International Journal of World Economic Research
ISSN: 3080-9886 (Print), ISSN: 3080-9894 (Online) | Vol. 3, No. 1, 2026

Mapping Patent Topics and Technological Relatedness in Next-Generation Information Technology: Evidence from the Chengdu-Chongqing Economic Circle

Dingxuan Huang, Yu Yang and Binjie Wang

Chongqing University of Technology, Chongqing, China

Abstract: Patent-based evidence is widely used to map regional technological capabilities, but topic identification is often separated from the analysis of technological relatedness. This study examines next-generation information technology in the Chengdu-Chongqing Economic Circle using 90,232 patent records downloaded from the China National Patent Database and covering 2014-2024. Chinese patent texts were encoded with the Sentence-BERT checkpoint all-MiniLM-L6-v2 and analyzed through BERTopic, UMAP, HDBSCAN, class-based TF-IDF (c-TF-IDF), cosine similarity, and hierarchical clustering. The six substantive topics contain 78,511 records; 11,721 observations assigned to the outlier/noise class were excluded from topic-level analysis. The identified topics are communication networks and data architecture (52,707), device structure and installation (10,949), optical sensing and imaging (7,689), metal preparation and processing (2,866), intelligent systems and information integration (2,308), and graphite development and application (1,992). Communication networks account for 67.1% of the classified records and show the strongest pairwise similarity with device technologies (0.96). Optical sensing forms an intermediate hardware-perception branch, while the two material topics are more distinct. Intelligent systems are strongly related to communication and device topics in pairwise comparisons but remain globally differentiated in the hierarchical structure. The findings support a layered interpretation of the regional portfolio spanning digital infrastructure, equipment, sensing, system integration, and enabling materials.

Keywords: Next-generation information technology, patent analytics, Sentence-BERT, BERTopic, technological relatedness, regional innovation.

1. Introduction

Digital transformation is reshaping how manufacturing firms organize production, integrate data, and create value. Rather than representing the adoption of a single tool, it combines information, computing, communication, and connectivity technologies with organizational change [1, 2]. In manufacturing, this transformation is associated with the convergence of the Internet of Things, cloud services, big-data analytics, artificial intelligence, cyber-physical systems, and networked automation [3, 4]. These technologies support real-time monitoring, distributed coordination, predictive decision making, and increasingly intelligent production systems [5-7].

The Chengdu-Chongqing Economic Circle is a major regional development initiative in western China. The national master plan emphasizes coordinated infrastructure, advanced manufacturing, scientific and technological innovation, and the formation of an integrated regional growth pole [8]. This setting is therefore suitable for examining how next-generation information technology is distributed across a large manufacturing and innovation system. A systematic map of the regional technology portfolio can inform specialization, cross-city coordination, and industrial-upgrading policy, but only if the underlying patent field and analytical procedures are transparently defined.

Patent documents are useful for this purpose because they contain standardized bibliographic data and detailed technical descriptions that often become public before technologies are widely commercialized. Patent information has long supported strategic technology management, forecasting, competitive intelligence, and policy evaluation [9-11]. The growth of full-text patent databases has encouraged text mining for concept extraction, technology mapping, and trajectory detection [12, 13]. More recently, text-based similarity measures and neural embeddings have improved the representation of technological proximity beyond classification-code overlap and citation links [14, 15].

Despite these advances, three limitations remain. First, many regional patent studies emphasize counts, classifications, or isolated hotspots rather than the semantic organization of the complete portfolio. Second, conventional bag-of-words topic models may struggle with technical synonyms, contextual meaning, and specialized patent language. Third, topic identification is often separated from the analysis of inter-topic relatedness, limiting the interpretation of how infrastructure, equipment, sensing, intelligent applications, and enabling materials are connected. A further practical problem is that studies frequently omit the search strategy, citation window, embedding model, and clustering parameters needed for replication.

This study addresses these limitations through a unified semantic workflow for patents associated with the Chengdu-Chongqing Economic Circle. It asks two questions: (1) What substantive technology topics characterize the regional portfolio of next-generation information technology patents? (2) How are these topics related, and what functional structure do the reported similarity and clustering patterns reveal? The analysis combines Sentence-BERT and BERTopic with UMAP, HDBSCAN, c-TF-IDF, cosine similarity, and hierarchical clustering. The contribution is primarily empirical and methodological: the study maps a large citation-screened regional portfolio, connects topic extraction with relatedness analysis, and develops a cautious layered interpretation of digital infrastructure, physical devices, sensing, intelligent integration, and material support. Because the available analysis pools 2014-2024, the study examines portfolio structure rather than temporal evolution.

