A Multi-Attribute Fusion Method for Mining Technological Innovation Themes from Patents: A Case Study of Chip Patents (Postprint)

Authors: Li Hui, Xuan Hongsheng

Date: 2023-04-01T00:00:00+00:00

Abstract

[Purpose/Significance] This study employs a quantitative method integrating multiple attributes to rapidly and effectively extract multiple technological innovation topics within a domain, providing reference for determining technological innovation directions. [Method/Process] We propose a quantitative method for mining technological innovation topics by combining the LDA (Latent Dirichlet Allocation) topic model with patent value evaluation indicators. First, we comprehensively utilize TF-IDF, perplexity, and the quartile method to construct an LDA topic model for domain patents. Then, using the probability distribution matrix output by LDA, combined with patent value evaluation indicators (claims and IPC), we construct a quantitative indicator system. Next, chip patents are selected for validation experiments, where quantitative indicators are calculated and visualized using heatmaps to identify technological innovation topics. Finally, based on the mapping relationships between patents, LDA output matrices, innovation topics, and quantitative indicators, patent screening and appropriate labeling of technological innovation topics are performed. [Results/Conclusion] The experimental results are evaluated through two approaches: inviting experts in the microelectronics field and referencing the latest domestic and international chip technologies. The results demonstrate that the domain technological innovation topic mining method integrating multiple attributes can rapidly and effectively extract multiple technological innovation topics, and at the practical level, can better provide ideas for relevant domain enterprises and scientific researchers to discover technological innovation topics.

Full Text

Multi-attribute Mining Method for Technology Innovation Themes from a Patent Perspective: The Case of Chip Patents

Li Hui, Xuan Hongsheng
School of Economics and Management, Xidian University, Xi' an 710126

Abstract

[**Purpose/Significance**] This study employs a quantitative method that integrates multiple attributes to rapidly and effectively mine multiple technological innovation themes within a domain, providing reference for determining directions of technological innovation. [**Method/Process**] We propose a quantitative approach for mining technological innovation themes by combining the LDA (Latent Dirichlet Allocation) topic model with patent value evaluation indicators. First, TF-IDF, perplexity, and the quartile method are comprehensively utilized to construct an LDA topic model for domain patents. Then, using the probability distribution matrices output by LDA combined with patent value evaluation indicators (claims and IPC), a quantitative indicator system is constructed. Next, chip patents are selected for verification experiments, where quantitative indicators are calculated and visualized using heatmaps to identify technological innovation themes. Finally, based on the mapping relationships between patents, LDA output matrices, innovation themes, and quantitative indicators, patent screening and reasonable labeling of technological innovation themes are performed. [**Result/Conclusion**] By inviting experts in microelectronics and referencing the latest domestic and international chip technologies to evaluate the experimental results, the findings demonstrate that the multi-attribute domain technological innovation theme mining method can rapidly and effectively extract multiple technological innovation themes, offering better practical guidance for enterprises and researchers in related fields to discover technological innovation themes.

Keywords: patent, perplexity, LDA, quantitative indicator system, technological innovation theme

---

1 Introduction

As China accelerates its national strategy of building an innovative country, technological innovation has been positioned at the core of national development, becoming a strategic support for improving social productivity and comprehensive national strength. In this national strategic context, scientific and technological intelligence has become particularly important as it provides global and strategic decision-making support for technological innovation. Consequently, scholars have focused on selecting appropriate data sources as reliable pathways for obtaining frontier technological innovation intelligence.

Among various scientific and technological literature data, patent data is the most commonly used indicator for measuring technological innovation in academia, making it an effective approach for acquiring technological innovation intelligence. Schmookler used patent data to characterize technological innovation, examining technological innovations in four U.S. industries (railway transportation, agriculture, papermaking, and petroleum processing) from 1836-1957 based on patent application volumes. Griliches proposed that patent data represents an important source of information for technological innovation, possessing unique advantages in completeness and technological innovation information disclosure compared to other indicators. Schmoch found that patent data can be used to analyze technological innovation levels in specific fields, compensating for the limitation that indicators such as R&D budgets and R&D personnel are typically only statistically analyzed at aggregate levels. The OECD noted that patents and patent data can reveal not only the technical fields to which inventions belong but also information about applicants, assignees, and inventors. Zhao Yang et al. pointed out that as patent data grows rapidly, academic methods for mining patent technological innovation information (patent retrieval, patent mapping, patent citation, patent network, and patent text mining) are also rapidly evolving.

