ChinaRxiv

AI translation · View original & related papers at
chinarxiv.org/items/chinaxiv-202310.03149

Research on Mapping Patent Documents to Industrial Categories: A Case Study of 2015 Chinese Academy of Sciences Patents and the Strategic Emerging Industries Classification Postprint

Authors: Tian Chuang, Zhao Yajuan

Date: 2023-10-08T00:00:00+00:00

Abstract

[Purpose/Significance] This study proposes a mapping model for patent literature and industry categories based on cosine similarity. The model exhibits advantages in accuracy, efficiency, and extensibility, thereby providing a valuable reference for subsequent research. [Method/Process] Existing methods for mapping patents to industry categories are systematically reviewed. Using the 2015 invention patents from Chinese Academy of Sciences institutes and the "Strategic Emerging Industry Classification" as case studies, a mapping model for patent literature and industry categories is designed, mapping experiments are conducted, and the model is evaluated based on the experimental results. [Results/Conclusion] The mapping model for patent literature and industry categories automatically generates mapping combinations between patent literature and industry categories through natural language processing technology, enabling bidirectional mapping from patents to industries and from industries to patents. The model significantly reduces labor costs and facilitates convenient fine-grained adjustments of mapping categories, making it applicable to most patent-to-industry category mapping scenarios. Finally, the limitations of the model are identified, and potential directions for future application expansion are discussed.

Full Text

Preamble

Research on Mapping Patent Documents and Industrial Categories
A Case Study of 2015 CAS Patents and the Classification of Strategic Emerging Industries

ChinaRxiv 中X

Tian Chuang, Zhao Yajuan
National Science Library, Chinese Academy of Sciences, Beijing 100190

Abstract

[Purpose/significance] This paper proposes a mapping model based on cosine similarity for mapping between patent documents and industrial classification. This model is accurate, efficient and scalable, which provides references for further research. [Method/process] After reviewing existing methods for mapping between patents and industrial classification, we designed a model for mapping between patent documents and industrial classification and completed the mapping between the 2015 annual patents of Chinese Academy of Sciences and the Classification of Strategic Emerging Industries. Then we evaluated this model according to the mapping results. [Result/conclusion] This model obtains mapping results between patent documents and industrial classification automatically through natural language processing technology and enables bidirectional mapping between patents and industries. The method saves substantial labor costs, facilitates convenient adjustment of mapping granularity at the category level, and is applicable to most patent-industry category mapping scenarios. Finally, we describe improvements needed for the model and briefly discuss potential future application areas.

Keywords: patent document; industry classification; classification mapping; mapping methods
Classification Number: G254.11

Introduction

As a product of technological innovation, patents are closely related to industrial innovation levels. Simultaneously, as an important tool for market competition, they are intimately connected with industrial economic activities. Effectively mapping patents to industrial classification systems is an indispensable step in technology transfer and patent intelligence research.

Based on a review of existing methods for mapping patents to industrial categories, this study establishes guiding principles for an accurate, efficient, and scalable model. Using 2015 invention patents from Chinese Academy of Sciences (hereinafter referred to as "CAS") institutes and the Classification of Strategic Emerging Industries as a case study, we propose a cosine similarity-based mapping model for patent documents and industrial categories, conduct mapping experiments, and comprehensively evaluate the model based on the results. The model automatically generates mapping combinations between patent documents and industrial categories through natural language processing technology, enabling bidirectional mapping from patents to industries and from industries to patents. It can save substantial labor costs and facilitate convenient adjustment of mapping granularity at the category level, making it applicable to most patent-industry category mapping scenarios. Finally, we identify areas for im-

provement and discuss potential application areas after refinement, providing references for future research.

1 Existing Methods for Patent-Industry Category Mapping

Currently, there are three primary methods for mapping patents to industries: expert judgment-based mapping, cross-retrieval-based mapping, and probability calculation-based mapping [1].

