Asia Pacific Science Press

Asia Pacific Economic and Management Review
ISSN(O):3005-9275
ISSN(P):3005-9267
Vol. 3 No.5 (2026)

Research on Identification of Key Technological Topics of Traditional Chinese Medicine Enterprises Based on LDA Model

Menglei Li*, Lei Yang, Min Lu

School of Economics and Management, Shaanxi Institute of International Trade & Commerce, Xi'an, Shaanxi, 712046, China

Corresponding author: Menglei Li, floralimenglei@126.com

Copyright: 2026 Author(s). This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY-NC 4.0), permitting distribution and reproduction in any medium, provided the original author and source are credited, and explicitly prohibiting its use for commercial purposes.

Abstract: Patent abstracts do more than record that an invention exists. They also leave a trace of the technical problems enterprises return to and the approaches they choose to protect. This study uses an LDA topic model to examine invention patent abstracts from Traditional Chinese Medicine (TCM) enterprises. After stop-word removal and jieba segmentation, key technical terms were extracted for modeling. The analysis produced four recurring topics, including Quality Detection, Intelligent Systems, Drug Development, and Production Equipment. Patent activity is therefore spread across product R&D, testing, digital production, and manufacturing hardware instead of being concentrated in one area. Reading the recurrent terms through these topics helps show where technical effort is currently directed and gives enterprises a practical basis for reviewing R&D priorities and technology portfolios.

Keywords: Traditional Chinese Medicine Enterprises; LDA Topic Model; Key Technologies; Topic Identification

Published: Nov 4, 2026

DOI: https://doi.org/10.62177/apemr.v3i5.1854

1. Introduction

Traditional Chinese medicine (TCM) has long been used in clinical practice, with holistic diagnosis and syndrome differentiation forming part of its distinctive approach. Herbal medicines, acupuncture, massage, and related therapies are used in prevention, treatment, and rehabilitation. Its present development, however, is no longer confined to China. TCM products and services have attracted wider attention in a number of countries and regions [1]. South Korea and Japan are established East Asian markets, while access conditions in some Western countries have gradually become less restrictive; selected TCM services have also entered local health-insurance arrangements in several places. For enterprises, this wider setting creates opportunity but also raises expectations for evidence, quality, standardization, and reliable production. TCM firms thus work between two demands: preserving established medical knowledge and turning that knowledge into products and processes that meet contemporary technical and market requirements [2]. That task involves specific choices about formulations, quality control, manufacturing, and digital technologies. Patents make many of those choices visible because each patent records a concrete technical solution. Viewed as a collection, patents can show not only what firms seek to protect but also the areas in which they repeatedly commit technical effort [3]. This makes patent documents a useful source for examining enterprise-level technological development. Existing TCM patents cover medicinal-material cultivation and processing, formulation research, production methods, and diagnostic or treatment-related technologies [4], providing a

relatively direct record of how traditional knowledge is being translated into modern technical practice.

The difficulty is that this record is technically diverse. A single patent abstract may combine ingredients, preparation procedures, testing indicators, equipment, and therapeutic use, so conventional classification can hide important distinctions. IPC classes are useful for broad grouping, yet they are often too coarse to separate narrower themes within the same category. Manual keyword coding reaches more detail but introduces another problem: results depend heavily on researcher judgment and are hard to apply consistently across thousands of documents. Text mining offers a different way to organize the material. Latent Dirichlet Allocation (LDA) is an unsupervised probabilistic topic model that identifies latent themes from patterns of word co-occurrence and has been widely used in document analysis [5]. Applied to TCM patent abstracts, it can reveal recurring combinations of technical terms and show where innovation activity clusters. The present study uses LDA for that purpose, identifying the main technological topics in the patent corpus and examining how they fit into the wider pattern of enterprise innovation [6].