Domestic and international scholars have fully recognized the feasibility of using patent data to measure technological innovation and have conducted extensive research on how to efficiently and accurately identify technological themes using patent data. According to relevant studies, technology theme identification mainly falls into two directions: one based on patent citation characteristics, and the other based on patent text content features. Methods based on patent citation characteristics attracted early scholarly attention. Choi et al. constructed patent citation networks and used main path analysis algorithms to identify technology themes. Kwon et al. built patent citation coupling networks and co-citation networks to comprehensively analyze patent distributions and identify technology themes. Zhang Xin et al. combined an improved PageRank algorithm with patent citation counts and patent age, applying it to the OLED field to identify core patents. With the development of natural language processing technologies such as text clustering, LDA topic models, and community detection, content-based technology theme identification methods have gradually gained attention. Hayoung et al. proposed an algorithm for identifying potential technological innovation themes in patents, using augmented reality patent abstracts to identify technology themes with potential innovation value. Yi Huifang et al. combined the LDA model with strategic coordinate mapping for patent text content analysis to identify technology themes and their structural characteristics. Fan Yu et al. proposed a topic model and clustering algorithm for patent content clustering, combining Latent Dirichlet Allocation (LDA) with the OPTICS algorithm for core technology theme analysis.

Comprehensive analysis of existing research reveals that although citation-based identification methods can effectively identify domain technology themes, citation analysis suffers from citation lag, resulting in deficiencies in timeliness and accuracy of identified themes. Furthermore, while text content-based methods have certain advantages over citation-based methods (no citation lag), they also have limitations. For example, mining technology themes from patent titles, abstracts, and other text content only considers natural language processing perspectives without accounting for the economic and technological attributes that technology themes should possess.

In summary, to address current limitations in using patent data for technology theme mining, this paper proposes a multi-attribute quantitative method to rapidly and effectively mine multiple technological innovation themes within a domain, where technological innovation themes refer to generalized technology themes that can be developed or improved. The main innovations are: (1) Avoiding the time lag of patent citation analysis. Using the number of claims and IPC classifications instead of citation counts can randomly select a large number of domain patents as a corpus for mining technology themes, minimizing human influence on final results. (2) Compensating for the lack of economic and technological attributes in technology themes. Research shows that valuable patents are characterized by numerous claims and broad technology coverage, where more IPC classifications indicate broader technical fields involved. Therefore, this paper introduces the number of claims and IPC classifications from patent value evaluation indicators. (3) Constructing a quantitative indicator system that integrates multiple attributes. By comprehensively studying LDA probability distribution matrices and patent value evaluation indicators, quantitative indicators are defined from multiple dimensions to construct a quantitative indicator system for identifying technological innovation themes.

---

2 Research Design

This research design consists of three parts: (1) technical feature vectorization, (2) quantitative indicator system construction, and (3) technological innovation theme mining. The quantitative indicator system construction is both the core of this theoretical research and the main innovation. The specific process of the research design is shown in Figure 1 [Figure 1: see original paper].

2.1 Technical Feature Vectorization

Technical feature vectorization consists of two parts: data preprocessing and vector space model construction.

(1) Data Preprocessing. First, the corpus is segmented, then stop words are removed, stemming is performed, and finally punctuation, special symbols, and numbers are eliminated. In the term frequency matrix, some noise words still appear, such as "method," "system," and "action," which are removed through programming.

(2) Vector Space Model Construction. First, based on the determined number of technical feature words, the preprocessed corpus is converted into a term frequency (TF) matrix. The TF matrix is then transformed into an inverse document frequency (IDF) matrix. Finally, the TF and IDF matrices are multiplied to generate a TF-IDF matrix.

2.2 Quantitative Indicator System Construction

Quantitative indicator system construction is the focus of this research outcome. By constructing a quantitative indicator system, technological innovation themes can be identified—not merely technology themes based on natural language processing attributes, but also incorporating economic and technological attributes. This is divided into two parts: constructing the LDA topic model and building the quantitative indicator system.