The expert judgment-based mapping method primarily relies on subjective expert evaluation to establish correspondences between categories. While this approach achieves relatively high accuracy, it is time-consuming and labor-intensive, excessively dependent on manual judgment, and unsuitable for large-scale datasets.

The cross-retrieval-based mapping method involves using one classification system to retrieve documents organized by another classification system in a corpus. By analyzing and statistically examining the categories identified in the retrieval results, it establishes mappings between the two classification systems [2-3]. This method has several limitations: first, it requires a substantial data volume, as insufficient data leads to low coverage; second, cross-retrieval yields one-to-many mappings that require further statistical analysis or manual intervention to determine final mapping relationships.

The probability calculation-based method decomposes the overall concept of a classification category into several sufficiently small unit concepts. The similarity between overall concepts can then be built upon the similarities of these unit concepts, with the probability between entire categories derived by calculating similarities between sub-categories, where the sum of sub-category probabilities should equal the overall probability. Unit concepts are typically represented by keywords, transforming category similarity into the sum of similarities of words expressing these unit concepts [4-5]. This method uses computers to replace manual semantic matching based on specific rules, saving time and effort, though mapping results still require some manual adjustment.

2 Guiding Principles for the Mapping Model

Existing domestic patent-industry mapping approaches rely excessively on manual judgment [6-7], lack universality, and require improvement in both methods and results. An ideal mapping method should satisfy accuracy requirements while ensuring efficiency and scalability. Based on this direction, we establish the following guiding principles:

2.1 Accuracy

We use title and abstract information from patent documents as patent features and official annotations from industrial categories as category features. After initial word segmentation, we extract verbs and nouns that more precisely reflect

the characteristics of patents and industries, and remove stop words without distinctive features to ensure mapping accuracy.

2.2 Efficiency

The 2012 edition of the Classification of Strategic Emerging Industries contains 100 categories at the third level [8]. Manually mapping each patent document to industrial categories would be extremely time-consuming. Therefore, we should maximize automation of the mapping process through computer technology to reduce manual involvement. Our model uses computer programming to rapidly extract features from both patent documents and industrial categories and calculate their similarity. It can be continuously updated as new patents and industrial classifications emerge, and allows efficient, flexible adjustment of the part-of-speech and quantity of extracted category feature words based on mapping results, relying on objective algorithms rather than subjective human judgment.

2.3 Scalability

In addition to exploring the mapping between patent documents and industrial categories, this mapping model focuses on constructing a comprehensive and rigorous mapping workflow to facilitate future expansion to other industrial classification systems and adjustment of category granularity.

2.3.1 Expansion to Other Industrial Classifications In addition to the Classification of Strategic Emerging Industries, China currently employs several other industrial classifications, including the National Economic Industry Classification, High-tech Industry Classification, Ten Key Industries Classification, Statistical Product Classification Catalog, and Industrial Structure Adjustment Catalog. Existing mapping achievements are primarily based on expert judgment and only cover partial industrial classifications [9-10]. When new versions of industrial classifications are released, re-mapping typically requires substantial human and material resources. To avoid redundant work, our model emphasizes designing a comprehensive mapping workflow that automatically extracts feature words from official category annotations through computer technology, enhancing scalability while maintaining accuracy and facilitating mapping to other industrial classifications.

2.3.2 Facilitating Category Granularity Adjustment The Classification of Strategic Emerging Industries comprises three hierarchical levels: 7 categories at the first level, 30 at the second level, and 100 at the third level. By effectively utilizing annotation information from each category, the hierarchical level of category mapping can be conveniently adjusted.

3 Cosine Similarity-Based Mapping Model

Based on the guiding principles of accuracy, efficiency, and scalability established for the mapping model, this section designs a mapping workflow for patent documents and industrial categories. Following this workflow, we initially implement mapping between 8,309 invention patents from CAS institutes in 2015 and the 100 categories at the third level of the Classification of Strategic Emerging Industries. Section 4 comprehensively evaluates the model based on the mapping results. The mapping workflow is illustrated in Figure 1 [Figure 1: see original paper].