2. Research Design

2.1 Research Framework

Three steps were used in the analysis. ① Patent titles and abstracts were collected and prepared for text analysis. Common stop words were removed, after which jieba was used to segment the Chinese text. A separate filter removed routine patent phrases with little thematic value, such as 'the invention,' 'relates to,' and 'particularly.' ② Word-frequency analysis was then applied to the cleaned corpus to identify prominent terms and compare their distribution across patents. ③ Finally, the segmented texts were entered into the LDA model, which was used to identify recurring technological themes and the keyword distribution associated with each topic.

2.2 Data Sources

The sample was defined with reference to the industry classification of the China Securities Regulatory Commission and relevant industry standards. It includes 72 enterprises whose main business is traditional Chinese medicine. The analysis uses invention-patent abstracts dated from January 1, 2016, to December 31, 2025. Records with missing, abnormal, or unverifiable information were removed when they could not be analyzed reliably. The screening left 9,756 valid patents retrieved from the Innojoy patent database.

3. Research Results and Topic Analysis

3.1 Core Technical Keyword Analysis

Before interpreting the topic model, Table 1 gives a first view of where technical vocabulary is concentrated. 'Preparation' has the largest TF-IDF weight among the top 20 terms (0.129014), followed by 'Method' (0.108819). The first result is consistent with the structure of TCM formulation, in which raw-material handling, extraction, processing, and final preparation are closely linked. 'Method' is less specific, but its prominence shows that many patents concern how research, production, or quality-control work is carried out rather than only the product itself. Other terms sharpen the picture. 'Detection' and 'Quality' are associated with testing and quality assurance; 'Combination,' 'Traditional Chinese medicine,' 'Drug,' and 'Extract' point more directly to formulation and drug development. 'Device,' 'Technology,' and 'Extraction' add process and equipment concerns. The keyword profile therefore mixes pharmaceutical development, quality control, and production-related innovation, with no single technical direction dominating the corpus.

Table 1: Top high-frequency keywords

No.	Keyword	Keyword Frequency	No.	Keyword	Keyword Frequency
1	Preparation	0.129014	11	Chromatogram	0.03282
2	Method	0.108819	12	Content	0.032721
3	Detection	0.068958	13	Chromatography	0.030019
4	Combination	0.058721	14	Technology	0.029904
5	Traditional Chinese medicine	0.057414	15	Ingredient	0.028507

No.	Keyword	Keyword Frequency	No.	Keyword	Keyword Frequency
6	Drug	0.046698	16	Device	0.028446
7	Solution	0.045913	17	Extraction	0.028257
8	Procedure	0.045323	18	Treatment	0.028064
9	Extract	0.038838	19	Formulation	0.027621
10	Quality	0.034232	20	Granules	0.024832

3.2 Core Technical Topic Identification

Table 2 shows the four topics produced by the LDA model. Each was interpreted from its ten highest-frequency keywords, read together because the meaning of a topic lies in the combination of terms rather than in any single word. On this basis, the four topics were labeled ‘Quality Detection,’ ‘Intelligent Systems,’ ‘Drug Development,’ and ‘Production Equipment.’

Table 2: Topic-keyword distribution

Topic	Topic Designation	Keywords
Topic 1	Quality Detection	Method, Detection, Solution, Quality, Content, Chromatogram, Chromatography, Control, Granules, Ingredient
Topic 2	Intelligent Systems	Data, Method, Information, Detection, Module, Model, System, Production, Image, Collection
Topic 3	Drug Development	Preparation, Combination, Drug, Traditional Chinese medicine, Method, Weight, Extract, Effect, Technology, Raw Material
Topic 4	Production Equipment	Device, Arrangement, Mechanism, Connection, Component, Installation, Fixation, Technology, Equipment, Control

The first cluster is built around ‘Method,’ ‘Detection,’ ‘Solution,’ ‘Quality,’ ‘Content,’ ‘Chromatogram,’ and ‘Chromatography,’ together with related control terms. The common thread is measurement: the patents deal with how TCM products are examined, compared, and described through testing results. The presence of chromatographic and content-related vocabulary points in particular to attempts to make quality assessment more consistent across products and production settings.