2.2.1 Constructing the LDA Topic Model First, the optimal number of topics is determined using a perplexity-based method. Currently, the biggest challenge in applying LDA is determining the optimal number of topics. This paper uses the perplexity method to determine the optimal number of topics for the LDA topic model. The dataset is divided into training and test sets, weighted using TF-IDF, and the LDA model is trained on the weighted training set. Since LDA has advantages in patent text analysis, we generate patent document-topic and topic-feature word probability distribution matrices based on LDA probabilistic topic modeling. After model training, the test set is used as a corpus to calculate the perplexity of the LDA model under different topic numbers, and the topic number with minimum perplexity is selected as the optimal number of topics. The LDA topic model is then formally constructed, ultimately generating document-topic and topic-feature word matrices.

2.2.2 Constructing the Quantitative Indicator System To ensure that technology themes possess not only natural language processing attributes but also economic and technological attributes, we apply quantitative methods to process LDA output probability distribution matrices, claim counts, and IPC classification numbers. Based on relevant theoretical research findings, quantitative indicators are defined to construct a three-level quantitative indicator system for identifying technological innovation themes, as shown in Figure 2 [Figure 2: see original paper].

The construction of the quantitative indicator system is divided into three parts: (1) Level III quantitative indicator definition, (2) Level II quantitative indicator definition, and (3) Level I quantitative indicator definition.

(1) Level III Quantitative Indicator Definition. After constructing the LDA topic model, two probability distribution matrices are generated (document-technology theme and technology theme-feature word). Based on these, we define quantitative indicators with natural language processing attributes (CTM, CTS, TWM) and define quantitative indicators with economic and technological attributes based on IPC classification numbers and claim counts (IPCc, IPCcn, IPCcns).

- **CTM (Corpus Topic Mean)** represents the mean probability of technology themes in the corpus, calculated as shown in Equation (1). CTM indicates the magnitude of a technology theme' s technical value within the current corpus. A larger CTM value indicates greater technical value.

$$CTM(j) = \frac{\sum_{i=1}^{N} t_{ij}}{N}$$

where N represents the number of patent documents in the corpus; M represents the number of themes; and t_{ij} represents the probability value of theme j in patent document i.

- **CTS (Corpus Topic Standard)** represents the standard deviation of technology theme probabilities in the corpus, calculated as shown in Equation (2). CTS indicates the stability of a technology theme' s technical value within the current corpus, measuring the dispersion degree of the theme' s technical value. A smaller CTS value indicates more stable technical value.

$$CTS(j) = \sqrt{\frac{\sum_{i=1}^{N} (t_{ij} - CTM_j)^2}{N}}$$

where N represents the number of patent documents in the corpus; M represents the number of themes; t_{ij} represents the probability value of theme j in patent document i; and CTM_j represents the mean of theme j in the corpus.

- **TWM (Topic Word Mean)** represents the mean probability of feature words for a technology theme, calculated as shown in Equation (3). When calculating TWM, to select feature words with strong explanatory power for the technology theme, the quartile method is introduced to rank feature words under each technology theme in descending order of probability values. The top quartile of feature words is selected to calculate TWM, indirectly optimizing the topic-feature word probability distribution matrix. TWM indicates the degree to which a technology theme is explained. A larger TWM value indicates more sufficient explanation and more convincing technical value.

$$TWM(j) = \frac{\sum_{i=1}^{K} t_{ij}}{K}$$

where K represents the number of feature words; M represents the number of themes; and t_{ij} represents the probability of the i-th feature word of theme j. $TWM(j)$ represents the mean probability of feature words for theme j.

IPC classification numbers and claim counts are evaluation indicators of patent technical value. Compared with citation counts, the former two do not change over time, while the latter grows dynamically over time. IPC classification numbers represent the coverage scope of patent technology. Research shows that larger IPC classification numbers indicate higher technical value and greater economic benefits. Claim counts represent the protection breadth of patent technology, and studies demonstrate a good correlation between claim counts and patent technical value. While citation counts play an irreplaceable role in evaluating patent technical value, they are far less effective than IPC classification numbers and claim counts for evaluating recently published patents, which is particularly suitable for the temporal characteristics of patent data in this study.

Original patent documents contain IPC classification numbers and claims entries. Using statistical methods, we calculate the number of IPC classification numbers and claims in each patent document. Using IPC classification numbers and claim counts can evaluate not only technical value but also potential technical value.

Applying quantitative methods, we multiply them by adjustment coefficients (α, β) respectively, and the sum represents the potential technical value of the patent. The adjustment coefficients are set based on the current corpus data.