2. Literature Review

2.1. Next-generation Information Technology and Manufacturing Transformation

Research on digital transformation and Industry 4.0 generally treats information technology as a system of mutually reinforcing capabilities rather than a collection of independent tools. Digital connectivity links machines, products, workers, and organizations; data infrastructures convert operational signals into analyzable resources; and artificial intelligence supports prediction, optimization, and autonomous decision making [1, 3, 5]. Reviews of Industry 4.0 identify a common technological core consisting of cyber-physical systems, industrial networks, cloud computing, data analytics, and intelligent automation [4, 5]. Firm-level evidence likewise shows that successful implementation depends on the systemic adoption of both front-end applications and foundational technologies such as connectivity, cloud services, and analytics [6].

This systemic view is particularly important at the regional level. A region may possess strong digital infrastructure without equivalent capability in sensing, equipment integration, or intelligent applications. Conversely, software capability may be constrained by weaknesses in devices, components, and materials. Patent portfolios can reveal such complementarities because they record technical activity across multiple layers of an industrial system. The resulting topic structure should nevertheless be interpreted as a map of documented inventive activity, not as direct evidence of commercialization, inter-firm collaboration, or industrial performance.

2.2. Patent Analytics, Topic Identification, And Technological Relatedness

Patent analytics has evolved from descriptive indicators toward integrated approaches that combine bibliometrics, network analysis, and natural-language processing. Early work established the strategic value of patent information for monitoring competitors, evaluating portfolios, and supporting R&D planning [9]. Bibliometric and patent indicators have also been used to forecast emerging technologies and estimate technological life cycles [10]. Reviews of the field show that modern patent analysis increasingly relies on text mining and visualization because patent documents are too large and technically heterogeneous for manual coding [11].

Text-mining studies have introduced term extraction, document clustering, technology maps, and patent networks [12, 13]. A central challenge is the measurement of similarity. Classification codes are consistent and interpretable, but they may be too coarse or slow to reflect emerging combinations. Citation links identify explicit knowledge relationships, but they are sparse and influenced by legal and examination practices. Text-based measures instead compare the substantive content of patents. Arts et al. demonstrated that text matching can provide a validated measure of technological similarity [14], while Hain et al. showed that embeddings can scale patent-to-patent similarity analysis to large technological networks [15]. Recent surveys further document the growing role of deep learning in patent classification, retrieval, and landscaping [16]. Fine-tuned BERT models have also improved semantic patent classification [17].

For regional innovation studies, technological relatedness also has a spatial and policy dimension. Innovation does not depend on geographic proximity alone; cognitive, organizational, and institutional proximity shape knowledge exchange [18]. Patent citations indicate that knowledge spillovers can be geographically localized [19], while regional knowledge-space research shows that cities diversify along paths conditioned by existing technological capabilities [20]. A topic-based representation can therefore identify closely related branches and more distant capabilities, although semantic similarity alone cannot establish knowledge transfer or policy effectiveness.

2.3. Transformer-Based Topic Modeling and Research Gap

Latent Dirichlet Allocation (LDA) remains a foundational probabilistic topic model, representing documents as mixtures of topics and topics as distributions over words [21]. Although LDA is effective in many applications, bag-of-words representations do not directly encode contextual semantics. Transformer models address this limitation by learning context-dependent representations through self-attention [22]. Sentence-BERT adapts BERT to produce semantically meaningful sentence embeddings that can be efficiently compared using cosine similarity [23].

BERTopic integrates transformer embeddings, dimensionality reduction, density-based clustering, and c-TF-IDF topic representation [24]. This modular design is attractive for patent analysis because semantic representation is separated from topic-word extraction and the number of clusters need not be fixed in advance. UMAP emphasizes local manifold structure during dimensionality reduction [25], while HDBSCAN identifies clusters of varying density and can label weakly connected observations as noise [26]. However, the resulting topics are sensitive to the selected encoder, preprocessing rules, dimensionality-reduction settings, density parameters, random seed, and topic-labeling procedure. Accordingly, a publishable application should report these choices and evaluate topic stability and interpretability rather than relying only on visual plausibility.

3. Data and Methods

3.1. Data Source and Sample Construction

The study uses 90,232 patent records downloaded from the China National Patent Database. The observation period is 2014-2024, the geographic scope is the Chengdu-Chongqing Economic Circle, and the inclusion threshold is at least ten forward citations. All records in the analytical dataset contain a Chinese title and abstract. Because forward citations accumulate over time, this threshold may favor older patents unless citation counts are normalized by filing year or observed over a common citation window.

