted in two aspects: (1) Distinguish patent domains during data preparation, as same-domain patents are more likely to share similar technical backgrounds and textual/semantic information that may affect each other's validity; (2) Improve data processing precision, such as segmentation accuracy and stop-word filtering, to minimize controllable errors.

Conclusion

This study's main contributions are: (1) Beyond text similarity at document level, we explore features at vocabulary level including co-occurring vocabulary and quantity; (2) We replace traditional information retrieval methods with machine learning classification, achieving good results with the GBDT model; (3) We provide detailed error analysis to identify misclassification causes and guide future improvements.

However, limitations remain: (1) Reference document retrieval involves two steps—first retrieving relevant documents, then determining reference documents among results. This study focuses on the second step; future work will address retrieval. (2) Current research uses a small dataset; scaling to large datasets for practical application is the next optimization direction. We plan to employ the enterprise search engine Elasticsearch to support the first-step retrieval. (3) This study uses text-based features, while patents contain rich fields like IPC, citations, and patent family information that could aid reference document identification, representing our team's future work.

Applying artificial intelligence to complex examination work is extremely challenging. We hope this study provides assistance and inspiration for academia and practice.

References

[1] National Intellectual Property Administration. 1985 Patent Statistics Annual Report [EB/OL]. [2020-08-05]. http://www.cnipa.gov.cn/tjxx/jianbao/1985-1999/85/1.1.htm. [2] BLOSSER GH, ARSHADI N, AGRAWAL S. A critical assessment of the USPTO policies toward small entity patent applications [J]. Technology and innovation, 2011, 13(3): 249-259. [3] National Intellectual Property Administration. Top 10 Patent Reexamination and Invalidation Cases

of 2018 [EB/OL]. [2020-08-05]. http://www.sipo.gov.cn/mtsd/1138630.htm. [4] National Intellectual Property Administration. Top 10 Patent Re-examination and Invalidation Cases of 2017 [EB/OL]. [2020-08-05]. http://www.sipo.gov.cn/mtsd/1123789.htm. [5] National Intellectual Property Administration. Shen Changyu emphasized improving patent examination quality and efficiency at the National Patent Examination Work Symposium [EB/OL]. [2020-08-05]. http://www.sipo.gov.cn/zscqgz/1138755.htm. [6] National Intellectual Property Administration. Press Conference on China's Intellectual Property Development in 2018 [EB/OL]. [2020-02-05]. http://www.sipo.gov.cn/zscqgz/1120594.htm. [7] State Intellectual Property Office of the P.R.C. Patent Examination Guidelines (2010) [M]. Beijing: Intellectual Property Publishing House, 2009. [8] China Patent Retrieval Skills Competition [EB/OL]. [2020-08-05]. http://www.ipsearch.top/home/index.htm. [9] Patent Reexamination Board of the National Intellectual Property Administration. Case-based Interpretation—Guidance on Patent Reexamination and Invalidation Typical Cases [M]. Beijing: Intellectual Property Publishing House, 2018: 1-446. [10] HUNT D, NGUYEN L, RODGERS M. Patent Searching: Tools & Techniques [M]. Translated by Beijing Intellectual Property Bureau, Chen Kenan. Beijing: Intellectual Property Publishing House, 2013. [11] CLARKE N S. The basics of patent searching [J]. World patent information, 2011, 33(1): 55-59. [12] LUPU M, MAYER K, TAIT J, et al. Current challenges in patent information retrieval [M]. Berlin: Springer, 2011. [13] GAO Jigang. Analysis of the role of computer keyword selection in patent retrieval [J]. Communication world, 2015(12): 257-257. [14] LU Shiyan, ZHU Jia, LI Jiao, et al. Application of tracking search in chemical patent application examination [J]. Guangdong chemical industry, 2019, 46(3): 131-132. [15] ZHU Jingjing, YANG Yi. Patent retrieval technique of "following the vine to get the melon" [J]. Science and technology guide (electronic edition), 2017(10): 218-220. [16] HUANG Wei. Retrieval and application of non-patent literature in patent examination [J]. Management & technology of SME, 2016(7): 118-119. [17] RAJSHEKHAR K, SHALABY W, ZADROZNY W. Analytics in post-grant patent review: possibilities and challenges (preliminary report) [J]. Social science electronic publishing, 2017. [18] LONG Jin. Research on patent invalidation reference documents and their acquisition—From the perspective of patent citation analysis [D]. Xiangtan: Xiangtan University, 2012. [19] ZHANG Jie, SUN Ningning, ZHANG Haichao. Chinese similar patent identification algorithm based on SAO structure [J]. Journal of the China Society for Scientific and Technical Information, 2016, 35(5): 472-482. [20] LIU Yuqin, WANG Xuefeng, LYU Lin. Chinese patent invalidity retrieval model based on claim structural information [J]. Application research of computers, 2008, 25(7): 2068-2070. [21] ZHAI Dongsheng, MA Wenshan. Research on Chinese patent claim segmentation algorithm [J]. Journal of intelligence, 2011, 30(11): 152-155. [22] MA Shuanggang. Research on Chinese patent text automatic classification based on deep learning theory and methods [D]. Zhenjiang: Jiangsu University, 2016. [23] LIAO Liefa, LE Fugang, ZHU Yalan. Application of LDA model in patent text classification [J]. Modern information, 2017, 37(3): 35-39. [24]