Information-processing terms dominate the second cluster. ‘Data,’ ‘Information,’ ‘Module,’ ‘Model,’ ‘System,’ ‘Image,’ and ‘Collection’ appear alongside ‘Method’ and ‘Detection.’ Rather than describing one manufacturing technique, this group is concerned with how information is collected, organized, and used during production and testing. The combination suggests a gradual embedding of digital tools and model-based processing in TCM manufacturing and management.

The third cluster turns toward the product itself. Its leading terms, ‘Preparation,’ ‘Combination,’ ‘Drug,’ ‘Traditional Chinese medicine,’ ‘Weight,’ ‘Extract,’ ‘Effect,’ and ‘Raw Material’, describe formulations, ingredient combinations, extracts, preparation methods, and material use. In this part of the patent portfolio, traditional prescriptions and raw materials are being expressed as technical solutions that can be specified, tested, and protected within modern patent and pharmaceutical-development procedures.

The fourth cluster is the most clearly engineering-oriented. Words such as ‘Device,’ ‘Arrangement,’ ‘Mechanism,’ ‘Connection,’ ‘Component,’ ‘Installation,’ ‘Fixation,’ and ‘Equipment’ refer to the physical structure and operation of production systems. These patents focus on how equipment is assembled, connected, controlled, or adapted for TCM manufacturing. Their contribution is practical, and they provide the hardware and process support for more standardized production and a gradual move toward automation.

The four clusters therefore occupy different points in the same technological chain. Drug Development is product-centered, Quality Detection deals with evaluation and control, Intelligent Systems concerns data and coordination, and Production Equipment provides the physical means of manufacturing. Their simultaneous presence shows that innovation in the sampled enterprises is distributed across linked stages of production rather than concentrated in one field.

4. Conclusions and Recommendations

Across the patent corpus, the four topics , 'Quality Detection,' 'Intelligent Systems,' 'Drug Development,' and 'Production Equipment,' form a connected picture of innovation. The recommendations below follow that pattern and focus on the practical issues implied by the patent results.

4.1 Deepen Innovation in Quality Detection and Improve the Standardized Quality Management System

Quality Detection stands out because testing-related terms recur throughout the patent results. A practical next step is to examine how well the methods already being developed perform outside a narrow laboratory setting. Chromatography and spectroscopy remain important, but reliability across medicinal materials, decoction pieces, and proprietary medicines matters as much as technical sophistication. Biosensors, fingerprint analysis, and similar tools are useful when they solve a defined measurement problem rather than simply adding novelty. Standardization is another concern. Patented methods with stable performance can provide evidence when national, industry, or local criteria need revision, especially where existing evaluation rules are inconsistent. Translation to routine production is a separate hurdle: laboratory methods may need adjustment for cost, testing speed, available equipment, and operator requirements. Collaboration among enterprises, universities, and research institutes is most valuable when it addresses these concrete constraints and turns mature methods into quality-control procedures that factories can actually use.

4.2 Enhance Intelligent System Innovation and Build a Data-Driven Manufacturing Model

In the Intelligent Systems cluster, data and production terms appear together rather than as separate strands. This matters because digitalization has value only when it improves a specific production or inspection task. Patent development can be linked to real-time data collection, online quality assessment, process monitoring, and adaptive control. A production-wide platform is useful when testing, manufacturing, and management information can move between functions and inform day-to-day decisions. The same standard should be applied to large models designed for TCM. Their usefulness depends less on novelty than on the quality of the underlying data and whether the output improves an identifiable inspection or process-control problem. In many enterprises, reducing fragmented data systems and improving consistency may therefore be a necessary step before more advanced algorithms can become part of routine production.