- **IPCc (IPC-Claim)** represents the potential technical value of a patent, calculated as shown in Equation (4). IPCc indicates the potential technical value of a patent. A larger IPCc value indicates greater potential technical value.

$$IPCc(i) = \alpha N\_Claimi + \beta N\_IPCi \quad (i \leq N, 0 < \alpha \leq 1, \beta \geq 1)$$

where N_Claimi represents the number of claims in patent document i; N_IPCi represents the number of IPC classifications in patent document i; $\sum Claim_i$ represents the total number of claims in the corpus; and $\sum IPC_j$ represents the total number of IPC classifications in the corpus.

- **IPCcn (IPC-Claim-Normalization)** normalizes IPCc as shown in Equation (7). Since large IPCc values affect subsequent definitions and scientific calculations of Topic Latent Value (TLC), we decided to apply min-max normalization to IPCc after experimentation and discussion. The calculation results meet experimental expectations. IPCcn still represents the potential technical value of a patent.

$$IPCcn(i) = \frac{IPCc_i - IPCc_{min}}{IPCc_{max} - IPCc_{min}} \quad (i \leq N)$$

where $IPCc_i$ represents the i-th value in IPCc; $IPCc_{min}$ represents the minimum value in IPCc; and $IPCc_{max}$ represents the maximum value in IPCc.

- **IPCcns (IPC-Claim-Normalization-Sum)** is the sum of IPCcn values for a technology theme, as shown in Equation (8). As observed in the research design diagram (Figure 1), the document-topic matrix points to the middle of the directed line segment connecting IPCcn and IPCcns. When defining IPCcns, to solve the problem of different potential technical values represented by technology themes, we set an appropriate threshold for the document-topic matrix through experimentation to filter typical technology themes in each document. In the new document-topic matrix, the number 1 indicates that a technology theme appears in the current patent document, while 0 indicates it does not. This establishes a mapping relationship between the new document-topic matrix and IPCcn, matching each technology theme with its corresponding IPCcn, thus solving the problem of representing technology theme potential technical value and completing the definition and calculation of IPCcns. IPCcns represents the sum of patent potential technical values for a technology theme. A larger IPCcns value indicates more patents with potential technical value for that theme.

$$IPCcns(j) = \sum_{i=1}^{P} IPCcn_{ij}$$

where P represents the number of patent documents belonging to each technology theme (varying by theme); M represents the number of technology themes; and $IPCcns(j)$ represents the cumulative sum of IPCcn for theme j.

(2) Level II Quantitative Indicator Definition.

- **TVC (Topic Value Centrality)** is the product of the reciprocal of CTS (Corpus Topic Standard) and TWM (Topic Word Mean), as shown in Equation (9). TVC represents the centrality strength of a technology theme, i.e., the technical value of the theme at the current stage. A larger TVC value indicates greater current technical value.

$$TVC(j) = \frac{1}{CTS_j} \times TWM_j \quad (i \leq M)$$

where M represents the number of themes; CTS_j represents the corpus topic standard deviation value of theme j; and TWM_j represents the mean feature word probability value of theme j.

- **TLV (Topic Latent Value)** is the product of CTM (Corpus Topic Mean) and IPCcns for a technology theme, calculated as shown in Equation (10). TLV represents the magnitude of a technology theme' s potential technical value, i.e., the technical value the theme will have in future stages. A larger TLV value indicates greater future technical value.

$$TLV(j) = CTM_j \times IPCcns_j \quad (j \leq M)$$

where M represents the number of themes; CTM_j represents the mean value of theme j in the corpus; and $IPCcns_j$ represents the sum of IPCcn for theme j.

(3) Level I Quantitative Indicator Definition.

- **TI (Topic Innovation)** is the product of TVC (Topic Value Centrality) and TLV (Topic Latent Value), and is also the quantitative indicator for identifying technological innovation themes by integrating multiple attributes, as shown in Equation (11). TI represents the strength of a technology theme' s innovativeness, i.e., the innovation value of the theme. A larger TI value indicates greater innovation value.

$$TI(j) = TVC_j \times TLV_j \quad (j \leq M)$$

where M represents the number of themes; TVC_j represents the topic value centrality value of theme j; and TLV_j represents the topic latent value of theme j.

2.3 Technology Innovation Theme Mining

2.3.1 Technology Innovation Theme Identification After constructing the quantitative indicator system, technological innovation themes can be identified through the Topic Innovation (TI) metric. However, numerical values alone present poor visualization results. Using knowledge graphs to present theme innovation results enables intuitive identification of technological innovation themes.

