BERT-LDA-Based Key Technology Identification Method and Its Empirical Study: A Case Study of Agricultural Robots (Postprint)

Authors: Wang Xiuhong, high sensitivity

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

Abstract

[Purpose/Significance] Effective key technology identification methods can better support the identification, prediction, and R&D of key technologies at all levels and tiers. [Method/Process] This paper proposes a key technology identification method based on the BERT-LDA model, which combines BERT with LDA to compensate for the deficiency of contextual semantic information in the standalone LDA topic model, and conducts an empirical study using agricultural robots as an example. The specific processes include: constructing BERT semantic feature vectors and LDA topic feature vectors based on Python, concatenating them in high-dimensional space, and using an autoencoder to learn the low-dimensional latent space representation of the concatenated vectors; implementing semantic association clustering using the K-means algorithm on the latent space representation to obtain a two-dimensional clustering visualization and key technology topic word cloud; conducting key technology determination; in the field of agricultural robot technology, comparing with patent analysis results based on Derwent TI patent software and the list of key common technologies for agricultural equipment in the "Made in China 2025" key area technology roadmap to empirically validate the effectiveness of this method. [Results/Conclusion] The research demonstrates that: the BERT-LDA model improves the coherence of topic clustering and the precision of fine-grained partitioning; it achieves excellent precision and recall rates for key technology identification; it exhibits good inclusivity and compatibility for identifying literature datasets from different databases and publication types, with strong adaptability; and it can be widely applied to the identification of various key technologies.

Full Text

Introduction

Since the 21st century, global technological innovation has entered an unprecedented period of intense activity, with a new round of technological revolution and industrial transformation reshaping the global innovation landscape and economic structure. Science and technology profoundly influence national destiny and people' s well-being. President Xi Jinping emphasized at the Academicians' Conference that China should "take key generic technologies, frontier leading technologies, modern engineering technologies, and disruptive technological innovations as breakthroughs, dare to forge paths never taken before, strive to achieve independent control of core technologies, and firmly grasp the initiative in innovation and development." Currently, China' s technological innovation faces increasingly prominent shortcomings in vision, capability, resource allocation, and institutional policies, and the situation of core technologies being controlled by others has not fundamentally changed. In this new round of technological revolution, identifying and forecasting key technologies has become particularly important for seizing opportunities, developing core technologies, and becoming a global science and innovation hub.

Numerous scholars have conducted research on key technology identification and forecasting, achieving substantial results. Studies based on indicator evaluation, patent data, and text mining have guided technological innovation to some extent. However, as demands for technology intelligence deepen, more stringent requirements are imposed on optimizing key technology identification methods. This paper proposes a BERT-LDA-based key technology identification method to improve topic coherence and fine-grained classification accuracy while ensuring precision and recall. This approach enhances inclusivity across different publication types, enabling identification from not only patent literature (like TI/Thomson Innovation) but also from scientific abstracts across different databases and publication types, thereby improving objectivity and timeliness.

2 Literature Review

Research on identifying key emerging technologies, generic technologies, core technologies, and breakthrough technologies has established a solid foundation, primarily employing three approaches: indicator-based evaluation, patent network analysis, and text mining.

2.1 Indicator-Based Identification Methods

Indicator-based methods systematically analyze key technology characteristics to construct multi-indicator evaluation frameworks. For example, Altuntas et al. evaluated technologies using four indicators: technology lifecycle, diffusion speed, patent rights, and expansion potential. Park et al. calculated a prospect

index based on growth potential, influence, and marketability to identify core patents in user interface and experience technologies. Lee et al. proposed machine learning methods using multiple patent indicators for early-stage emerging technology identification. Liu et al. developed a three-dimensional evaluation framework integrating persistence, community, and growth. Jiang and Wei constructed a generic technology identification framework based on patent analysis, examining technological impact scope and research stage across four dimensions: fundamentality, externality, integrability, and advancement. Yang and Yang explored core technology identification using patent data and indicator systems, while Song et al. built an identification index system for novelty, persistence, community, and growth, introducing emerging scores and LDA models to identify emerging terms and topics.