3.1 Data Acquisition

Considering the volume of patent data and the operability of mapping result evaluation, the experiment selected 8,309 invention patents from various CAS institutes in 2015 for testing, with industrial categories mapped at the third level (100 categories). The experiment requires two major types of data:

3.1.1 Patent Document Data The "CAS Patent Online Analysis System" developed and maintained by the National Science Library, Chinese Academy of Sciences, collects all Chinese patent information published by the State Intellectual Property Office since 1985. The patent data is standardized and comprehensive, meeting the requirements of this experiment. The following search criteria were established: (1) filing country is China; (2) application date is between January 1, 2015 and December 31, 2015; (3) applicant includes CAS; (4) patent type is invention patent and invention authorization. This yielded 8,309 Chinese invention patents with CAS-related units as applicants in 2015. We downloaded the main classification number, applicant, title, and abstract information for all patents.

3.1.2 Official Annotations of Industrial Categories The Classification of Strategic Emerging Industries covers seven national strategic emerging industries: new generation information technology, biotechnology, high-end equipment manufacturing, new energy, new materials, and new energy vehicles. The version used in this experiment is the "2012 edition," which contains 100 categories at the third level. We compiled the complete classification annotation information.

3.2 Keyword Extraction

The experiment uses Python and natural language processing technology to extract keywords from raw data. The process consists of the following three steps:

3.2.1 Word Segmentation We used the open-source Chinese word segmentation component "Jieba Chinese Word Segmentation" [11] for tokenization.

This tool implements efficient word graph scanning based on a prefix dictionary, generating all possible word formations for Chinese characters in a sentence and composing them into a directed acyclic graph. It employs dynamic programming to find the maximum probability path, identifying the optimal segmentation combination based on word frequency. For words not included in the segmentation lexicon, it uses an HMM model based on Chinese character word-formation capability with the Viterbi algorithm, demonstrating strong segmentation performance.

3.2.2 Part-of-Speech Tagging The "Jieba Chinese Word Segmentation" tool adopts the part-of-speech tagging system from the Institute of Computing Technology, Chinese Academy of Sciences, which includes 22 major categories such as nouns, temporal words, locational words, and verbs. In practice, we found that nouns and verbs more accurately reflect the characteristics of patents and industries. Therefore, after segmentation, we apply conditional filtering to extract only nouns and verbs from the segmentation results.

3.2.3 Stop Word Removal Stop words refer to neutral words in annotations without strong categorical characteristics, such as common high-frequency terms like "similar," "for," "provide," and "can." These words lack distinctive features, and words that appear frequently in a corpus are generally considered noise words [12]. To ensure the accuracy of keyword extraction, we remove these noise words. The stop word list used in this experiment was determined based on word frequency statistics from the corpus and the Baidu stop word list [13].

After the above processing, we can batch-extract relatively accurate keywords. To illustrate the effectiveness, we list the keyword extraction results for the first and last categories in the Classification of Strategic Emerging Industries, as shown in Table 1 .

Table 1 Keyword Extraction Results from the Classification of Strategic Emerging Industries (Examples)

