Prediction of Patent Value Based on Machine Learning Algorithms

Haiying Ren*[a], Chuangchuang Sun[b]

* Corresponding author: renhaiying@bjut.edu.cn

[a,b] School of Economics and Management, Beijing University of Technology, Beijing, China

ABSTRACT: Predicting the value of patents is crucial for individuals and enterprises to make informed decisions during the patent application and commercialization process. But little research has considered the role of prior knowledge play in the prediction of patent value. This paper selects and designs variables of the knowledge network embeddedness to represent the association between focal patents' knowledge and the prior domain knowledge from the knowledge recombination perspective. Then use multiple machine learning models to predict patent value proxied by patent transfers. The feasibility of this method is tested with sample patents in the neural network field. The results show that the ExtraTrees achieves the best prediction accuracy of 84.4%.

Key words: Patent value prediction; Knowledge network embeddedness; Machine learning; ExtraTrees

1. INTRODUCTION

Predicting the value of patents is crucial in today's fiercely competitive global science and technology landscape. With the development of computer technology, data mining and machine learning, predicting the value of patents has becoming an important tool in patent management and application. These technological advances have significantly improved the accuracy and efficiency of prediction. According to the knowledge recombination theory, new technology is recombination of existing knowledge developed from prior technology [1][2][3]. However, the relationship between patent value and prior knowledge, particularly the texts of technical documents, has not been fully explored in current research. Further research in this area could lead to a better understanding of the value of patents and how they can be leveraged in the development of intellectual property, and enable more accurate patent value prediction.

A patent comprises a compilation of knowledge elements that are conveyed through a blend of specialized terminology and specific concepts. As a result, the content of the patent text, especially the abstract section, carries immense significance in determining the patent's value [4]. The patent abstract usually encompasses details such as the technical scheme, technical effect, application field, market prospects, and other pertinent information. These factors serve as a pivotal basis for evaluating the technical and commercial value of the patent. The analysis and extraction of valuable information from patent text can enable the prediction of the patent's worth, thereby facilitating informed decisions concerning patent investments and technology transfers.

ICIDC 2023, June 02-04, Nanchang, People's Republic of China

DOI 10.4108/eai.2-6-2023.2334602

Numerous studies have been conducted to evaluate the value of patents [5][6], with some scholars utilizing patent indicator systems for this purpose [7]. With the advancements in machine learning and natural language processing technologies, research on predicting patent value has emerged [8]. Tools such as BERT are utilized to transform text data into directly computable structured data, aiding in the prediction of patent value [9]. However, scholars have not often explored the relationship between the value of patents and prior textual knowledge.

In this study, we draw on the knowledge recombination theory to posit that there is a significant relationship between patent value and prior knowledge; Patent abstracts are not merely random word combinations, but instead contain valuable word patterns that can be extracted from historical technical data to collect indicators or variables related to the value formation process. Aided by complex network theory, knowledge networks are constructed, and metrics are designed to generate predictive models of patent value using various machine learning algorithms. Given that neural networks have demonstrated impressive results in various fields of emergent technologies, this paper uses patent data in the field of neural networks to validate the proposed method.

2. METHODOLOGY AND PROCEDURE

We present a new method for predicting patent value that takes into account the relationship between a sample patent's knowledge embeddedness in its prior knowledge. Figure 1 illustrates the overall process of our proposed approach. The approach consists of the following steps: collecting and preprocessing patents, extracting IPO structure and construct knowledge network, designing and calculating knowledge network embeddedness, and building machine learning models to predict patent value.