4.3 Strengthen Drug Development Innovation and Break Through Bottlenecks in New Drug Creation

Drug Development brings together two tasks that are related but not identical: creating new products and modernizing established prescriptions. In both cases, the clinical purpose needs to be clear enough to guide formulation work and evidence collection. Major chronic diseases, difficult or rare diseases, and emerging infectious diseases may be relevant priorities when they fit the theoretical basis and practical strengths of TCM. Development routes also need to reflect product type. Compound formulas, effective extracts, classic prescriptions, and experienced doctors' formulas do not begin with the same evidence base and should not be treated as if they did. Digital tools can support particular stages, for example, knowledge graphs can organize relationships among prescriptions and ingredients, human-use data can make accumulated clinical experience easier to examine, and decision models may reduce repeated screening early in development. When a product moves toward translation, policy becomes more important. R&D incentives and review procedures should support promising candidates without weakening requirements for safety, quality, and evidence; for classic formulas and hospital preparations with substantial human-use experience, clearer review pathways may remove avoidable delay while keeping those standards in place.

4.4 Advance Production Equipment Upgrading to Support Large-Scale and Intelligent Production

Production Equipment is not a background topic in the patent results; it forms part of the technological core of TCM manufacturing. Material handling, extraction, preparation, testing, and transfer between stages all depend on equipment performance, so weaknesses at one point can affect both quality and throughput elsewhere. Enterprises can work with equipment manufacturers and research institutes on specialized devices, conveyance systems, and control technologies that match actual TCM production conditions. Priority is best given to recurring operational bottlenecks, particularly where manual handling, unstable process conditions, or poorly connected equipment cause problems. Upgrading should also

be considered across the chain rather than machine by machine. Automated lines and shared workshops are useful when they improve the link between raw-material processing and later manufacturing instead of creating isolated smart units. Demonstration factories can still be valuable, but mainly when they generate practices that other firms can adapt and when equipment innovation is tied to production and quality standards.

Funding

Shaanxi Institute of International Trade & Commerce (2025): Research on the Technological Innovation Model of Shaanxi TCM Enterprises Based on Social Network Analysis (SMXY25017); Research on the High-Quality Development of Shaanxi's Health Industry Empowered by New Quality Productive Forces (SMXY25016); and Digital Trade and Internationalization of TCM Science and Technology Innovation Team (2025KCTD04).

Conflict of Interests

The authors declare that there is no conflict of interest regarding the publication of this paper.

Reference

[1] Ma, Y., Chang, Z., Li, Q., Fang, Z., & Dong, W. (2026). Inheritance and innovation of traditional Chinese medicine in the new era of AI: An exploration of the synergy between intelligent agents and legal frameworks. Research in Globalization, 12, 100369.

[2] Guo, D., Qi, L., & Song X. (2025). Examining how fiscal policies influence innovation in TCM enterprises: the role of R&D investment and executives with pharmaceutical backgrounds. Frontiers in Public Health, 2025, 13, 1531622.

[3] Cao, Y., Gong, R., Chen, Ning., & Liu, W. (2019). Research on patent quality of Chinese listed traditional Chinese medicine enterprises from R&D investment perspective. China journal of Chinese materia medica, 44 (6), 1284-1288.

[4] Li, X. (2024). Influencing factors, evolution, and strategies for improving the quality of traditional Chinese medicine patents: an empirical study based on patent invalidation cases. Annals of medicine, 56 (1), 2422571.

[5] Lu, K., Xie, M., Sun, W., & Liu, Y. (2025). Evolution of traditional Chinese medicine registration review and approval policies: research based on the LDA topic model. Frontiers in Public Health, 13, 1655636.

[6] Li, M., Xia, L., Wu, Q., Wang, L., & Zhu, J. (2025). et al. MD-LDA: a supervised LDA topic model for identifying mechanism of disease in TCM. Data Technologies and Applications, 59 (1): 1-18.