Comparative Study on Technology Gaps in Animal Husbandry and Veterinary Genomics between China and Abroad: Postprint

Authors: Wu Lei, Li Xiaojie, Ding Qian, Sun Wei, Zhou Zhengkui

Date: 2024-04-03T00:00:00+00:00

Abstract

[Purpose/Significance] To identify technological gaps in China' s key agricultural domains and predict the future development trends of these gaps, thereby providing effective consulting recommendations on technological development opportunities for science and technology management decision-makers. [Method/Process] First, mine the abstract information from papers and patents using key sentence embedding methods and sentence vector clustering methods; then conduct comparative analysis of topic clustering to discover technological gaps; second, construct semantic similarity networks and classification similarity networks to identify thematic directions that can readily form cross-fusion with these gaps. [Results/Conclusion] An empirical analysis of genomics technology was conducted in the field of animal husbandry and veterinary medicine. The results demonstrate that this method can identify technological gaps, and combined with expert analysis, can interpret the current development status and predict future trends of genomics technology in the field of animal husbandry and veterinary medicine, providing methodological and data support for think tank consulting on genomics technology in China' s animal husbandry and veterinary domain.

Full Text

Preamble

Title: A Comparative Study of Technological Gaps in Genomics for Animal Husbandry and Veterinary Medicine Between China and Foreign Countries

Authors: WU Lei[1,2], LI Xiaojie[3], DING Qian[1,2], SUN Wei[1,2], ZHOU Zhengkui[4]

Affiliations: 1. Agricultural Information Institute, Chinese Academy of Agricultural Sciences, Beijing 100081 2. Key Laboratory of Agricultural Big Data, Ministry of Agriculture and Rural Affairs, Beijing 100081 3. Shenzhen Agricultural Genome Research Institute, Chinese Academy of Agricultural Sciences, Shenzhen 518120 4. Institute of Animal Sciences, Chinese Academy of Agricultural Sciences, Beijing 100193

Abstract:
[Purpose/Significance] To identify technological gaps in China' s key agricultural fields and predict future development trends of these gaps, providing effective consultation for science and technology management decision-makers. [Method/Process] First, key sentence embedding and sentence vector clustering methods were used to mine abstract information from papers and patents. Comparative analysis of topic clustering was then performed to discover technological gaps. Second, semantic similarity networks and classification similarity networks were constructed to identify thematic directions that could easily form cross-fusion with gap points. [Results/Conclusions] An empirical analysis of genomics technology in animal husbandry and veterinary medicine demonstrates that this method can effectively identify technological gaps. Combined with expert analysis, it enables interpretation of current development status and prediction of future trends in genomics technology for animal husbandry and veterinary medicine in China, providing methodological and data support for think tank consultation in this domain.

Keywords: technology gap discovery; key sentence extraction; sentence vector clustering; genomics; intellectual property

---

1 Introduction

Technological innovation refers to changes occurring during the progress and development of science and technology. Science and technology management decision-makers need to understand the processes and dynamics of technological innovation, grasp and even predict opportunities for such innovation, and thereby utilize limited resources to support and promote scientific and technological progress. Technological gaps refer to undeveloped areas within a technological field with strong potential for technological innovation, manifested as concepts or combinations of concepts not yet laid out in existing patents.

Current research on technological gaps primarily focuses on gap identification, but few studies examine future development trends of these gaps, thus failing to provide further consultation for science and technology management decision-makers. Building upon gap discovery, this study further identifies mature directions with similar basic research classifications among numerous established research and technological directions both domestically and internationally. Mature directions refer to themes that have formed independent topics in both papers and patents, where basic research and technological applications have formed complete systems. Due to their similarity with technological gaps in basic research, these mature directions may provide references in experimental conditions, methodologies, and other areas, thereby forming cross-fusion with gap points in basic research and further promoting the development of gap research.

This paper uses key sentences rather than keywords to characterize papers and patents, as they provide more semantic information and facilitate interpretation of each cluster' s theme by domain experts. Through clustering methods, we can discover China' s technological gaps and their development and fulfillment by comparing thematic clustering distributions of Chinese and foreign papers and patents. This study further extracts other thematic directions with similar basic research to mature directions, which may also provide references for technological gap development and support subsequent development predictions.