The downloaded records included Chinese patent titles and abstracts. According to the supplied modeling script, the text was preprocessed with a regular-expression filter that retained Chinese and English alphabetic characters while removing digits, punctuation, and other symbols. Chinese text was segmented with jieba, after which a custom stop-word list and a one-character-token filter were applied. The processed texts were then used in the semantic topic-modeling workflow.

BERTopic and HDBSCAN assigned 78,511 records to six substantive topics; the remaining 11,721 records were treated as Topic -1 outliers/noise and excluded from topic-level percentages and substantive interpretation.

3.2. Semantic Embedding and Topic Identification

The cleaned texts were encoded with Sentence-BERT. At a high level, the transformer encoder constructs contextual token representations through multi-head self-attention. A simplified attention operation is expressed as:

$$\text{Attention}(Q, K, V) = \text{softmax}(QK^T / \text{sqrt}(d_k))V \quad (1)$$

Where Q, K, and V denote the query, key, and value matrices, and d_k is the dimension of the key vectors. Sentence-BERT produces independent sentence-level embeddings that can be reused for clustering and similarity calculation [23]. In the supplied analysis code, sentence embeddings were generated with the pretrained all-MiniLM-L6-v2 checkpoint from the SentenceTransformers library.

The high-dimensional embeddings were projected with UMAP, which constructs a fuzzy topological representation of local neighborhoods and optimizes a lower-dimensional embedding [25]. HDBSCAN was then applied within the BERTopic workflow to identify density-based clusters and isolate weakly connected observations [26]. For the two-dimensional document map, UMAP was configured with n_components = 2 and random_state = 42.

For each cluster, BERTopic concatenates the documents belonging to that cluster and calculates a class-based variant of TF-IDF. In simplified form, the importance of term t in topic c is:

$$\text{c-TF-IDF}(t,c) = \text{tf}(t,c) \times \log(1 + A / f_t) \quad (2)$$

Where tf(t,c) is the frequency of term t in topic c, f_t is its frequency across topic-level documents, and A is the average number of words per topic-level document [24]. The highest-scoring terms were used to assign substantive labels. c-TF-IDF identifies vocabulary that is frequent within a topic and distinctive across topics, consistent with established term-weighting approaches [27]. Topic labels were interpreted with reference to the representative terms and the substantive content of the associated patent records.

3.3. Topic Similarity and Hierarchical Clustering

Pairwise relatedness was measured using cosine similarity. For topic vectors x and y, cosine similarity is defined as:

$$\cos(x,y) = (x \cdot y) / (\|x\| \|y\|) \quad (3)$$

The measure ranges from -1 to 1, with higher values indicating more similar vector directions. Pairwise topic relatedness was calculated from BERTopic's topic_embeddings_ output using the cosine_similarity function in scikit-learn, with Topic -1 excluded. The hierarchical dendrogram was generated with BERTopic's visualize_hierarchy() function. The cosine matrix and dendrogram are therefore interpreted as complementary descriptive outputs: the former captures pairwise semantic proximity, whereas the latter summarizes the global clustering structure.

3.4. Analytical Scope and Interpretation

The empirical design pools records from 2014-2024 and therefore provides a cross-sectional map of topic structure and relatedness. It does not estimate annual topic prevalence, topic birth or decline, or transitions between topics. Claims of technological evolution are consequently avoided. The two-dimensional UMAP plot is used only as an exploratory visualization of local neighborhoods; global distances, cluster areas, and apparent overlap in the plot should not be interpreted as formal statistical tests. Likewise, cosine similarity indicates semantic proximity, not collaboration, knowledge flow, complementarity, or causal synergy.

4. Results

4.1. Patent Topic Structure

The analysis identifies six substantive topics (Table 1), together containing 78,511 classified records. Communication networks and data architecture dominate the classified portfolio, followed by device structure and optical sensing. Intelligent systems form a smaller integration-oriented topic, while metal and graphite technologies form a distinct materials branch. The labels were assigned from the c-TF-IDF terms and the technical content represented in each topic.