2.3.2 Technology Innovation Theme Labeling The technology innovation theme labeling stage is the convergence phase of this research, with the previous four stages preparing for this phase. Although technological innovation themes have been identified, each theme lacks appropriate labeling. The task at this stage is to label innovation themes using data results from the previous four stages. Different thresholds are set for patents belonging to each innovation theme based on IPCc counts to select an appropriate number of patent documents. After discussion, we decided to use IPC classification descriptions and feature words to define technological innovation themes. However, during the experimental phase, we found that some feature words of innovation themes were not sufficiently professional. After repeated discussion and experimentation, we determined to segment and deduplicate titles of innovative patent documents, then select appropriate words from them to replace less professional feature words in innovation themes. Finally, labeling of technological innovation themes is completed based on IPC classification descriptions and optimized feature words.

---

3 Experimental Validation

3.1 Data Acquisition and Preprocessing

The experimental validation corpus consists of patent literature in the chip field. English patent documents in the chip field from 2014-2018 were downloaded from the TotalPatent database, totaling 9,197 documents. The search expression was: Ti:(integrated circuit OR microcircuit OR microchip OR chip). Downloaded patent document entries included title, abstract, IPC classification number, and claims.

The corpus for constructing the LDA topic model uses abstracts from chip patent document entries. Python' s Natural Language Toolkit (NLTK) was used to complete preprocessing of the abstracts.

3.2 Topic Model Construction

3.2.1 Determining the Optimal Number of Topics Before constructing the LDA topic model, in addition to preparing preprocessed corpora, the optimal number of topics must be determined. This paper uses the sklearn toolkit to calculate perplexity values for topic numbers between 0 and 100 to determine the optimal number of topics, as shown in Figure 3 [Figure 3: see original paper].

As shown in Figure 3, perplexity is minimized when the number of topics is around 31, so 31 is determined as the optimal number of topics for the LDA topic model.

3.2.2 Constructing the LDA Topic Model Vector space model construction uses the sklearn toolkit, setting the number of feature words to 2,000, then generating term frequency (TF) matrix, inverse document frequency (IDF) matrix, and TF-IDF matrix. All three matrices have dimensions of 9,197 $\times$ 2,000. After reading the TF-IDF matrix, the LDA topic number is set to 31, iteration count to 100. LDA initialization information is shown in Table 1 . LDA model construction uses the LDA toolkit to generate document-topic (Doc-Topic) and topic-feature word (Topic-Term) probability distribution matrices, with partial matrices shown in Table 2 and Table 3 .

3.3 Calculating Quantitative Indicators

3.3.1 Calculating Quantitative Indicators Based on LDA Probability Distribution Matrices (1) Calculating CTM (Corpus Topic Mean) and CTS (Corpus Topic Standard). Using the document-topic matrix, the mean (Equation (1)) and standard deviation (Equation (2)) of probabilities for each chip technology theme in the corpus are calculated. Final results are rounded to two decimal places, as shown in Table 4 , Table 5 , and Figure 4 [Figure 4: see original paper].

(2) Calculating TWM (Topic Word Mean). Using the topic-feature word matrix, the mean probability of feature words for each chip technology theme is calculated (Equation (3)). First, the quartile method is used to select effective feature words for each technology theme, with 32 feature words most representative of each theme selected, as shown in Table 6 . Then, the mean probability of feature words under each theme (TWM) is calculated, with final results rounded to two decimal places, as shown in Table 7 .

3.3.2 Calculating Quantitative Indicators Based on Patent Literature Evaluation Indicators

(1) Calculating IPCc (IPC-Claim) and IPCcn (IPC-Claim-Normalization). IPC and claims entry data from the chip patent corpus are read, then IPC classification numbers and claims are segmented into single IPCs and claims according to their organizational characteristics. Using statistical knowledge, the quantities after segmentation are counted to replace original IPC and claims entries, with updated entries shown in Table 8 . Finally, IPCc for each patent is calculated (Equation (4)), and a new IPCc entry is created in the chip corpus, as shown in the new entry IPCc in Table 8.

(2) Calculating IPCcn (IPC-Claim-Normalization). Using min-max normalization, IPCc entry data is normalized, and a new IPCcn entry is created in the chip corpus using the resulting data, with calculation results (Equation (7)) shown as the new entry IPCcn in Table 8.

