Rapid Patent Quality Evaluation Method Based on Big Data Analysis: Chinese Invention Patents as Sample

Yide Yang[1,a,*]

[1]Faculty of Science, University of Alberta, Edmonton, AB, T6G2R3, Canada
a. yide1@ualberta.ca
**corresponding author*

Abstract: High-quality patents have high technical value and market competitive advantage. Faced with the huge number of patent data, how to rapidly and efficiently identify the quality of patents from the patent announcement is a crucial research issue at present. Therefore, it is reasonable to predict that, big data based techniques will be the best method to exploit this kind of data. The patents authorized by CNIPA (China National Intellectual Property Administration) are taken as the research object. This study chooses several types of patent evaluation indicators and uses EWM (The Entropy Weight Method) to calculate the weight of each indicator. The study determines a correction coefficient to enhance the usability and provides the final quality score of each patent. The evaluating formula is provided. In this study, easily accessible patent indicators are used, which makes it easier to evaluate the quality of patents. By this method, rapidly evaluating the patent quality only by its basic announcement data is feasible, which solves the limitation that laborious access to advanced indicators.

Keywords: patent quality, big data, the entropy weight method, evaluation indicator

1. Introduction

With the increasing importance of intellectual property rights' economic status and strategic position globally, the number of authorized patents has increased year by year. However, patents vary substantially in their qualities, from major breakthroughs to negligible improvements [1]. Of the huge number of patents, only a small fraction has high quality and plays a major part in the development and contest in their fields.

The research on high-quality patent distinguishing is minimal. According to the existing research, the identification of high-quality patents mostly relies on expert qualitative analysis [2] and the manual use of statistical methods and quantitative measurement models [3]. Some scholars have proposed methods to evaluate patents by building index systems, and some patent-database have their own evaluation analysis systems but no evaluation method is provided. Those methods are costly for individuals and small businesses with the fee, time, and effort. Even some methods do not consider that some indicators have negative correlation rather than positive correlation at all and provided an unreliable formula. With the increasing number of authorized patents, the cost of screening high-quality patents with strong influence among a large number of patents will be incalculable.

Based on this problem, rapidly and efficiently identifying the quality of patents from the patent announcement without extra cost is a crucial research issue at present.

This study analyzes several types of collected patent indicators and uses the entropy weight method to calculate the patent quality score. The provided method has the ability to rapidly and efficiently evaluate the quality of patents, which makes individuals and small businesses distinguish high-quality patents only from the patent announcements possible.

2. Methodology

2.1. Indicator Selection and Data Collection

Data of patent quality calculation indicators used in this study are from the IncoPat global patent database, which contains worldwide patent numeral and text data, legal status, and extended information. Eight important indicators for patent quality are selected to evaluate patent quality in four dimensions: patent claims, patent participants, patent citations patent families and patent life. All of the eight indicators are accessed easily on the patent announcement.

2.1.1. Patent Claims

The patent first claim contains the recognition of the technological and commercial value of the patent, the fewer words in the first claim, the larger the patent protection range [4]; the number of patent claims is in direct proportion to the quality of patents [5], which can reflect the quality of patents.

2.1.2. Patent Participants

Under normal conditions, the number of patent inventors is in direct proportion to the patent value [6].

2.1.3. Patent Life

According to the first definition of high-value invention patents by CNIPA in March 2021, the service life of more than 10 years is one of the characteristics of high-value patents.

2.1.4. Patent Citations

Forward citations in the patent literature have frequently been used as a measure of patent value [7], and backward citations show the probability that the technology be widely promoted [8].

2.1.5. Patent Families

The number of patents in the patent family is an important indicator of the private value of patents, family size is positively correlated with patent or firm value [9]. There are 6 types of patent families, simple families and extended families are used in this study.

2.2. Research Method

2.2.1. EWM

Entropy is a thermodynamic concept, and is now widely used in many other fields, the entropy value in statistics indicates the relative importance of a parameter [10]. Generally, for an index, the larger the entropy, the more chaotic the data is, the less information it carries, the smaller the utility value, and therefore the smaller the weight.

The concrete steps of EWM are shown below:

Label the mth set of indicators with nth set of data as $X_1, X_2, X_3, \ldots, X_m$, where $X_i = \{x_{i1}, x_{i2}, x_{i3}, \ldots, x_{in}\}$. Normalize the data firstly, then, according to EWM (The Entropy Weight Method), the entropy of Xi is:

$$E_i = -\frac{\sum_{j=1}^{n} \ln(p_{ij})}{\ln(n)} \tag{1}$$

where

$$p_{ij} = \frac{x_{ij}}{\sum_{j=1}^{n} x_{ij}} \tag{2}$$

The weight of each indicator can be calculated based on its entropy:

$$W_i = \frac{1-E_i}{m-\sum_{i=1}^{n} E_i} \tag{3}$$

It is easy to reach this purpose by using the SPSSAU program.