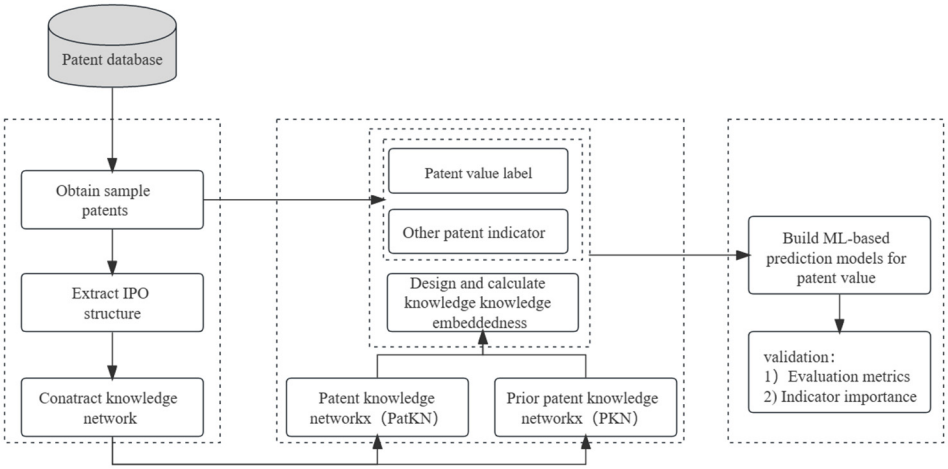

Figure 1. Framework for predicting the patent value

2.1 Obtain sample patents

The IncoPat patent database was the data source for this study. The search formula employed for neural network technology was "IPC-LOW=(G06N3/02)", and the search scope was limited to patents from 1973 to 2012. Given that the abstracts of patents contain critical information, we selected the abstracts of these patents as the research object. The abstracts were preprocessed to exclude irrelevant information, ensuring that only pertinent data was utilized in our analysis.

2.2 Extract IPO structure and construct knowledge network

Initially, the patent abstract text is parsed using Python and its natural language processing tools to generate a phrase structure tree. Next, according to the research [10], utilize regular expressions to isolate verbs, prepositions, and conjunctions from the phrase structure tree and combine adjacent process elements (P). Starting from the root node of the phrase structure, the nearest parent and child nodes corresponding to the NP blocks are identified according to the P elements. These are respectively denoted as NP1 and NP2, and categorized as either input element (I) or output element (O). The triplet of IPO structure is then represented as [NP1, P, NP2]. The IPO structure is standardized through parts of speech restoration, lowercase conversion, and removal of stop words, among other techniques. By obtaining the IPO structure of each individual sentence in the abstract, the IPO list of a single patent can be obtained. The repeated IPOs of a single patent can be removed, and the elements I and O are treated as nodes of the knowledge network, while the elements P serve as the edges of the network. Assuming that a patent, denoted as *i,* was filed in a given year *t*, this procedure constructs a patent knowledge network (PatKN) for patent *i*. Finally, the prior knowledge network (PKN) for *i* is constructed in the above manner using all patents published in the focal field in the years up to year *t* - 1.

2.3 Design and calculation knowledge network embeddedness

The network embeddedness of patent *i* refers to the network characteristics of its PatKN in the PKN. This approach aims to measure the "position" and "relationship" of the knowledge elements of *i* in its PKN. In this study, patents applied in the year 2012 are randomly selected as sample patents and a patent knowledge network is constructed for each patent. Then, all patents that were made public prior to 2012 are used to construct the PKN for all the sample patents, which is then used to embed each sample PatKN.

Previous studies have demonstrated that novel knowledge elements and high-importance conventional knowledge elements are likely to impact patent value. Furthermore, the distribution of knowledge elements in the knowledge network is non-uniform, and the network structure features, such as centrality, reflecting the number of direct connections with other nodes, indicates the potential for this knowledge element to be combined with other knowledge elements. In light of these findings, we define S_i as the PaKN of patent *i.* We then design and calculate the knowledge network embeddedness.

1) *n_recent.* The recent node means that the application year of a node in S_i is greater than or equal to 60% of the public year of all nodes in the previous knowledge network. It is used to represent novelty. Literatures have confirmed that patents with relatively recent nodes tend to have higher value[11]. In this paper, this embeddedness feature is expressed as the arithmetic square root of the number of recent nodes.

$$n\_recent = \sqrt{number\ of\ recent\ nodes\ in\ s_i}$$

2) *e_novel*, *n_novel*. Novel edges are the edges that do not exist in the knowledge network but exist in S_i, which reflects the frontier of technology development and has a positive impact on the value of patents. In this paper, the arithmetic square root is used to represent the number of novel edges.

$$e\_novel = \sqrt{number\ of\ novel\ edeges\ in\ s_i}$$

n_novel is treated similarly to *e_novel*.

3) *e_range_mean.* The year span of the edge refers to the difference between the oldest and earliest years of the edge. This variable is an indicator of conventionality, and the edge with a long year span has a relatively lasting relationship in the field, indicating that the technical content is practical. The average span of edge years in S_i is utilized as a representation of this embeddedness feature.

4) *eigen_mean.* Eigenvector Centrality reflects the structural importance of nodes in S_i. The higher the eigenvector centrality a node has, the more important its position in knowledge network; the more core nodes S_i has, the more significant impact it has on patent value [12]. In this paper, the average eigenvector centrality of nodes in S_i are retained as an embeddedness feature.