3.4 Identifying Technology Innovation Themes

This section' s experiments consist of two parts: calculating Topic Value Centrality (TVC), Topic Latent Value (TLV), and Topic Innovation (TI), and visualizing the identification of technological innovation themes.

(1) Referring to Equations (9), (10), and (11), TVC, TLV, and TI are calculated respectively. Results are rounded to two decimal places, as shown in Table 9 .

(2) Visualizing Technology Innovation Theme Identification. The central idea of this study is to propose a multi-attribute quantitative method to quickly and effectively mine multiple technological innovation themes. The specific approach is to first use LDA to mine technology themes, then integrate multiple attributes to mine technological innovation themes from these themes, and finally invite domain experts to professionally evaluate the mined themes. If expert scores indicate that the technology themes have value, this research is meaningful; conversely, if expert scores indicate no value, the proposed method is erroneous. Selecting 5 technology themes based on TI values aligns with this study' s purpose—rapidly and effectively mining multiple potentially most valuable technological innovation themes.

During visualization, heatmaps are used to present TVC, TLV, and TI calculation results. During the experiment, we found that TI values are relatively large, making differences in TVC and TLV among technology themes less distinct in the heatmap. After several experimental validations, scaling TI values down by 5 times yields the best heatmap effect, with visualization results shown in Figure 5 [Figure 5: see original paper].

Through intuitive heatmaps and TI value ranking tables (see Table 10), the top 5 technology themes by TI value are selected as technological innovation themes: Topic-5, Topic-27, Topic-25, Topic-28, and Topic-8.

3.5 Labeling Technology Innovation Themes

(1) Data Preparation. Read the chip corpus with the new IPCc entry, then read the document-topic matrix.

(2) Data Processing. First, extract technological innovation theme patent documents. Construct a new document-innovative theme matrix from the portions corresponding to the 5 technological innovation themes in the LDA-generated document-topic matrix, establishing a mapping relationship between the corpus and the new matrix to extract qualified patent documents. Due to the large number of qualified patent documents, different thresholds are set to select patent documents. Taking innovative theme Topic-5 as an example, the numbers of extracted patent documents are 5,284, 3,256, and 122 when probability values in the new matrix are rounded to two decimal places, one decimal place, and not rounded, respectively. The other four themes show the same decreasing pattern. After discussion, we decided to use the matrix without decimal places to generate Boolean indexes for mapping with the corpus to extract a reasonable number of patent documents. The numbers of documents corresponding to the five innovation themes (Topic-5, Topic-8, Topic-25, Topic-27, Topic-28) are 122, 87, 121, 134, and 165, respectively.

Second, segment and deduplicate IPC and titles. After identifying patent documents corresponding to each technological innovation theme, extract their IPC and title entry contents, segment them according to their content characteristics, and perform deduplication. Count the number of IPCs after deduplication for each theme, as shown in Table 11 . The large quantity makes labeling technological innovation themes difficult.

Third, filter patent documents for technological innovation themes. Due to the quantity issue of IPCs, effective labeling of technological innovation themes is hindered. The solution, determined through experimentation and discussion, is to first sort patent documents of innovation themes by IPCc quantity, group and count patent documents and IPCs by IPCc quantity, then for Topic-5, Topic-8, Topic-25, Topic-27, and Topic-28, select groups with IPCc quantities greater than 32, 29, 29, 31, and 30, respectively. Finally, the numbers of patent documents and IPCs after filtering for the five innovation themes are counted, as shown in Table 12 .

(3) Labeling Results. Labeling of technological innovation themes is based on optimized patent IPC descriptions and innovative feature words. Innovative feature words are determined based on segmentation results of patent document titles and feature words from the quartile-optimized topic-feature word matrix, with 30 innovative feature words labeling each innovation theme. Topic-5, Topic-27, Topic-25, Topic-28, and Topic-8 are renamed as C-Topic-1, C-Topic-2, C-Topic-3, C-Topic-4, and C-Topic-5 according to their TI values. Labeling of technological innovation themes is completed based on IPC descriptions and feature words. The labeling results are shown in Table 13 .