Category Code	Category Name	Extracted Keywords
1.1.1	High-Efficiency Energy-Saving General Equipment Manufacturing	boiler, manufacturing, energy-saving, power station, boiler, energy-saving, industrial boiler, energy-saving, marine, steam boiler, economizer, fluidized bed, oil shale, boiler, straw, boiler, coal slime, fluidized bed, boiler, heat storage, blast furnace, gas, boiler, boiler, device, coal powder, industrial boiler, pump, vacuum equipment, manufacturing, energy-saving, energy-saving, vacuum, drying equipment, energy-saving, vacuum furnace, energy-saving, vacuum, gas, compression, energy-saving, refrigeration, compressor, energy-saving, refrigeration, compressor, hydraulic, pneumatic, power machinery, component, manufacturing, energy-saving, hydraulic, component, manufacturing, energy-saving, pneumatic, component, manufacturing, furnace, melting furnace, electric furnace, manufacturing, energy-saving, furnace, burner, energy-saving, mechanical, feeding, device, energy-saving, industrial electric furnace, energy-saving, electric heating, metal, furnace, energy-saving, roller, kiln, energy-saving, tunnel kiln, energy-saving, shuttle, kiln, energy-saving, push plate, kiln, energy-saving, protective, atmosphere, kiln, energy-saving, nitriding, kiln, energy-saving, firing, kiln, energy-saving, baking, steel billet, stepping, heat storage, heating furnace,

Category Code	Category Name	Extracted Keywords
7.3.3	New Energy Vehicles	new energy, vehicle, electric, integration, technology, new energy, vehicle, complete vehicle, technology, new energy, vehicle, complete vehicle, matching, technology, new energy, vehicle, complete vehicle, lightweight, technology, new energy, vehicle, complete vehicle, production process, technology, new energy, vehicle, complete vehicle, technology, new energy, vehicle, vehicle, complete vehicle, product, quality inspection, evaluation, new energy, vehicle, battery, management system, integration, technology, new energy, vehicle, large capacity, power battery, new energy, vehicle, battery, management system, new energy, vehicle, motor, technology, new energy, vehicle, motor, manufacturing, technology, new energy, vehicle, power, motor, development technology, new energy, vehicle motor, controller, development, technology, new energy, vehicle, sensor, development technology, new energy, vehicle, power, device, technology, new energy, vehicle, welding, process, development technology, electric vehicle, sensor, electronic component, technology, power, technology

Category Code	Category Name	Extracted Keywords

3.3 TF-IDF Processing

To calculate similarity between categories more accurately, the extracted keywords undergo TF-IDF processing. TF-IDF is a commonly used weighting technique where TF (term frequency) represents word frequency and IDF (inverse document frequency) represents inverse document frequency, with TF-IDF being their product [14]. This method evaluates the importance of a word to a specific document within a corpus. The importance of a word increases proportionally with its frequency in the document but decreases inversely with its frequency across the corpus [15]. High word frequency in a particular document combined with low frequency of that word in the entire corpus produces high TF-IDF weights. Consequently, TF-IDF tends to filter common words while retaining words with strong category discrimination capability.

3.4 Cosine Similarity Calculation

In the previous step, each keyword has been assigned a TF-IDF value. This section uses cosine similarity to calculate the similarity between patent documents and industrial categories. The process involves:

3.4.1 Vectorization Each patent document's keywords and each industrial category's query terms can be represented as vectors:

$$Patent_i = (w_{1,i}, w_{2,i}, w_{3,i}, \dots, w_{t,i}) \quad (1)$$

$$Industry_q = (w_{1,q}, w_{2,q}, w_{3,q}, \dots, w_{t,q}) \quad (2)$$

Each dimension represents the TF-IDF value w of a word appearing in a patent document or industrial category. The dimensionality of the vector equals the number of distinct words appearing in a patent document or industrial category.

3.4.2 Cosine Similarity Calculation Each patent document or industrial category is represented by a corresponding high-dimensional vector, with each word assigned to a different dimension and each dimension's value being its TF-IDF value. In other words, each patent document or industrial category has been transformed into a vector containing word importance information. Cosine similarity measures the similarity between two vectors by calculating the cosine of the angle between them. A cosine value closer to 1 indicates an angle closer to 0 degrees, meaning the two vectors are more similar. Therefore, cosine similarity can provide similarity measures between patent documents and industrial categories, a method commonly used for similarity comparison in text mining [16]. The formula for calculating cosine similarity between patent documents and industrial categories is as follows:

3.5 Category Similarity Matrix

After the above steps, we ultimately obtain a 100 × 8,309 similarity matrix between patent documents and the Classification of Strategic Emerging Industries, with rows representing the 100 third-level categories of the Classification and columns representing the 8,309 invention patents. Each patent document corresponds to a similarity value with each industrial category. Positive values indicate positive correlation, with larger positive values indicating stronger correlation. This matrix can be used to determine the mapping between patent documents and industrial categories. Additionally, the matrix can be transposed to examine the patent documents corresponding to each industrial category, thereby enabling statistical analysis of patent portfolios of various CAS research institutes under each industrial category.