5) *pr_mean*. Some studies demonstrate that the PageRank value of patents in citation network is an effective indicator to predict patent value [13]. Therefore, this paper takes it as an variable that may affect the value of the patent, and uses the average PageRank value of the nodes in S_i as an embeddedness feature.

6) Degree_mean. The degree centrality refers to the number of links connected to a node, and the larger the degree of a node, the higher its degree centrality, which means it is more important in the network. In this paper, the mean of the degree centrality in S_i is used to represent it.

2.4 Build ML-based prediction models for patent value

In this stage, the predictor, or input variables for the patent value prediction model are collected based on the knowledge network embeddedness in Section 2.3 together with the number of inventors, the number of applicants, the number of claims, the coverage of technology, whether there is an agency, and the number of literature pages. Patent transfer is taken as the predicted, or output variable. Table 1 shows the input and output variables, with embeddedness variables in bold.

Table 1. Variables and measurement method for patent value prediction

Property	Patent variable	Measurement method
Input variables	*family country*	number of countries
	claims	Number of claims
	Applicant	Number of applicants
	pat_back_cite	Number of backward citations
	Agency	1 if a patent has agency; 0 otherwise
	Page	Number of literature pages
	inven_num	Number of inventors

	IPC_4	Number of IPC_4
	degree_mean	**Mean node degree in S_i**
	PR_mean	**Mean node PageRank in S_i**
	e_range_mean	**Mean time range of edges in S_i**
	eigen_mean	**Mean node eigenvector centrality in S_i**
	non_pat_back_cite	**Non-patent backward citations**
	n_recent	**Number of recent nodes in S_i**
	e_novel	**Number of novel edges in S_i**
	n_novel	**Number of novel nodes in S_i**
Output variable	*patent transfer*	1 if a patent has been transferred; 0 otherwise

1045 patents applied in 2012 were taken as the data set, among which 349 patents had transfer and 696 patents had not transfer. The training set and the test set are randomly divided according to 7:3, five-fold cross-validation is used. To prevent imbalance in the distribution of output categories, we employed SMOTE undersampling for negative samples. This technique ensures that the number of negative samples is reduced while maintaining the balance between positive and negative samples.

The patent value prediction model is built with various machine learning algorithms such as logistics regression (LR), Support Vector Machine (SVM), decision tree (DT), random forest (RF), eXtremeGradient Boosting (XGB), Extremely randomized trees (ExtraTrees) and Artificial Neural Network (ANN), using the scikit-learn library in Python. The model parameters with the best performance on the test set were obtained by using mesh parameter tuning.

3. RESULTS AND EVALUATION OF MACHINE LEARNING METHODS

3.1 Comparison of experimental results

The key evaluation metrics of prediction models constructed by the machine learning algorithms were examined separately. We chose accuracy, precision, recall, F1-score, and AUC as the evaluation metrics. Table 2 presents the results, which clearly indicate that, among all machine learning models, ExtraTrees outperformed other algorithms across all evaluation metrics. Consequently, we selected ExtraTrees to predict patent value, and Figure 2 depicts the ranking of variable importance.

Table 2. Key metrics of patent value prediction models

	Test set						
Evaluation metrics	**LR**	**SVM**	**DT**	**RF**	**XGB**	**ExtraTrees**	**ANN**
accuracy	0.752	0.787	0.818	0.838	0.825	0.844	0.828
precision	0.653	0.698	0.802	0.742	0.733	0.754	0.751
recall	0.606	0.679	0.633	0.817	0.780	0.817	0.792
f1	0.629	0.688	0.708	0.777	0.756	0.784	0.789
auc	0.789	0.847	0.792	0.901	0.886	0.902	0.834

3.2 Variables importance ranking

The height of the blue bars in Figure 2 explains the degree to which each variable contributes to patent value. It is evident that certain metrics, such as the number of patenting countries (regions), claims, applicants, agency presence, pages, inventors, and novel edges, have a significant impact on the model's overall efficacy. Furthermore, values of weighted degree centrality, PageRank, and eigenvector centrality are associated with an increased likelihood of the given sample being classified as a high-value patent.

This discovery strengthens the viewpoint that in patent abstract texts, phrases can largely represent the content of the patent, and a new patent is created through the recombination and development of prior knowledge. The knowledge elements represented by novel edges/nodes reflect the innovation point and higher level of technological innovation of the patent, and are therefore more likely to be high-value patents. Additionally, recent and novel knowledge elements may have a significant impact on the growth of the knowledge network.