To measure relationships between technologies, researchers primarily employ co-classification analysis, using classification attributes including International Patent Classification (IPC) codes, standard industry classification codes, and keywords. Some studies construct domain co-occurrence networks and use cosine similarity to calculate technology domain fusion at the IPC classification granularity. However, such classification granularity is relatively coarse, revealing broad technology domain fusion directions. This paper uses WOS categories for papers and IPC classifications for patents to obtain basic research classification distributions and technology application classification distributions for each cluster. By calculating vector similarity of classification distributions, we avoid the problem of coarse classification granularity. Additionally, this paper simultaneously uses embedding vector similarity to measure semantic similarity between cluster themes.

By comparing Chinese and foreign genomics technology distributions in animal husbandry and veterinary medicine, we can identify Chinese technological gaps and predict future trends in both basic research and technology application domains, helping science and technology management decision-makers grasp the entire process and new directions of scientific and technological development, which is significant for guiding scientific research and technological development.

2 Related Work

With the rapid development of science and technology, technological improvement increasingly exhibits recursive characteristics—new technologies cannot emerge from nothing. Current technological gap discovery methods mainly include two approaches. The first is patent map-based methods, which require constructing keyword association matrices or keyword vectors, then performing principal component analysis or using self-organizing maps, generative topological mapping, and other methods for mapping to ultimately generate patent maps. These methods analyze current patents in a field from technology and efficacy dimensions to find undeveloped areas. However, the disadvantage is that efficacy definition is sometimes difficult to distinguish.

The second approach uses technology efficacy matrices, which require manual indexing of technology types and efficacy for each patent, with technology types as the vertical axis and efficacy as the horizontal axis to draw matrices. This method can find new areas not yet researched, but the drawback is the substantial workload and time consumption for manual indexing, as well as subjective errors.

Some studies use co-classification analysis to identify technology fusion opportunities. When a technology gap exists only in China, foreign technologies with similar technology application domains can provide references for China' s gap development and fulfillment. When a technology gap restricts global technological development, seeking support in basic research may be a solution. Some studies construct domain co-occurrence networks and use cosine similarity to calculate technology domain fusion at the IPC classification granularity, but these reveal relatively broad technology domain fusion directions due to coarse classification granularity.

3 Research Methods

This study uses key sentence embedding methods to mine key sentences, clusters sentence vectors to discover technological gaps, performs comparative analysis and interpretation of clustering results, and uses similarity analysis to identify thematic directions that can easily form cross-fusion with technological gaps, ultimately completing technology gap discovery and prediction to achieve consultation purposes.

3.1 Key Sentence Embedding and Clustering

The TextRank algorithm is used to extract key sentences from abstracts of papers and patents. Key sentences represent semantic key information extracted from text collections. TextRank is an unsupervised key sentence extraction method that treats sentences in abstracts as nodes in a network and co-occurrence relationships between sentences as edges, using co-occurrence similarity between sentence pairs as edge weights. The algorithm is commonly used to extract keywords, keyword groups, and key sentences from given texts.

The extracted key sentences serve as input for the Sentence-BERT algorithm. Similar to word embedding algorithms like word2vec and FastText, Sentence-BERT can represent text as numerical vectors to obtain key sentence embeddings for subsequent learning tasks. Compared to BERT, Sentence-BERT uses siamese and triplet networks on top of BERT, employs pooling layers to fix sentence vector length, and generates semantic sentence vectors rather than using contextual representations of each token. This provides higher computational efficiency and faster sentence similarity calculation.

K-means clustering algorithm is used for thematic clustering. K-means is a common thematic clustering algorithm that ensures minimal distance between sentence vectors within each cluster while maximizing distance between clusters. Although simple, K-means requires pre-setting the number of clusters, and its accuracy depends heavily on data structure characteristics. However, K-means has the advantage of low computational cost and can mine more meaningful thematic categories through manual parameter adjustment. Since patent technology domain data volume is relatively small, DBSCAN methods can only complete partial data clustering, leaving large amounts of data unassigned to any cluster, making K-means more suitable for this research.

Domain experts use key sentences and keywords/IPC classification information from clusters to perform semantic annotation of themes. By comparing Chinese and foreign paper and patent clustering distributions, we can discover Chinese technological gaps and predict their development.

3.2 Technology Gap Discovery and Prediction

Considering that science and technology management decision-makers need to grasp scientific and technological opportunities from a national perspective, this study conducts comparative research on Chinese and foreign science and technology thematic distributions from the national dimension to help decision-makers understand science and technology development trends and formulate science and technology policies.