4 Evaluation of Mapping Results

4.1 Evaluation of Patent-to-Industry Category Mapping

4.1.1 Macro-Level Analysis In the final similarity matrix, each patent document corresponds to similarity values with industrial categories. The experiment defines similarity values greater than 0 as positive correlation. Therefore, for each patent document, we recommend all industrial categories with positive similarity values, sorted in ascending order by similarity value. To observe the overall distribution of recommended categories, we conducted statistical analysis. Table 2 shows the average number of recommended industrial categories for the 8,309 patent documents.

Table 2 Statistics on Recommended Industrial Categories for 8,309 Patent Documents

Metric	Value
Total Categories	100
Average Recommendations	[value not provided in text]
Median	[value not provided in text]
Lower Quartile	24
Upper Quartile	[value not provided in text]
Minimum	[value not provided in text]
Maximum	[value not provided in text]

The frequency distribution histogram of patent document to industrial category mapping is shown in Figure 2 [Figure 2: see original paper]. The histogram exhibits a right-skewed distribution, with recommended mapping quantities concentrated between “10 to 60,” which satisfies certain mapping quantity requirements. Further analysis reveals that the mapping results ensure 99% of patent documents receive recommendations for 5 or more industrial categories, and

96% receive recommendations for 10 or more categories, providing sufficient options for expert judgment and selection. Table 3 shows the recommendation satisfaction rates.

Table 3 Recommendation Satisfaction Rates for 8,309 Patent Documents

Recommendation Threshold	Satisfaction Rate
At least 1 category	99.9%
At least 3 categories	99.6%
At least 5 categories	99.0%
At least 10 categories	95.8%

4.1.2 Micro-Level Analysis Section 4.1.1 evaluated the overall mapping situation. This section delves deeper into the recommended mapping results for specific patent documents. To ensure objectivity and reasonableness, we selected the first patents that correspond to the lower quartile, median, and upper quartile of recommended mapping quantities. For example, the lower quartile of recommended mappings between patent documents and the Classification of Strategic Emerging Industries is 24, with 143 patents meeting this number of mappings. We selected the first patent among the 8,309 patents that satisfies this number, analyzing a total of 3 patents. For each patent, we list the top 5 industrial categories by similarity value. Detailed results are shown in Table 4 .

Table 4 Micro-Level Analysis of Patent Document to Industrial Category Mapping Results

Patent (Quartile)	Top 5 Categories by Similarity
Lower Quartile: Forward Osmosis Separation Method, Application and Recovery Method of Sodium Polystyrene Sulfonate	6.1.2 New Membrane Materials Manufacturing, 1.3.2 Industrial Solid Waste, Waste Gas, and Waste Liquid Recovery and Resource Utilization, 1.4.1 Energy Conservation and Environmental Protection Scientific Research, 6.3.1 High-Performance Fiber Composite Materials Manufacturing, 6.4.3 Smart Materials Manufacturing
Median: A Preparation Method for Vanadium Dioxide/Zinc Oxide Nanocomposite Powder	2.1.1 New Generation Mobile Communication Network Services, [additional categories not fully provided]
Upper Quartile: Wireless-Controlled Ultrathin Section Automatic Collection System	6.3.2 Other High-Performance Composite Materials Manufacturing, 6.2.2 High-Quality Metal Materials Processing and Manufacturing, [additional categories not provided]