Table 1. Patent topics in next-generation information technology

Topic	Interpretive label	Representative keywords	Patents	Share (%)
Topic 0	Communication networks and data architecture	connection; data; configuration; installation; mechanism	52,707	67.13
Topic 1	Device structure and installation technology	configuration; connection; level; installation; mechanism	10,949	13.95
Topic 2	Optical sensing and imaging technology	laser; optical fiber; image; detection; optics	7,689	9.79
Topic 3	Metal preparation and processing	metal; preparation; alloy; structure; material	2,866	3.65
Topic 4	Intelligent systems and information integration	intelligent; module; data; information; smart	2,308	2.94
Topic 5	Graphite development and application	graphite; preparation; rock; material; field	1,992	2.54

Note: Percentages use 78,511 classified records as the denominator. The initial dataset contains 90,232 records; 11,721 records assigned to the Topic -1 outlier/noise class are excluded from the six substantive topics.

Communication networks and data architecture. Topic 0 contains 52,707 records, representing 67.13% of the classified set and 58.43% of the reported initial corpus. Its leading terms emphasize connection, data, configuration, installation, and mechanisms. The topic is therefore interpreted as a broad infrastructure and data-architecture cluster. Because several leading terms are generic engineering expressions, representative-patent and IPC/CPC checks are needed before linking the topic to specific technologies such as 5G, 6G, or industrial Internet systems.

Device structure and installation technology. Topic 1 contains 10,949 records (13.95% of the classified set). Shared terms such as connection, configuration, installation, and mechanism indicate semantic proximity to Topic 0, while the emphasis shifts toward physical equipment and deployment. The topic is cautiously interpreted as a hardware-and-installation layer through which information technologies may be embedded in manufacturing systems.

Optical sensing and imaging technology. Topic 2 contains 7,689 records (9.79%) and is characterized by laser, optical fiber, image, detection, and optics. These terms support an interpretation centered on optical communication, imaging, measurement, and detection. The topic may function as a perception layer linking physical equipment with data-intensive applications, although this functional role is inferred from the vocabulary rather than directly observed in firm or product data.

Metal and graphite materials. Topics 3 and 5 contain 2,866 (3.65%) and 1,992 (2.54%) records, respectively. Their leading terms concern material preparation, alloy structures, graphite, and processing. These topics may capture enabling materials and device-level components, but their presence may also reflect a broad or imperfectly bounded search strategy. Without the original query and IPC/CPC composition, direct attribution to next-generation information technology should remain tentative.

Intelligent systems and information integration. Topic 4 contains 2,308 records (2.94%) and includes intelligent, module, data, information, and smart. The vocabulary supports an interpretation centered on system integration and data-enabled applications. Its relatively small size should not be treated as evidence of low importance, because cluster size reflects corpus composition and model granularity rather than technological quality or economic value.

4.2. Topic Distinctiveness and Pairwise Relatedness

Figure 1 plots the c-TF-IDF scores of the ten leading terms in each topic. Topic 4 has the most distinctive first-ranked term, followed by Topics 3 and 5. Topics 0 and 1 rely more heavily on generic engineering terms such as connection, configuration, installation, and mechanism, producing lower individual c-TF-IDF scores despite much larger cluster sizes. Topic 2 displays a more gradual decline across its optical and imaging terms. These scores measure term distinctiveness within the fitted topic model; they do not measure patent quality, technological importance, or market value.