These methods integrate multiple patent indicators to form evaluation systems, offering scientific validity. However, some indicator rules require expert-defined thresholds and scoring, making results dependent on subjective expert judgments and compromising objectivity.

2.2 Patent Network-Based Identification Methods

Patent network analysis combines social network theory with patent analysis, using citation, co-citation, and coupling algorithms to analyze technology evolution and knowledge flow networks. Cho and Shih identified five core and emerging technologies in Taiwan using USPTO patents (1997-2008) through citation network analysis. Ho et al. constructed patent citation networks to identify frequently cited core patents through path analysis. Kuusi et al. used patent coupling networks to forecast breakthrough technologies in nanotechnology. You et al. proposed a two-layer citation network model based on knowledge transfer between patents and subclasses to predict technology trends. Li et al. built an emerging technology identification model using patent citation coupling clustering and applied it to nanotechnology. Yang et al. developed a key technology identification framework using patent co-citation clustering and portfolio analysis.

These methods avoid subjective cognitive differences and objectively identify key technologies but rely heavily on actual citation data. The lag between patent application, publication, and citation creates delays, raising questions about the validity and accuracy of identification results.

2.3 Text Mining-Based Identification Methods

Text mining approaches analyze paper and patent content using natural language processing techniques like text clustering, SAO structures, and LDA topic models to extract implicit technical knowledge. Chen et al. used topic models to generate topic-year weight matrices and trend coefficient sequences to quantitatively estimate development trends. Yang et al. employed semi-supervised topic clustering to distinguish new from traditional topics in 3D printing. Zhou

et al. proposed a data augmentation and deep learning method to address training sample scarcity. Li et al. extracted SAO structures from patent claims and clustered patents using improved semantic similarity algorithms. Zhou et al. developed a machine learning topic model approach for high-throughput fusion processing of paper and patent data. Chen et al. established a key generic technology identification framework using text mining and technology evolution analysis.

Text mining offers strong replicability, cost-effectiveness, and objectivity but often suffers from weak semantic associations between keywords, ignoring contextual semantics and producing less interpretable results. Effective methods must incorporate semantic relationships to improve interpretability.

2.4 The BERT Model

Recent natural language processing models have achieved excellent results in text semantic analysis. In 2018, Google released BERT (Bidirectional Encoder Representations from Transformers), a deep bidirectional, unsupervised language representation model pre-trained on plain text corpora. Unlike traditional unidirectional left-to-right models, BERT' s bidirectional architecture better learns word relationships and detects linguistic nuances.

BERT has been widely applied in document semantic research, Chinese word segmentation, part-of-speech tagging, named entity recognition, and topic extraction. Asgari-Chenaghlu et al. used BERT to provide contextual semantic relationships for social network data, enhancing topic visualization. Thompson and Mimno demonstrated that BERT with clustering performs as well as or better than LDA, even with large datasets. Abuzyed et al. showed that Arabic BERT models outperform LDA and NMF for topic modeling.

Some studies have combined BERT with LDA for sentiment analysis, text classification, and machine translation, but few have empirically validated the approach for technology identification. While LDA is widely used for identifying key technology topics, it struggles with contextual semantics, word ambiguity, and limited semantic expression. This study addresses these limitations by proposing a BERT-LDA model for key technology identification, using agricultural robotics patents for empirical validation. The approach compares results with LDA, TF-IDF, Word2Vec, and BERT alone using topic coherence and silhouette coefficients, and validates effectiveness against TI text mining methods and the "Made in China 2025" roadmap.

3 Theory and Methods

The BERT-LDA model for key technology identification involves four systematic processes: dataset construction and preprocessing, BERT-LDA text vectorization, semantic association clustering and visualization, and key technology determination [Figure 1: see original paper].

3.1 Dataset Construction and Preprocessing

Identify appropriate databases and construct search queries to retrieve literature in the target technology domain. Extract key information including titles, abstracts, patent numbers, IPC codes, and inventors. Preprocess text content through tokenization, stopword removal, and stemming using Python' s NLTK and stopwords packages.