The final score of each patent is:

$$Z_i = \sum_{i=1}^{m} X_i * W_i \tag{4}$$

2.2.2. Weight Correction

Generally, indicators 'Words in First Claim', 'Backward Citations', 'Patent Families'. etc. are in the same hierarchy [11], 'Simple Patent Families' and 'Extended Patent Families' are belonging to the parent indicator 'Patent Families'.

For the purpose of 'rapid', this study does not use the 'Analytic Hierarchy Process' to make a more complex model, but 'Simple Patent Families' and 'Extended Patent Families' are two of the important information shown on the CNIPA search, EWM might cause the weights of 'Simple Patent Families' and 'Extended Patent Families' that belongs to the parent indicator 'Patent Families' are overestimated. To solve this problem, since there are 6 type of patent families in the parent indicator 'Patent Families', the weights of 'Simple Patent Families' and 'Extended Patent Families' will be divided by 3 after the EWM. The sum of weight after correction will be less than 1, but it will not affect the determination of the quality level.

2.2.3. Quality Level

According to the report in recent years of CNIPA, the proportion of high-quality patent in all patents is slightly higher than 50%. After considering both the average and median score, a theoretical score of high-quality patents will be determined. Patents above and below this score will be recognized as high-quality and low-quality patents

3. Empirical Analysis

3.1. Data Collection

Since the effective patent life is a core indicator to evaluate, this study chooses patent data in the IncoPat database in 2007 with the IPC number H04L, which is the field of Digital information transmission.

8 patent indicators are downloaded from the IncoPat database. A total of 3150 sets of data have been obtained. Generate the patent indicator matrix as Table 1.

Table 1: Quality Indicators Matrix (Part).

Patent Sample	Words in First Claim	Number of Claims	Patent Inventors	Patent Life	Forward Citations	Backward Citations	Simple Patent Families	Extended Patent Families
1	313	25	2	2.33	4	0	4	40
2	275	8	7	10.00	4	0	29	29
3	168	28	1	6.00	3	0	11	12

3.2. Data Preprocess

Reverse processing (NMMS) the data of ‘Words in First Claims’ indicators since the words in the first claim is negatively correlated with patent quality. Transfer the unit of patent life value from months to years (round down). The patent life in IncoPat is the time from application to invalidation, this study assigns the value of 30 years to all the still valid patents to highlight the stability of these patents. Then, normalize the indicators (MMS) matrix to range (0, 1). A part of the preprocessed indicators matrix is shown in Table 2.

Table 2: Preprocessed Indicators Matrix (Part).

Patent Sample	Words in First Claim (NMMS)	Number of Claims (MMS)	Patent Inventors (MMS)	Patent Life (MMS)	Forward Citations (MMS)	Backward Citations (MMS)	Simple Patent Families (MMS)	Extended Patent Families (MMS)
1	0.9505	0.1165	0.0233	0.1481	0.2857	0	0.0162	0.0244
2	0.9580	0.0340	0.1340	1.0000	0.2857	0	0.1513	0.0174
3	0.9791	0.1311	0	0.5556	0.2143	0	0.0541	0.0064

3.3. Positive Analysis

Using the SPSSAU program to calculate the weight of each indicator by EWM, then, correct the weight of ‘Simple Patent Families’ and ‘Extended Patent Families’. The summary results are shown in Table 3.

Table 3: Indicators Weight Summary Table.

Patent Sample	Mean Value	Standard Deviation	Information Entropy	Information Utility Value	Indicators Weight (%)
Words in First Claim (NMMS)	0.943	0.054	0.9998	0.0002	0.09
Number of Claims (MMS)	0.063	0.063	0.9672	0.0328	12.49
Patent Inventors (MMS)	0.041	0.048	0.9552	0.0448	17.09
Patent Life (MMS)	0.491	0.299	0.9781	0.0219	8.36

Table 3: (continued).

Forward Citations (MMS)	0.253	0.113	0.9885	0.0115	4.39
Backward Citations (MMS)	0.003	0.021	0.9657	0.0343	13.06
Simple Patent Families (MMS)	0.045	0.072	0.9434	0.0566	7.20
Extended Patent Families (MMS)	0.009	0.041	0.9399	0.0601	7.64

Using the weight provided to calculate the Final quality score of each patent, a part of the results is shown in Table 4.

Table 4: Final Score (Part).

Patent Sample	Quality Score
1	7.461
2	7.758
3	6.161

The average score of 3150 patent samples is 5.028 and the median score is 3.913, the dividing score between high-quality and low-quality patents is determined as 3.9.

The difference between average and median score comes from the sharp rise in scores caused by abnormal data that has a huge patent family, this cannot be avoided by weight correction. The value of these two indicators is in the range (0,20) and (0,30), the average value of these two indicators is 9.241 and 15.602, but some of the patent have abnormal data about more than 100. A set of typical comparison is shown in table 5.