The results demonstrate that the model can automatically recommend a certain number of industrial categories for each patent, with most categories aligning with professional judgment. For example, for the patent "Forward Osmosis Separation Method, Application and Recovery Method of Sodium Polystyrene Sulfonate," the complete abstract states: "The present invention provides a forward osmosis separation method, application and recovery method of sodium polystyrene sulfonate. The application includes using sodium polystyrene sulfonate as a draw solute in the forward osmosis process. The invention solves problems existing in current draw solutes, including low osmotic pressure, severe reverse osmosis, difficult recovery, toxicity, and poor membrane compatibility." The recommended top 5 categories from the Classification of Strategic Emerging Industries are "6.1.2 New Membrane Materials Manufacturing, 1.3.2 Industrial Solid Waste, Waste Gas, and Waste Liquid Recovery and Resource Utilization, 1.4.1 Energy Conservation and Environmental Protection Scientific Research, 6.3.1 High-Performance Fiber Composite Materials Manufacturing, and 6.4.3 Smart Materials Manufacturing." These recommended categories show certain similarities with the patent, providing options for expert judgment and selection while saving substantial preliminary labor costs.

4.2 Evaluation of Industry Category-to-Patent Mapping

4.2.1 Macro-Level Analysis By transposing the final similarity matrix, we obtain a matrix with rows representing the 8,309 invention patents from CAS institutes in 2015 and columns representing the 100 third-level categories of the Classification of Strategic Emerging Industries. Each industrial category corresponds to similarity values with patent documents. This experiment defines similarity values greater than 0 as positive correlation. Therefore, for each industrial category, we recommend all patent documents with positive similarity values, sorted in ascending order by similarity value. To observe the overall distribution of recommended patent quantities, we conducted statistical analysis. Table 5 shows the average number of 2015 CAS patents recommended for the 100 industrial categories.

Table 5 Statistics on Recommended 2015 CAS Patents for 100 Industrial Categories

Metric	Value
Total Patents	8,309
Average Recommendations	[value not provided in text]
Median	[value not provided in text]
Lower Quartile	[value not provided in text]
Upper Quartile	[value not provided in text]
Minimum	8

The frequency distribution histogram of industry category to patent document

mapping is shown in Figure 3 [Figure 3: see original paper]. The histogram shows no obvious distribution characteristics, as the results are closely related to the selected patent dataset. Different patent datasets will produce different frequency distributions. In this experiment, the minimum recommended number of patents is 8, ensuring that 99% of industrial categories receive recommendations for 10 or more patent documents. This enables statistical analysis of patent portfolios of various CAS research institutes under each industrial category. Table 6 shows the recommendation satisfaction rates.

Table 6 Recommendation Satisfaction Rates for 100 Industrial Categories

Recommendation Threshold	Satisfaction Rate
At least 1 patent	[value not provided]
At least 3 patents	[value not provided]
At least 5 patents	[value not provided]
At least 10 patents	99%

4.2.2 Micro-Level Analysis Section 4.2.1 evaluated the overall situation. This section selects the first two categories from the Classification of Strategic Emerging Industries—"1.1.1 High-Efficiency Energy-Saving General Equipment Manufacturing" and "1.1.2 High-Efficiency Energy-Saving Special Equipment Manufacturing"—to statistically analyze patent portfolio information of various CAS research institutes under these specific industrial categories from the applicant perspective. The statistical data is shown in Table 7 .