3.2 BERT-LDA Text Vectorization

Train LDA topic feature vectors and BERT semantic feature vectors on large unlabeled datasets, then fuse them to generate BERT-LDA representations.

(1) Constructing BERT Semantic Feature Vectors: Use BERT for word embedding of preprocessed data. In the Transformer encoder, multi-head attention processes tokens into vectors passed through residual connections and normalization layers, then through feedforward and residual networks to extract BERT semantic feature vectors. For document d_i, each token is mapped to three vectors: token embedding ω, segment embedding σ, and position embedding ρ. The BERT semantic feature vector d_m is defined as:

$$d_m = w_{ij}(\omega + \sigma + \rho)$$

(2) Constructing LDA Topic Feature Vectors: LDA is a three-level Bayesian probability model containing words, topics, and documents. It reduces dimensionality by representing a tokenized document as a topic distribution (n_0 feature vector topics). The joint distribution of the LDA topic model is defined as:

$$P(w_i, z_i, \theta_i, \Phi|\alpha, \beta) = \prod_{j=1}^{N_i} P(w_{ij}|\phi_{z_{ij}})P(z_{ij}|\theta_i) \cdot P(\theta_i|\alpha) \cdot P(\Phi|\beta)$$

where α is the hyperparameter for per-document topic prior distribution, θ_i is Dirichlet-sampled from α, β is the hyperparameter for per-topic word prior distribution, and Φ is Dirichlet-sampled from β. Gibbs sampling estimates parameters iteratively until convergence. The model outputs topic distribution matrices with the same dimensionality as BERT semantic feature vectors. Topic feature vector μ is calculated from cosine distances between high-frequency topic words and documents.

(3) Vector Concatenation: The Transformer encoder learns semantic and syntactic relationships, which are preserved by concatenating BERT semantic and LDA topic feature vectors into a new input vector d'_m containing both word-level and document-level semantic features:

$$d'_m = \{\mu; d_m\}$$

where ";" denotes vector concatenation.

3.3 Semantic Association Clustering and Visualization

Since concatenation occurs in high-dimensional sparse space, an autoencoder learns a low-dimensional latent representation that concentrates information. K-means clustering is applied in this latent space to group semantically and thematically similar terms. K-means is ideal for large-scale clustering with minimal parameters (only K, the number of clusters), simple implementation, high efficiency, and no need for preliminary distance matrix computation. Parameter K naturally corresponds to the number of topics.

The optimal K is determined using perplexity, which decreases with more topics and stabilizes at the optimal number. Top-ranked topic words are visualized to identify key technologies. The BERT-LDA model combines contextual semantics with LDA' s strengths, producing superior topic vectors with better granularity and clustering precision.

4 Empirical Study

Using patent literature, this study collects and processes data from authoritative Derwent patent databases and TI analysis software, comparing results with TI' s thematic identification and the "Made in China 2025" agricultural equipment key generic technology list to validate precision and recall.

4.1 Data Collection and Preprocessing

Derwent patents in agricultural robotics serve as the dataset. Based on literature review and expert knowledge, the search query was constructed as: ((TS=(agricult* OR crop OR crops OR fruit OR fruits OR vegetable* OR harvest* OR seedling*)) OR (MAN=(X25-N* OR X22-X11 OR X22-P09 OR Q19-G OR T06-D01* OR A12-W04* OR X25-X02)) OR (IP=(A01B* OR A01C* OR A01D* OR A01F* OR A01G* OR A01M-021))) AND ((TS=(robot* OR manipulator* OR "mechanical arm" OR "mechanical arms" OR "mechanical hand" OR "mechanical hands")) OR IP=(B25J*) OR MAN=(X25-A03E* OR T06-D07B* OR V03-U14* OR V04-M30R* OR V04-Q30R* OR V06-U05* OR V04-R04F1* OR X27-U* OR S05-B07)) NOT (IP=(A01G-005* OR A01G-023)) OR (MAN=(X25-N02* OR T06-D01C*))). The search covered all agricultural robotics patents published before December 6, 2020, yielding 8,957 patents after processing.