3.6 Results Validation

This study selects patent data from the chip field for 2014-2018 and evaluates the effectiveness of mined technological innovation themes using two scientific approaches:

3.6.1 Expert Evaluation in the Chip Field Since technology themes within a domain overlap to some extent, the proposed technological innovation themes also exhibit some overlap. Individual scoring does not conform to validity criteria. Experts integrate the proposed technological innovation themes and evaluate them holistically. Five experts in microelectronics were invited, all possessing professional knowledge and rich experience in chip technology innovation. The scoring rule is: 1 to 10 points represent low to high quality of technological innovation themes, comprehensively considering three aspects: technical value, innovation value, and application value. Each microelectronics expert provided professional evaluations for the five proposed technological innovation themes. Final scores are shown in Figure 6 [Figure 6: see original paper], with statistical results in Table 14 .

3.6.2 Latest Domestic and International Chip Technology Research Chip technology has entered the post-Moore' s Law era, with both Moore' s Law (deep Moore) for chip manufacturing and beyond-Moore for chip packaging having arrived. Deep Moore primarily involves innovative R&D in materials, device structures, and interconnect wires, with more applications in digital circuits. Beyond-Moore mainly involves heterogeneous integration of different modules using packaging technology in the same package, with more applications in analog circuits. The technological innovation themes mined in this study align with the latest chip technology scope, where C-Topic-1, C-Topic-2, and C-Topic-3 belong to chip packaging innovation themes, while C-Topic-4 and C-Topic-5 belong to chip manufacturing innovation themes, as shown in Table 15 .

This paper proposes a multi-attribute quantitative method to rapidly and effectively mine multiple technological innovation themes. Table 14 shows that experts affirm the value of the mined technological innovation themes. Table 15 shows that all five technological innovation themes belong to chip technology innovation R&D directions in the post-Moore era. Based on these two evaluation methods, the effectiveness of the proposed multi-attribute rapid mining method for domain technological innovation themes is confirmed. A potential controversy may lie in the number of technological innovation themes selected in Section 3.4(2). Selecting five technology themes as technological innovation themes aligns with this study' s purpose (rapidly and effectively mining multiple technological innovation themes). The optimal number for "multiple technological innovation themes" will be determined through further research.

References

[1] Wen Jun, Zhang Sen. Patents, technological innovation and economic growth[J]. East China Economic Management, 2019, 33(8): 152-158.

[2] SCHMOOKLER J. Changes in industry and in the state of knowledge as determinants of industrial invention[C]// NELSON R R. The rate and direction of inventive activity. Princeton: Princeton University Press, 1962: 195-232.

[3] GRILICHES Z. Patent statistics as economic indicators: a survey[J]. Journal of economic literature, 1990, 28(4): 1661-1707.

[4] SCHMOCH U. Indicators and the relations between science and technology[J]. Scientometrics, 1997, 38(1): 103-116.

[5] OECD. Patent statistics manual[M]. Paris: OECD Publishing, 2009.

[6] Zhao Yang, Wen Tingxiao. Research progress on patent technology information mining[J]. Library, 2018(4): 28-33.

[7] CHOI C, PARK Y. Monitoring the organic structure of technology based on the patent development paths[J]. Technological forecasting & social change, 2009, 76(6): 754-768.

[8] KWON O, SEO J, NOH K, et al. Categorizing influential patents using bibliometric analysis of patent citations network[J]. Information, 2017, 10(3): 313-326.

[9] Zhang Xin, Ma Ruimin. Research on core patent discovery based on improved PageRank algorithm[J]. Library and Information Service, 2018, 62(10): 106-115.

[10] WANG Y, BAI H J, STANTON M, et al. PLDA: parallel latent Dirichlet allocation for large-scale applications[C]// International conference on algorithmic aspects in information and management. San Francisco: Springer-verlag, 2009: 301-314.

[11] NEWMAN M E J, GIRVAN M. Finding and evaluating community structure in networks[J]. Physical review, 2004, 69(2): 108-155.

[12] BLONDEL V D, GUILLAUME J L, LAMBIOTTE R, et al. Fast unfolding of communities in large networks[J]. Journal of statistical mechanics: theory and experiment, 2008, 30(2): 155-168.

[13] HAYOUNG C, SEUNG H Y O, SUNGCHUL C, et al. Innovation pattern of technology: the case of augmented reality patents[J]. IEEE access, 2018(6): 16119-16137.

[14] Yi Huifang, Wu Hong, Ma Yongxin, et al. Patent technology theme analysis based on LDA and strategic coordinates: a case study of graphene[J]. Journal of Intelligence, 2018, 37(5): 97-102.