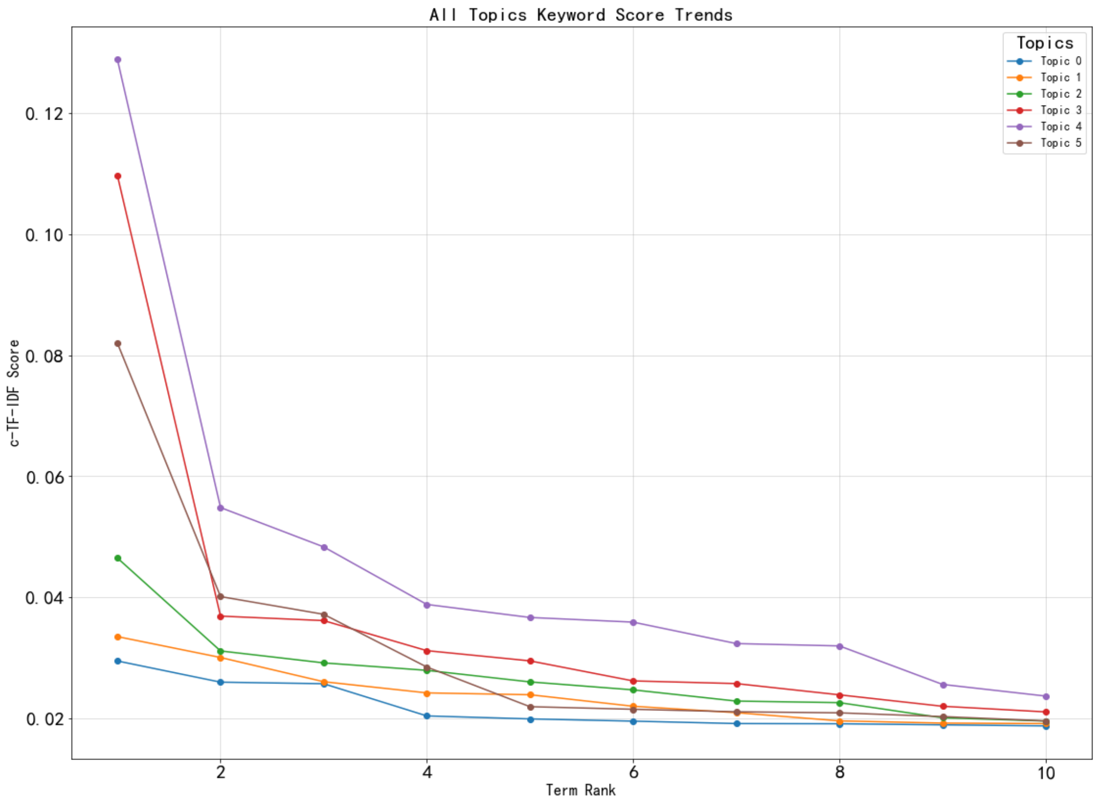

Figure 1. c-TF-IDF score trends for the six patent topics

The cosine-similarity matrix in Figure 2 indicates a highly connected semantic system. The strongest reported relationship is between communication networks (Topic 0) and device structure (Topic 1), with a similarity of 0.96. Topic 0 is also similar to optical sensing (0.88), intelligent systems (0.86), and graphite materials (0.82). Topic 1 is similar to intelligent systems (0.84), optical sensing (0.81), and graphite materials (0.86). These values indicate overlap in the selected topic-vector representation; they should not be interpreted as direct evidence of collaboration, complementarity, or technology transfer.

The optical-sensing topic shows moderate-to-high similarity with communication and device topics and lower similarity with the material topics. Topic 3 is the least connected overall in the displayed matrix, with similarities ranging from 0.55 with Topic 5 to 0.70 with Topic 0. The high similarities between graphite materials and the device (0.86) and communication (0.82) topics may reflect shared hardware terminology or genuine component-level relationships. Distinguishing these explanations requires inspection of representative patents and classification profiles.