Preprocessing involved: (1) Data cleaning of DWPI abstracts, removing 45 patents with missing abstracts (8,912 retained); (2) Tokenization, punctuation/number filtering, and noise reduction including lowercase conversion, spell checking, singular/plural unification, synonym merging, full name/abbreviation standardization, stopword removal (e.g., a, for), proprietary descriptors (e.g., comprise, involve), academic vocabulary (e.g., novelty, use, advantage), and domain-specific high-frequency interference words (e.g., robot, agriculture); and (3) Stemming extraction.

4.2 BERT-LDA Model Results and Analysis

To enhance BERT' s adaptability, the base Google BERT model was fine-tuned on the agricultural robotics patent abstract corpus (768-dimensional embeddings). The improved BERT model and LDA were trained on cleaned DWPI abstracts, with vectors concatenated and clustered to identify key technology topics. Perplexity analysis [Figure 2: see original paper] indicated optimal topic number $K = 10$ when perplexity stabilized.

UMAP (Uniform Manifold Approximation and Projection) was used for visualization due to its superior performance in preserving global structure and scalability [Figure 3: see original paper]. The 10 identified topics showed clear distribution with high intra-cluster coherence, demonstrating excellent clustering performance.

The top 50 topic words per theme were visualized, with the top 10 words and probabilities shown in Table 1 and word clouds in Figure 4 [Figure 4: see original paper]. Each topic represents a research hotspot in agricultural robotics.

Topic Interpretation: - **Topic 1** (connect, fruit, pick, harvest, arm, mechanism, automatic, control, collect, motor) corresponds to "Automatic Picking Devices," validated by patents like Jiangsu University' s CN101273688-A for orange-picking robots. - **Topic 2** (sensor, position, device, signal, direction, navigation, boundary, distance, detect, data) maps to "Target Position Detection and Localization," matching patents like Husqvarna' s EP3346348-A1 for robot guidance systems. - **Topic 3** (vehicle, control, unit, autonomous, drive, wheel, direction, motor, path, position) represents "Autonomous Navigation and Path Planning," consistent with patents from Bosch, iRobot, and John Deere. - **Topic 4** (method, vehicle, autonomous, area, path, unmanned, navigate, mobile, signal, system) also relates to navigation. - **Topics 5-10** correspond to "Seedling Transplanting Mechanisms," "Manipulator Control Devices," "Grafting," "Lawn Mowers," "Pruning Tools," and "Irrigation Equipment" respectively .

4.3 Key Technology Determination

The BERT-LDA model identified three core technologies in agricultural robotics:

(1) End-Effectors: Integrating Topics 1, 5, 6, 7, and 10 reveals end-effector technology as critical. Comprising mechanical devices and sensors (grippers, collision sensors, rotary connectors, pressure tools), end-effectors perform picking, transplanting, and spraying operations. Their design must account for the complex, specialized agricultural environment to ensure operation quality, requiring innovation for versatility, precision, flexibility, and controllability.

(2) Target Detection and Localization: Combining Topics 2, 5, and 7 shows this as a prerequisite for robotic operations. Primarily using machine vision (originating in the US), this technology faces challenges from lighting

conditions, occlusion, and individual differences. Future research should integrate machine vision with other technologies to improve image acquisition and processing algorithms for enhanced accuracy.

(3) Autonomous Navigation and Path Planning: Synthesizing Topics 3, 4, 8, and 9 identifies this as essential for robotic mobility. Current methods mainly use visual navigation and hybrid approaches. Complex environments, random target distribution, and unpredictable dynamics demand more robust navigation and path planning capabilities.

Model evaluation using topic coherence (C_V Coherence) and silhouette coefficients confirms BERT-LDA' s superiority. With coherence of 0.508 and silhouette coefficient of 0.15, it significantly outperforms TF-IDF (0.458, 0.006), Word2Vec (0.478, 0.071), LDA (0.481, 0.054), and BERT alone (0.453, 0.150) . The 2D clustering visualization [Figure 5: see original paper] shows clear inter-topic boundaries for BERT-LDA, while other methods exhibit overlapping clusters.