Table 5: Typical Comparison Between Normal Data and Abnormal Data (Part).

Patent Sample	Words in First Claim	Number of Claims	Patent Inventors	Patent Life	Forward Citations	Backward Citations	Simple Patent Families	Extended Patent Families	Score
8	299	25	4	10	3	0	32	111	15.827
851	239	27	2	10	4	0	20	24	8.214

These two patents are both stable and progressive, but their score has a huge difference due to the number of patent families. Some patents have more than 100 inventors, that is also abnormal data.

The patent quality calculation formula is determined as:

$$Z = a * 0.09 + b * 12.49 + c * 17.09 + d * 8.36 + e * 4.39 + f * 13.06 + g * 7.2 + h * 7.64 \quad (5)$$

Table 6: Coefficient Comparative Table.

Words in First Claim	Number of Claims	Patent Inventors	Patent Life	Forward Citations	Backward Citations	Simple Patent Families	Extended Patent Families
a	b	c	d	e	f	g	h

4. Conclusion

Using EWM, this study provides a feasible formula to rapidly and efficiently evaluate the quality of patents, and determined the dividing score between high-quality and low-quality patents.

The indicators used in this formula are all accessible from patent announcements from CNIPA, the problem of high-cost to for individuals and small businesses to distinguish the high-quality and low-quality patents. For the patents in the country other than China, this formula is also quite feasible.

This study also introduced a new direction of using EWM to research patent quality to the followings.

For some patent which has some abnormal huge value of indicators, this formula may be unable to distinguish the patent quality accurately, can use the average value of that indicator to estimate the patent quality.

Also, with the establishment of more reliable databases and the analysis of long-term data, the weight provided in this study might be improved.

Acknowledgment

I thank my classmate Mr. Jinglong (University of Alberta) for judgments that improved this paper. I would also like to extend my thanks to the SPSSAU program for the easy use programing.

References

[1] *Shu, T., Tian, X., & Zhan, X. (2022). Patent quality, firm value, and investor underreaction: Evidence from patent examiner busyness. Journal of Financial Economics, 143(3), 1043-1069. https://doi.org/10.1016/j.jfineco.2021.10.013*

[2] *Wu, J., Gui, L., & Liu, P. (2022). Indicator and Textual Features-Based Patent Evaluation with Graph Convolutional Networks [J]. Journal of Intelligence, 41(1):88-95,124. https://doi.org/10.3969/j.issn.1002-1965.2022.01.014*

[3] *Donato, C., Lo Giudice, P., Marretta, R., Ursino, D., & Virgili, L. (2019). A well-tailored centrality measure for evaluating patents and their citations. Journal of Documentation, 75(4), 750–772. https://doi.org/10.1108/jd-10-2018-0168*

[4] *Marco, A. C., Sarnoff, J. D., &amp; deGrazia, C. A. W. (2019). Patent claims and patent scope. Research Policy, 48(9), 103790. https://doi.org/10.1016/j.respol.2019.04.014*

[5] *Lanjouw, J. O., & Schankerman, M. (2004) Patent quality and research productivity: Measuring innovation with multiple indicators. The Economic Journal, 114(495), 441-465.*

[6] *Ernst, H. (2003). Patent information for strategic technology management. World Patent Information, 25(3), 233–242. https://doi.org/10.1016/s0172-2190(03)00077-2*

[7] *Hall, B., Thoma, G., & Torrisi, S. (2007). The market value of patents and R& D: Evidence from European firms. https://doi.org/10.3386/w13426*

[8] *Cativelli, A. S., Pinto, A. L., & Sanchez, M. L. L. (2020). Patent value index: Measuring Brazilian green patents based on family size, grant, and backward citation. Iberoamerican Journal of Science Measurement and Communication, 1(1), 004-004. https://doi.org/10.47909/ijsmc.03*

[9] *Harhoff, D., Scheerer, F., & Vopel, K. (2002). Citations, family size, opposition and value of patent rights. Research Policy, 32(8), 1343-1363. https://doi.org/ 10.1016/S0048-7333(02)00124-5*

[10] *Jianjun, Yang., Shanshan, Xing., Ruizhi, Qiu., Yimeng, Chen., Chunrong, Hua., & Dawei, Dong. Mathematical Problems in Engineering Decision-Making Based on Improved Entropy Weighting Method: An Example of Passenger Comfort in a Smart Cockpit of a Car. MATHEMATICAL PROBLEMS IN ENGINEERING. https://doi.org/ 10.1155/2022/6846696*

[11] Heng, Deng., & Zhaoyang, Dong. (2021). Research on Identification and Evaluation Model and Application of High-Quality Patents —Research Paradigm Based on Patent Information Analysis. China Invention & Patent, 18(11), 3-9. https://doi.org/10.3969/j.issn.1672-6081.2021.11.001