Table 7 Patent Layout of CAS Institutes Under Selected Industrial Categories

Category	Top Institutes (by Patent Count)
1.1.1 High-Efficiency Energy-Saving General Equipment Manufacturing	Changchun Institute of Optics, Fine Mechanics and Physics, CAS; Hefei Institutes of Physical Science, CAS; Shanghai Institute of Optics and Fine Mechanics, CAS; Institute of Semiconductors, CAS; Ningbo Institute of Materials Technology and Engineering, CAS; Technical Institute of Physics and Chemistry, CAS; Guangzhou Institute of Energy Conversion, CAS; Institute of Engineering Thermophysics, CAS; Institute of Process Engineering, CAS; Shanghai Institute of Technical Physics, CAS

Category	Top Institutes (by Patent Count)
1.1.2 High-Efficiency Energy-Saving Special Equipment Manufacturing	Changchun Institute of Optics, Fine Mechanics and Physics, CAS; Hefei Institutes of Physical Science, CAS; Ningbo Institute of Materials Technology and Engineering, CAS; Shanghai Institute of Optics and Fine Mechanics, CAS; Institute of Semiconductors, CAS; Institute of Process Engineering, CAS; Technical Institute of Physics and Chemistry, CAS; Shanghai Institute of Ceramics, Chinese Academy of Sciences; Shanghai Institute of Microsystem and Information Technology, CAS; Changchun Institute of Applied Chemistry, CAS

This model enables observation of applicant distribution under specific industrial categories from a text similarity perspective. For example, in 2015, the "Changchun Institute of Optics, Fine Mechanics and Physics, CAS" and "Hefei Institutes of Physical Science, CAS" filed relatively large numbers of patents in the "1.1.1 High-Efficiency Energy-Saving General Equipment Manufacturing" and "1.1.2 High-Efficiency Energy-Saving Special Equipment Manufacturing" domains, leading among CAS research institutes.

5 Conclusion and Future Improvements

5.1 Conclusion

As a strategic information resource integrating technical, legal, and economic information, patent information serves as a crucial link between science and technology and the economic domain. Effectively utilizing patent data and transforming it into patent indicators to evaluate and monitor industrial development status can significantly contribute to promoting rapid and healthy industrial development. Using 2015 invention patents from Chinese Academy of Sciences institutes and the Classification of Strategic Emerging Industries as a case study, this paper proposes a cosine similarity-based mapping model for patent documents and industrial categories, conducts mapping experiments, and automatically generates mapping combinations between patent documents and industrial categories through natural language processing technology. The model enables bidirectional mapping between patents and industries, saves substantial labor costs, facilitates convenient adjustment of mapping granularity at the category level, and is applicable to most patent-industry category mapping scenarios.

5.2 Improvement Directions

Although the model can produce relatively reliable results, its accuracy can be further improved. We propose two directions for improvement: (1) Enhance

semantic understanding of words. The model currently calculates similarity only at the word level without considering word semantics, such as synonyms and antonyms. If semantic issues can be properly addressed by designing algorithms that assign different weights to synonyms and antonyms, the model's accuracy will improve. (2) Incorporate official annotation features of corresponding patent categories. When calculating similarity, the model does not consider category features of the patent documents' corresponding patent categories. The official category annotations in the International Patent Classification contain relatively accurate word information. If an algorithm can be designed to comprehensively consider both patent document title/abstract information and features of the patent's classification category, the model's accuracy can be further enhanced.

5.3 Expanded Applications

After properly addressing current accuracy issues, this mapping model can be expanded to the following application areas due to its good scalability: (1) Category granularity adjustment. This experiment selected the finest granularity at the third level. With further refinement of the method, the hierarchical level of category mapping can be conveniently adjusted to the first and second levels. (2) Mapping to other industrial classifications. The feature words of industrial categories in this experiment are derived from official category annotations. Different industrial classifications such as the National Economic Industry Classification and High-tech Industry Classification also meet the data requirements of this model and can be adapted for expansion. (3) Mapping of other patent datasets. The patent data in this experiment consists of Chinese invention patents filed in 2015 with CAS-related units as applicants. Since patent information is standardized, any patent set can be selected for experimentation, such as invention patents from a specific institute over the past 10 years, 2015 invention patents from a particular university, or invention patents in a specific technical field.