This paper compares clustering distributions of Chinese and foreign papers and patents to identify Chinese technological gaps. By determining whether technological applications exist in patents, we can judge whether a cluster theme direction constitutes a technological gap. The mean of paper and patent abstract key sentence embedding vectors represents each cluster' s technical semantic distribution, while vectors of disciplines or IPC classifications represent basic research distribution or technology application distribution. Classification vector length is determined by the total number of possible disciplines or IPCs across all clusters, with values in each dimension representing the number of papers or patents in that cluster belonging to the corresponding classification.

To explore semantic similarity between paper and patent themes and further identify mature directions, this study first uses Spearman correlation coefficient to measure semantic similarity between Chinese and foreign paper and patent abstract clusters, constructing a semantic similarity network describing thematic similarity between multi-source data. Research shows that for word/sentence vectors, rank-based correlation calculation methods are more robust than other numerical similarity algorithms. Using Spearman correlation coefficient to measure similarity between two cluster center vectors, where X and X' represent center vectors of two clusters (the average of all abstract sentence vectors for papers or patents in the cluster), and R_X and R_X' represent ranks of X and X' sorted from smallest to largest, the coefficient _X represents semantic similarity between the two clusters.

Spearman correlation coefficient is also used to measure classification similarity between Chinese and foreign papers and patents, constructing paper discipline classification similarity networks and patent IPC classification similarity networks. Y and Y' represent discipline/IPC classification vectors, R_Y and R_Y' represent their ranks, and _Y represents classification similarity between two clusters. Smaller classification similarity values indicate more similar discipline/IPC classifications between themes, while larger values indicate greater differences.

From the semantic similarity network, we identify Chinese and foreign paper and patent clustering themes with strong similarity, extract mature directions, and further explore basic research and technology application directions in discipline classification networks and IPC classification networks that may help develop or even fill technological gaps.

4 Experimental Results and Analysis

Using abstract information from SCI papers collected through the Web of Science platform and patent data from the Derwent Innovations Index, we conducted collaborative analysis of basic research and technology applications in genomics between China and foreign countries. Chinese and foreign papers were divided based on address fields, while Chinese and foreign patents were divided based on applicant country code fields. The retrieval time range was 2001-2022.

4.1 Technology Gap Discovery

China has published approximately one-third of global papers in animal husbandry and veterinary genomics, with an even more significant advantage in patent quantity. However, by analyzing thematic distributions of Chinese and foreign papers and patents, we can identify Chinese technological gaps.

Paper thematic analysis reveals three main directions: 1. **Genome sequencing:** Both Chinese and foreign scholars have produced substantial basic research results involving sequencing of livestock, poultry, aquatic species, and viruses. 2. **Genome-wide association analysis (GWAS):** Chinese basic research primarily focuses on livestock and poultry, fecal microorganisms, and genetic diversity and important trait gene mining in wild relatives. Foreign GWAS basic research mainly centers on three aspects: association between genomics and genetic development, improving meat, egg, and milk quality and yield through genomic breeding, and disease resistance molecular mechanisms. 3. **Multi-omics integration analysis:** Foreign research proportion in this area is significantly larger than China' s. Although currently only involving integration of genomics, transcriptomics, and proteomics, it is ahead of Chinese basic research in this aspect.

Patent thematic analysis reveals: - Chinese patents mainly cover major Chinese livestock and poultry species and aquatic products, focusing on detection technologies. - Foreign patents have more comprehensive layout, with 8 themes related to detection and 6 related to transgenesis, covering gene integration, modification, selection methods, and viral genetic engineering. - China has only 2 transgenesis-related patent themes with small application numbers. - Foreign countries have corresponding patent layouts for livestock genomic selection breeding technology, while China remains relatively passive in this technology market.

Key findings: 1. **Transgenic breeding technology** represents a technological gap in China. 2. **Multi-omics integration analysis technology** is another gap area, lacking sufficient basic research support in China. 3. Although China has not yet formed favorable basic research support in multi-omics integration, future development may fill this research gap through talent introduction and frontier tracking.

The two strongest similarity edges from the semantic similarity network are: - Chinese paper theme 1 → Foreign paper theme 1 → Chinese patent theme 2 → Foreign patent theme 9 - Chinese paper theme 8 → Foreign paper theme 6 → Chinese patent theme 5 → Foreign patent theme 0

These edges characterize mature research and technology directions. The first edge represents livestock and poultry GWAS, while the second represents livestock and poultry virus genomics analysis.