4.4 Validation of Results

Derwent Innovation' s TI platform provides authoritative patent data and analysis, offering intelligent search, analysis, and visualization across 156 countries/regions. Its manually rewritten English abstracts eliminate cross-language barriers. Since TI' s algorithms are proprietary, comparison focuses on results.

The TI patent map [Figure 6: see original paper] visualizes technology landscapes, with peak heights representing document density. Comparing TI' s text clustering with BERT-LDA results shows 90% content consistency across the top 10 key technology themes (fruit/vegetable picking devices, target detection/localization, autonomous navigation/path planning, seeding/transplanting devices, manipulator control, grafting, lawn mowers, pruning tools, and liquid spraying equipment). Detailed feature word comparison confirms BERT-LDA' s effectiveness and high precision/recall.

Further validation against the "Made in China 2025" agricultural equipment key generic technology list shows alignment with end-effectors, detection/localization, and navigation/path planning technologies, confirming result consistency with national strategic priorities.

5 Conclusion and Outlook

The BERT-LDA-based key technology identification method demonstrates enhanced topic coherence, fine-grained classification accuracy, and high precision/recall. Compared with LDA, TF-IDF, Word2Vec, and BERT alone, the fused model leverages contextual semantics to produce more coherent and interpretable results. Validation against TI analysis and the "Made in China 2025" roadmap confirms its effectiveness.

The model exhibits strong inclusivity and compatibility across different databases and publication types (patents, journal articles, conference papers, dissertations, reports). It can integrate multilingual datasets from WOS, EBSCO, ScienceDirect, IEEE, and other sources for comprehensive key technology identification, substantially improving recall while maintaining precision.

5.1 Research Conclusions

BERT-LDA considers textual semantics and context, producing more semantically coherent feature words within topics and improving result interpretability. The method ensures high precision and recall while demonstrating better inclusivity and compatibility than existing models. Empirical results in agricultural robotics validate its effectiveness through topic coherence, silhouette coefficients, and 2D clustering visualization. Comparison with authoritative TI results and national technology roadmaps confirms alignment with strategic priorities.

5.2 Research Outlook

To enable comparison with TI, this study used Derwent patent abstracts. Future research will integrate multi-source literature (journal articles, conference papers, reports) using BERT-LDA for more comprehensive identification. Data collection and preprocessing will be optimized through enhanced cleaning, expanded stopword lists, and improved stemming. To further improve interpretability, SAO structures may be introduced to explicitly represent problem-solution relationships in technical aspects. Subsequent studies will also incorporate expert surveys and indicator evaluation methods to refine key generic technology determination.

References

[1] ALTUNTAS S, DERELI T, KUSIAK A. Forecasting technology success based on patent data[J]. Technological forecasting and social change, 2015, 96(7): 202-214.

[2] PARK I, PARK G, YOON B, et al. Exploring promising technology in ICT sector using patent network and promising index based on patent information[J]. ETRI journal, 2016, 38(2): 405-415.

[3] LEE C, KWON O, KIM M, et al. Early identification of emerging technologies: a machine learning approach using multiple patent indicators[J]. Technological forecasting and social change, 2018, 127(2): 291-303.

[4] LIU X, PORTER A L. A 3-dimensional analysis for evaluating technology emergence indicators[J]. Scientometrics, 2020, 124(1): 27-55.

[5] JIANG X, WEI F. Research framework for generic technology identification based on patent analysis[J]. Journal of intelligence, 2015, 34(12): 79-84.

[6] YANG W, YANG D. Research on core technology identification in industries based on patent data: a case study of 5G mobile communication industry[J]. Journal of intelligence, 2019, 38(3): 39-45, 51.

[7] SONG X, GUO Y, XI X. Multi-indicator emerging technology identification based on patent literature[J]. Journal of intelligence, 2020, 39(6): 76-81, 88.

[8] CHO T S, SHIH H Y. Patent citation network analysis of core and emerging technologies in Taiwan: 1997-2008[J]. Scientometrics, 2011, 89(3): 795-811.