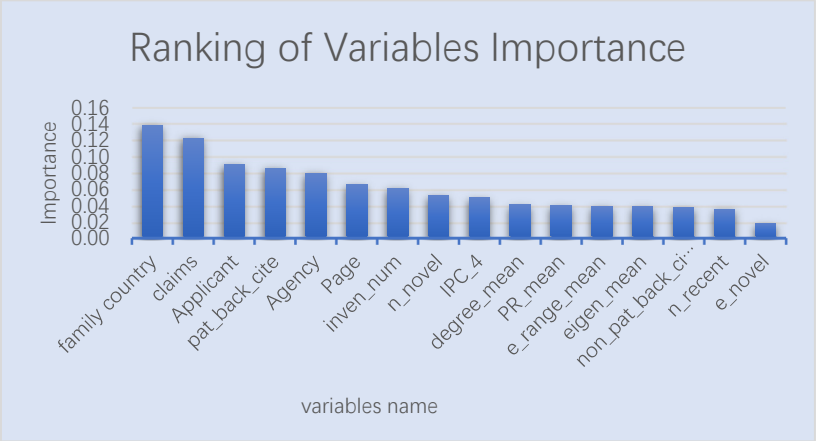

Figure 2. Variables importance ranking based on ExtraTrees

4. CONCLUSIONS

Accurate prediction of patent value is crucial for informed decision-making in patent application, commercialization, and management. However, the patent value prediction system based on patents' filing information is in early stage. This study incorporated knowledge network embeddedness by establishing the relationship between sample patent knowledge (PaKN) and prior domain knowledge (PKN), applied several machine learning models to predict the patent value and used neural network patents for validation. The results revealed that the ExtraTrees outperformed others, achieving an accuracy of 84.4%, and provides a new perspective of patent value research. The indicator importance demonstrated that the proposed knowledge network embeddedness was a new contributing factor to the prediction of patent value.

As future work, we plan to test the universality of the model by using data from other domains. This will help to evaluate the model's performance and applicability in diverse patent-related scenarios.

REFERENCES

[1] Jung H J, Lee J J. The quest for originality: A new typology of knowledge search and breakthrough inventions[J]. Academy of Management Journal, 2016, 59(5): 1725-1753.

[2] Savino T, Messeni Petruzzelli A, Albino V. Search and recombination process to innovate: a review of the empirical evidence and a research agenda[J]. International Journal of Management Reviews, 2017, 19(1): 54-75.

[3] MA R K, TAO X L, LI S M, et al. Network Search for Knowledge Elements and the Creation of Breakthrough Inventions: the Moderating Roles of Inventor Team Characteristics[J]. Studies in Science of Science, 2021, 39(5): 794.

[4] Han E J, Sohn S Y. Patent valuation based on text mining and survival analysis[J]. The Journal of Technology Transfer, 2015, 40: 821-839.

[5] Fischer T, Leidinger J. Testing patent value indicators on directly observed patent value—An empirical analysis of Ocean Tomo patent auctions[J]. Research policy, 2014, 43(3): 519-529.

[6] Grimaldi M, Cricelli L. Indexes of patent value: a systematic literature review and classification[J]. Knowledge Management Research & Practice, 2020, 18(2): 214-233.

[7] Li J, Li B, Fang H, et al. Evaluation of Invention Patent Value Based on AHP-Entropy Weight Method-Taking Toyota's Open-source Patent as an Example[J]. Journal of Intelligence, 2020, 39(5): 59-63.

[8] Mariani M S , Medo, Matú, Lafond F . Early identification of important patents: Design and validation of citation network metrics[J]. Technological Forecasting and Social Change, 2019:S0040162517314312.

[9] Chung P, Sohn S Y. Early detection of valuable patents using a deep learning model: Case of semiconductor industry[J]. Technological Forecasting and Social Change, 2020, 158: 120146.

[10] Ren Haiying, Li Zhen. Research on Identification Method of Core Technology Chain Based on an Input-Output SAO Network: A Case of Quantum Computing Technologies[J]. Library and Information Service, 2021, 65(19): 117-129.

[11] Strumsky D, Lobo J. Identifying the sources of technological novelty in the process of invention[J]. Research Policy, 2015, 44(8): 1445-1461.

[12] Caviggioli F, Scellato G, Ughetto E. International patent disputes: Evidence from oppositions at the European Patent Office[J]. Research Policy, 2013, 42(9): 1634-1646.

[13] Yang G C, Li G, Li C Y, et al. Using the comprehensive patent citation network (CPC) to evaluate patent value[J]. Scientometrics, 2015, 105(3): 1319-1346.