HU Jie, LI Shaobo, YU Liya, et al. Patent text classification model based on convolutional neural network and random forest algorithm [J]. Science technology and engineering, 2018, 18(6): 268-272. [25] GUO M, YUAN H, QIAN Y. A new method for rare feature extraction in patent documents [C]//2016 13th international conference on service systems and service management. Kunming: IEEE, 2016: 687-692. [26] ZHU F, WANG X, ZHU D, et al. User demand-driven patent topic classification using machine learning techniques [C]//The 11th conference on international fuzzy logic and intelligent technologies in nuclear science. Joao Pessoa: World Scientific, 2014: 657-662. [27] CHEN X, DENG N. A semi-supervised machine learning method for Chinese patent effect annotation [C]//2015 international conference on cyber-enabled distributed computing and knowledge discovery. Xi'an: IEEE, 2015: 243-250. [28] KREUCHAUFF F, KORZINOV V. A patent search strategy based on machine learning for the emerging field of service robotics [J]. Scientometrics, 2017, 111(2): 743-772. [29] LEE J. Predicting bad patents [EB/OL]. [2020-08-05]. https://www2.eecs.berkeley.edu/Pubs/TechRpts/2017/EECS-2017-57.pdf. [30] WINER D. Predicting bad patents: employing machine learning to predict post-grant review outcomes for US patents [EB/OL]. [2020-08-05]. https://www2.eecs.berkeley.edu/Pubs/TechRpts/2017/EECS-2017-60.pdf. [31] HO W. Predicting bad patents [EB/OL]. [2020-08-05]. https://www2.eecs.berkeley.edu/Pubs/TechRpts/2017/EECS-2017-63.pdf. [32] YE W T. Predicting bad patents [EB/OL]. [2020-08-05]. https://www2.eecs.berkeley.edu/Pubs/TechRpts/2017/EECS-2017-66.pdf. [33] RYAN L, MARCOS T. Predicting patent outcomes with text and attributes [EB/OL]. [2020-08-05]. http://cs230.stanford.edu/projects_{{{spring}}*{{2019}}}/reports/18681598.pdf. *[34] RAJSHEKHAR K, ZADROZNY W, GARAPATI S. Analytics of patent case rulings: empirical evaluation of models for legal relevance [C]//Proceedings of the 16th international conference on artificial intelligence and law. London: Elsevier, 2017: 1-9. [35] DENG Jie, YU Xiang, CUI Ligang. Empirical research on invention patent invalidation behavior in China based on patent information [J]. Journal of intelligence, 2014, 33(8): 52-58. [36] LI Hang. Statistical Learning Methods [M]. Beijing: Tsinghua University Press, 2019. [37] FRIEDMAN J H. Greedy function approximation: a gradient boosting machine [J]. Annals of statistics, 2001, 29(5): 1189-1232. [38] APACHE CN. scikit-learn (sklearn) official documentation Chinese version [EB/OL]. [2020-08-05]. https://sklearn.apachecn.org/. [39] GENSIM. Core concepts [EB/OL]. [2020-08-05]. https://radimrehurek.com/gensim/auto*{examples}/core/run_{{{core}}}_{{concepts}}}.html#core-concepts-document. [40] Wanxiangyun. Wanxiangyun patent retrieval [EB/OL]. [2020-08-05]. https://www.wanxiangyun.net/search/Index. [41] State Intellectual Property Office of the P.R.C. Patent Examination Guidelines (2010) [M]. Beijing: Intellectual Property Publishing House, 2010.

Author Contributions

Guo Shiqi: Model debugging, data analysis, initial draft writing, and paper revision. Yun Qiang: Research topic guidance, data collection, and data process-

ing. Chen Liang: Proposed overall research framework and model construction. Zhou Jie: Topic determination and paper revision.

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