[9] HO M H C, LIN V H, LIU J S. Exploring knowledge diffusion among nations: a study of core technologies in fuel cells[J]. Scientometrics, 2014, 100(1): 149-171.

[10] KUUSI O, MEYER M. Anticipating technological breakthroughs: using bibliographic coupling to explore the nanotubes paradigm[J]. Scientometrics, 2007, 70(3): 759-777.

[11] YOU H, LI M, HIPEL K W, et al. Development trend forecasting for coherent light generator technology based on patent citation network analysis[J]. Scientometrics, 2017, 111(1): 297-315.

[12] LI B, CHEN X. Emerging technology identification in nanotechnology based on patent citation coupling clustering[J]. Journal of intelligence, 2015, 34(5): 35-40.

[13] YANG Y, DONG Y, HAN T. Research on key technology identification method based on patent co-citation clustering and portfolio analysis: a case study of crop breeding technology[J]. Library and information service, 2016, 60(19): 143-148, 124.

[14] CHEN H, ZHANG G, ZHU D, et al. Topic-based technological forecasting based on patent data: a case study of Australian patents from 2000 to 2014[J]. Technological forecasting and social change, 2017, 119(6): 39-52.

[15] YANG C, ZHU D, WANG X, et al. Requirement-oriented core technological components' identification based on SAO analysis[J]. Scientometrics, 2017, 112(3): 1229-1248.

[16] ZHOU Y, DONG F, LIU Y, et al. Forecasting emerging technologies using data augmentation and deep learning[J]. Scientometrics, 2020, 123(1): 1-29.

[17] LI X, WANG J, YANG Z, et al. Emerging technology identification based on SAO structure semantic analysis[J]. Journal of intelligence, 2016, 35(3): 80-84.

[18] ZHOU Y, LIU Y, XUE L. A machine learning-based method for emerging technology identification: a case study of robotics technology[J]. Journal of the China Society for Scientific and Technical Information, 2018, 37(9): 939-949.

[19] CHEN W, LIN C, KONG L, et al. Key generic technology identification based on patent literature mining[J]. Information studies: theory & application, 2020, 43(2): 92-99.

[20] DEVLIN J, CHANG M W, LEE K, et al. BERT: pre-training of deep bidirectional transformers for language understanding[J]. arXiv preprint, 2018, arXiv: 1810.04805.

[21] ASGARI-CHENAGHLOU M, FEIZI-DERAKHSHI R, FARZIN-VASHEH L, et al. TopicBERT: a cognitive approach for topic detection from multimodal poststream using BERT and memory-graph[J]. Chaos, solitons & fractals, 2021, 151(10): 111274.

[22] THOMPSON L, MIMNO D. Topic modeling with contextualized word representation clusters[J]. arXiv preprint, 2020, arXiv: 2010.12626.

[23] ABUZAYED A, AL-KHALIFAH H. BERT for Arabic topic modeling: an experimental study on BERTopic technique[J]. Procedia computer science, 2021, 189(11): 191-194.

[24] FU J, GONG Y, LIAN X, et al. News short text classification method based on BERT-LDA[J]. Information technology and informatization, 2021(2): 127-129.

[25] ZHUANG M, LI Y, TAN X, et al. COVID-19 epidemic network public opinion evolution simulation based on BERT-LDA model[J]. Journal of system simulation, 2021, 33(1): 24-36.

[26] LI Y, MAO C, YU Z, et al. Chinese-Burmese bilingual word extraction method fusing topic and contextual features[J]. Small microcomputer systems, 2021, 42(1): 91-95.

[27] BLEI D M, NG A Y, JORDAN M I. Latent Dirichlet allocation[J]. The journal of machine learning research, 2003, 3(1): 993-1022.

[28] MCINNES L, HEALY J, MELVILLE J. UMAP: uniform manifold approximation and projection for dimension reduction[J]. arXiv preprint, 2018, arXiv: 1802.03428.

Author Contributions:

Wang Xiuhong: Conceptualization, methodology, experimental design, supervision, writing-review & editing.

Gao Min: Framework design, data collection, data processing, original draft preparation.

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