4.2 Technology Gap Prediction

For GWAS direction: The theme is derived from genetics, dairy science, and animal science. From a technology application perspective, GWAS technologies are mainly applied in enzyme, microorganism detection, testing, and preparation methods. Related themes include Chinese patent theme 8 (aquatic microsatellite marker detection), Chinese patent theme 7 (genotype detection of major genes for important traits in pigs), foreign patent theme 6 (livestock breeding methods), and foreign patent theme 1 (cloning of aquatic actin genes and promoters). China has breeding directions for multiple livestock and aquatic species that may form cross-fusion with GWAS in basic research, promoting each other' s development.

For virus genomics analysis direction: The theme is derived from virology and veterinary medicine. Related themes include Chinese paper theme 7 (multi-omics integration research), foreign paper theme 7 (livestock and poultry miRNA sequence research), Chinese paper theme 4 (microbial research), and foreign paper theme 9 (Illumina sequencing research). Virology and veterinary medicine have already formed cross-fusion in multiple research directions, and future multi-omics integration will likely first merge with livestock and poultry virus genomics analysis. From a technology application perspective, virus genomics analysis is mainly applied in mutation or genetic engineering and pharmaceutical preparations containing antigens or antibodies. Related themes include Chinese patent theme 6 (transgenic and cloned embryo methods) and foreign patent theme 1 (isolation of viral genomes).

Although China has not yet formed independent technology application themes in transgenic breeding, by drawing on foreign experience through talent introduction and other means, and with further cross-fusion between transgenic technology and GWAS in basic research, China will have significant development opportunities in transgenic breeding technology. This may promote GWAS to gain greater application space in breeding. Similarly, although multi-omics integration technology has not yet formed independent technology application themes globally, through cross-fusion with livestock and poultry virus genomics analysis in basic research, it will likely make significant contributions in livestock and poultry virus genomics technology, potentially even promoting transgenic livestock and poultry multi-omics integration technology to become a future R&D hotspot.

4.3 Results Validation

To validate the effectiveness of our method, we compared it with other scholars' research findings. Article [25] identified hotspots such as genome assembly, genome sequencing, and database construction through scientific paper analysis, which echo our clustering themes. Their predictions on theoretical innovation and holistic research in genomics align with our identified technology gaps in multi-omics integration research. Article [26] introduced genomics applications in developing country livestock industries and future opportunities, discussing GWAS, pathogen detection, vaccine development, and related technologies that correspond to the technology gaps discovered by our method. These studies can corroborate our conclusions and validate our method from different perspectives. Additionally, our identification of technology gaps is based on diverse data sources (both Chinese and foreign papers and patents), providing a more solid data foundation and thus greater specificity and persuasiveness.

5 Conclusion

This study conducted a comparative analysis of technological gaps in animal husbandry and veterinary genomics between China and foreign countries, identifying Chinese technological gaps and providing consultation suggestions for future development.

Key findings: - China has published numerous papers and patents, but its technical architecture layout is less complete than foreign countries. - Chinese paper topics are more comprehensive than patent topics. - China lacks sufficient basic research support in multi-omics integration and association research. - China's transgenic breeding technology field is also a technological gap. - Future technology fusion points may emerge between transgenic breeding and GWAS, as well as between multi-omics integration and livestock/poultry virus genomics analysis.

Limitations: 1. Manual analysis and interpretation of relationships between scientific papers and technological patents remains time-consuming and laborintensive. Future research will design more automated methods to construct correlations between these two data objects. 2. Expert interpretation of clustering themes can be improved. Future work may incorporate more data sources and label information to reduce manual annotation while providing quantitative accuracy metrics for result validation.

[Figure 1: see original paper] [Figure 2: see original paper] [Figure 3: see original paper]