[15] Fan Yu, Fu Hongguang, Wen Yi. Patent information clustering technology based on LDA model[J]. Computer Applications, 2013, 33(S1): 87-89, 93.

[16] Lü Xiaorong. Patent value evaluation indicator system and empirical research on patent technology quality evaluation[J]. Science & Technology Progress and Policy, 2014, 31(20): 113-115.

[17] LANJOUW J, SHANKERMAN M. Stylized facts of patent litigation: value, scope and ownership[R/OL]. [2019-10-19]. https://www.nber.org/papers/w6297.pdf.

[18] Sun Wei, Liu Wenjing, Ge Lige, et al. A patent literature classification method based on word-weighted LDA model[J]. Computer Technology and Development, 2019(3): 23-29.

[19] Pang Jianfeng, Bu Dongbo, Bai Shuo. Research and implementation of text automatic classification system based on vector space model[J]. Computer Applications Research, 2001(9): 23-26.

[20] BLEI D M, NG A Y, JORDAN M I. Latent Dirichlet allocation[J]. Journal of machine learning research, 2003(3): 993-1022.

[21] Zhang Han, Xu Shuo, Qiao Xiaodong. A review of topic model development integrating internal and external features of scientific literature[J]. Journal of the China Society for Scientific and Technical Information, 2014, 33(10): 1108-1120.

[22] Liao Liefa, Le Fugang, Zhu Yalan. Application of LDA model in patent text classification[J]. Modern Information, 2017, 37(3): 35-39.

[23] Yang Chao, Zhu Donghua, Wang Xuefeng. Patent technology theme analysis: LDA topic model method based on SAO structure[J]. Library and Information Service, 2017, 61(3): 86-94.

[24] WON S L, EUN J H, SO Y S. Predicting the pattern of technology convergence using big-data on large-scale triadic patents[J]. Technological forecasting & social change, 2015, 100: 317-329.

[25] Zhang Wenjun, Gu Xingfa, Chen Liangfu, et al. K-means initial clustering center selection algorithm based on mean-standard deviation[J]. Journal of Remote Sensing, 2006, 10(5): 715-721.

[26] PAOLA D R, SABRINA S, VINCENZO L. A semantic-grained perspective of latent knowledge modeling[J]. Information fusion, 2017, 36: 52-67.

[27] Li Qinghai, Liu Yang, Wu Sizong, et al. Overview of patent value evaluation indicators and hierarchical analysis[J]. Science Research Management, 2007, 25(2): 281-286.

[28] Yan Mingyi. Statistical analysis of functional data: ideas, methods and applications[J]. Statistical Research, 2007(2): 87-94.

[29] Wen Ying, Zhou Xin, Zhao Wenming. Quantitative evaluation of professional quality for software majors in vocational colleges[J]. Computer Engineering and Design, 2017, 38(9): 2586-2590.

[30] BIRD S, KLEIN E, LOPER E. Natural language processing with python[M]. New York: O' Reilly Media Press, 2009: 41-134.

[31] PEDREGOSA F, VAROQUAUX G. Scikit-learn: machine learning in python[J]. Journal of machine learning research, 2011, 12: 2825-2830.

[32] LDA Developers. LDA: topic modeling with latent Dirichlet allocation[EB/OL]. [2019-11-22]. https://lda.readthedocs.io/en/latest/.

[33] National Intellectual Property Administration - International Patent Classification Table (2008.01 edition)[S/OL]. [2019-09-01]. http://www.sipo.gov.cn/wxfw/zlwxggfw/zsyd/bzyfl/gjzlfl

[34] FRANK H, JOCHEN G, MICHAEL H. Memetic search for overlapping topics based on a local evaluation of link communities[J]. Scientometrics, 2016, 11(2): 1089-1118.

[35] Zhang Baishang, Shang Huimin. Analysis of current status and trends of chip industry technology at home and abroad[J]. Science and Technology Management Research, 2019(17): 131-134.

[36] Wang Lina, Tang Chuan, Fang Junmin, et al. Analysis of global semiconductor planning and development trends in 2018[J]. World Science and Technology Research and Development, 2019, 41(2): 120-126.

Author Contributions

Li Hui: Proposed research ideas, revised the paper.

Xuan Hongsheng: Collected data, designed and implemented the method, analyzed experimental results, drafted the paper.

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