In future work, we will continue to refine this model by attempting to integrate features from International Patent Classification category annotations to improve mapping accuracy. Establishing a comprehensive patent-industry mapping system can enable the connection between patents and industries, thereby facilitating patent statistical analysis from an industrial perspective combined with scientific, technological, and economic data. This will be of significant importance for decision-making regarding innovation activities and economic development policies in relevant industries.

References

[1] Tian Chuang, Zhao Yajuan. Research progress on patent-to-industry mapping[J]. Library and Information Service, 2016, 60(1): 135-141.
[2] VERSPAGEN B, MOERGASTEL T V, SLABBERS M. MERIT concordance table: IPC-ISIC (rev. 2)[R]. Maastricht: UNU-MERIT, 1994: 1-20.
[3] SCHMOCH U, LAVILLE F, PATEL P, et al. Linking technology areas to

industrial sectors: final Report to the European Commission[R]. Karlsruhe: Fraunhofer ISI, 2003: 36-52.
[4] EUROSTAT. Patent Statistics: Concordance IPC V8 – NACE REV.2[EB/OL]. [2015-09-03]. https://circabc.europa.eu/sd/a/d1475596-1568-408a-9191-426629047e31/2014-10-16-Final%20IPC_{{{NACE2}}{{2014}}}.
[5] LYBBERT T J, ZOLAS N J. Getting patents and economic data to speak to each other: an 'algorithmic links with probabilities' approach for joint analyses of patenting and economic activity[J]. Research policy, 2014, 43(3): 530-542.
[6] State Intellectual Property Office. China's patent literature national economic industry classification indexing work achieves phased results[EB/OL]. [2016-06-24]. http://www.sipo.gov.cn/ghfzs/zltjjb/201503/P020150325567300995160.pdf.
[7] Beijing Intellectual Property Information Center. Research on data organization methods serving industrial sectors[EB/OL]. [2016-06-24]. http://www.sipo.gov.cn/wxfw/zlxxyjjl/hgzdyj/201505/P020150827328075817725.pdf.
[8] National Bureau of Statistics. Classification of Strategic Emerging Industries (2012)[EB/OL]. [2016-06-24]. http://www.stats.gov.cn/zjtj/tjbz/201301/U020131021375903103360.pdf.
[9] State Intellectual Property Office. 2014 statistical analysis report on invention patents in strategic emerging industries[EB/OL]. [2016-06-24]. http://www.sipo.gov.cn/tjxx/yjcg/201504/P020150422347216350682.pdf.
[10] State Intellectual Property Office. Compilation notes for the International Patent Classification and National Economic Industry Classification reference table (trial version)[EB/OL]. [2016-06-24]. http://www.sipo.gov.cn/tjxx/zltjjb/201512/P020151221492994057
[11] Jieba Chinese Word Segmentation[EB/OL]. [2016-06-24]. https://github.com/fxsjy/jieba.
[12] Gu Yijun, Fan Xiaozhong, Wang Jianhua, et al. Automatic selection of Chinese stop word list[J]. Journal of Beijing Institute of Technology, 2005, 25(4): 337-340.
[13] Patent{{{To}}{{Industry}}}: stopwords[EB/OL]. [2016-06-24]. https://github.com/littlewilliam/Patent{{{To}}_{{Industry}}}/blob/master/StopWords.txt.
[14] SALTON G, WONG A, YANG C S. A vector space model for automatic indexing[J]. Communications of the ACM, 1975, 18(11): 613-620.
[15] SALTON G, BUCKLEY C. Term-weighting approaches in automatic text retrieval[J]. Information processing & management, 1988, 24(5): 513-523.
[16] TAN P N, STEINBACH M, KUMAR V. Introduction to data mining[M]. Boston: Pearson Addison Wesley, 2006.

Author Contributions

Tian Chuang: Responsible for literature collection, organization, and paper writing;
Zhao Yajuan: Responsible for overall research direction and framework, and paper review

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

Source: ChinaXiv — Machine translation. Verify with original.