References: [1] CUI Y Y, ZHAO B B, XIA F. Technology opportunity recognition algorithm and decision assistance for non-drug antidepressant field in China[J]. Mathematical problems in engineering, 2022, 2022: 1-13. [2] WU F F, CHEN M, HUANG L C. Identification of patent vacuums in 3D bioprinting based on GTM[J]. Journal of intelligence, 2015, 34(3): 58-64. [3] ZHAO L, LI P, LI P G. Development trend of seawater source heat pump technology in China based on patent IPC co-occurrence analysis[J]. Refrigeration & air conditioning, 2023, 37(2): 320-325. [4] GE X P. Research of the patent map and its application in the biomedical field[D]. Suzhou: Soochow University, 2010. [5] LEE S, YOON B, PARK Y. An approach to discovering new technology opportunities: Keyword-based patent map approach[J]. Technovation, 2009, 29(6/7): 481-497. [6] YOON B U, YOON C B, PARK Y T. On the development and application of a self-organizing feature map-based patent map[J]. R&D management, 2002, 32(4): 291-300. [7] GONG X. Analysis of the technology hotspot and the technical gap in the technical field of wastewater treatment[J]. China chemical trade, 2017, 9(24): 91-93. [8] LUO K, YUAN X D. A study on the technology convergence trend of patent based on LDA and social network—An example of joint robot[J]. Journal of intelligence, 2021, 40(3): 89-97. [9] MIHALCEA R, TARAU P. TextRank: Bringing Order into Texts[C]// Proceeding of the 2004 Conference on Empirical Method in Natural Language Processing, Association for Computational Linguistics Special Interest Group on the Lexicon, 2004: 404-411. [10] MIKOLOV T, SUTSKEVER I, CHEN K, et al. Distributed representations of words and phrases and their compositionality[C]// Proceedings of the 26th International Conference on Neural Information Processing Systems—Volume 2. New York: ACM, 2013: 3111-3119. [11] BOJANOWSKI P, GRAVE E, JOULIN A, et al. Enriching word vectors with subword information[J]. Transactions of the association for computational linguistics, 2017, 5: 135-146. [12] REIMERS N, GUREVYCH I. Sentence-BERT: Sentence embeddings using Siamese BERT-networks[C]// Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP). Stroudsburg, PA, USA: Association for Computational Linguistics, 2019: 3982-3992. [13] KIROS R, ZHU Y K, SALAKHUTDINOV R, et al. Skip-thought vectors[J]. Advances in neural information processing systems, 2015, 28(1): 1-11. [14] LOGESWARAN L, LEE H. An efficient framework for learning sentence representations[J]. arXiv: 1803.02893, 2018. [15] PAEA S, BAIRD R. Information architecture (IA): Using multidimensional scaling (MDS) and K-means clustering algorithm for analysis of card sorting data[J]. Journal of usability studies archive, 2018, 13: 138-157. [16] ESTER M, KRIEGEL H P, SANDER J, et al. A density-based algorithm for discovering clusters in large spatial databases with noise[C]// Proceedings of the Second International Conference on Knowledge Discovery and Data Mining. New York: ACM, 1996: 226-231. [17] ZHANG Z Y, WANG W F, MA L R, et al. A review of pre-training text summarization studies[J]. Journal of Yan' an university (natural science edition), 2022, 41(1): 98-104. [18] YAN D W, MEI X R, YANG X F, et al. Research on microblog text topic clustering based on the fusion of topic model and word embedding[J]. Journal of modern information, 2021, 41(10): 67-74. [19] BI DOKI M, MOOSAVI M R, FAKHRAHMAD M. A semantic approach to extractive multi-document summarization: Applying sentence expansion for tuning of conceptual densities[J]. Information processing & management, 2020, 57(6): 102341. [20] DE MOURA VENTORIM I, LUCHI D, RODRIGUES A L, et al. BIRCHSCAN: A sampling method for applying DBSCAN to large datasets[J]. Expert systems with applications, 2021, 184: 115518. [21] ZHELEZNIAK V, SAVKOV A, SHEN A, et al. Correlation coefficients and semantic textual similarity[J]. arXiv: 1905.07790, 2019. [22] LI X Y, ZHU Z M, YANG X Q. Comparative analysis on the research fronts of genomics between China and foreign[J]. China basic science, 2016, 18(3): 42-46, 50. [23] BOADU F, DU Y F, XIE Y. Knowledge transfer received, entrepreneurial opportunity type, environmental dynamism, and innovative performance by overseas subsidiaries in China[J]. Technology analysis & strategic management, 2023, 35(3): 237-254. [24] WANG D X, WANG Z, ZHAO F, et al. State-of-the-art and prospect of technical standards for the ships powered by hydrogen fuel cells[J]. Journal of transport information and safety, 2023, 41(2): 157-167, 178. [25] LIU X Y, ZHANG S W, SHAN X H. Research on complementarity and similarity of technology convergence[J]. Complex systems and complexity science, 2023, 20(1): 81-87. [26] Research on technology fusion analysis framework based on patent co-classification: Taking synthetic biology as an example[J]. Information studies: Theory & application, 2021, 44(10): 179-184.

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