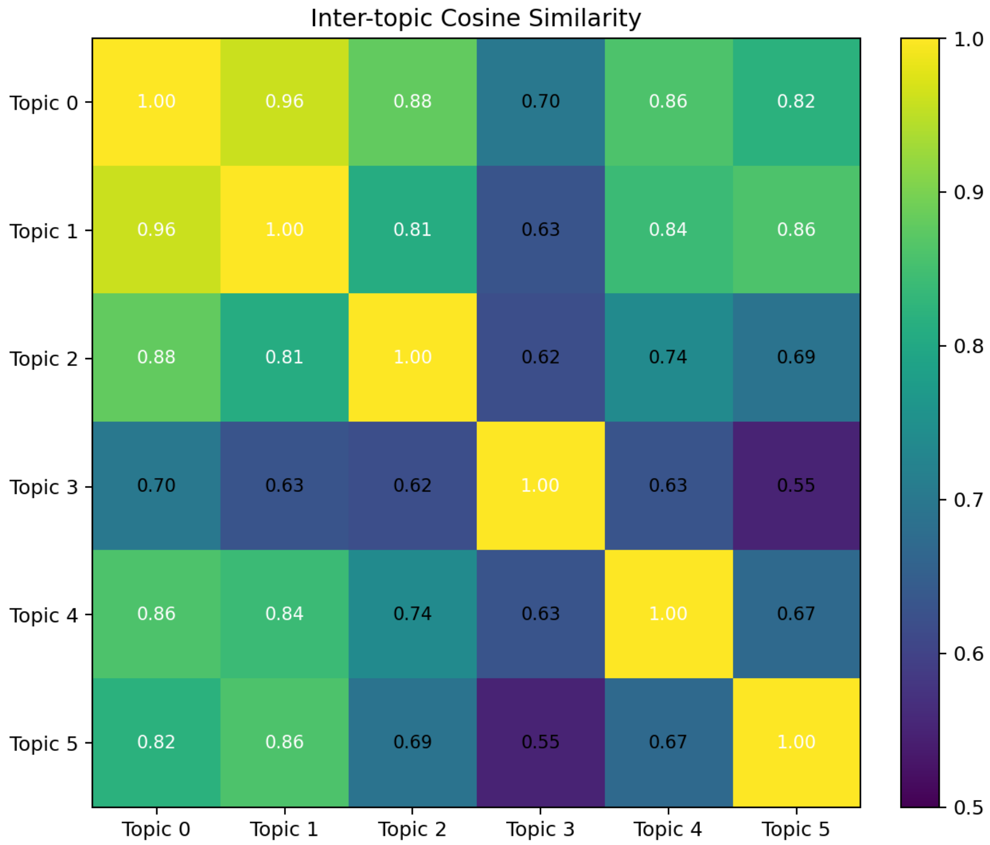

Figure 2. Cosine similarity among the six patent topics

4.3. Topic-space Distribution and Hierarchical Structure

The UMAP visualization in Figure 3 provides an exploratory document-level view of the fitted semantic space. Topic 0 occupies the dense central region, Topic 1 forms a compact cluster on the left, Topic 2 forms a separate optical-sensing cluster, Topic 4 is concentrated above the center, and Topics 3 and 5 appear toward the right. This pattern is consistent with a broad infrastructure cluster and more specialized peripheral groups. However, UMAP prioritizes local neighborhood preservation; apparent global distances, cluster sizes, and overlaps in the two-dimensional plot should not be treated as formal measures of relatedness.

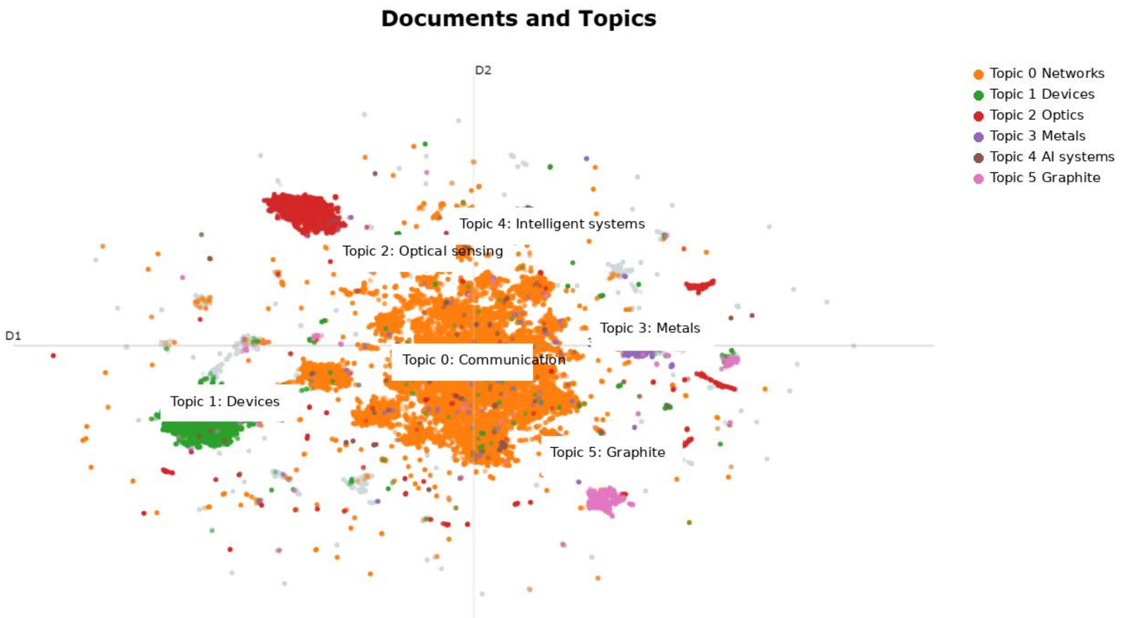

Figure 3. UMAP distribution of patent documents and topics

The reported dendrogram in Figure 4 places Topics 0 and 1 in the closest branch, with Topic 2 joining next. Topics 3 and 5 form a separate materials branch, while Topic 4 joins at the largest reported linkage distance. This global profile is compatible with a communication-device-sensing branch, a materials branch, and a differentiated intelligent-systems topic. Because the distance metric, linkage rule, and topic-vector representation are not documented, the exact linkage distances should be treated as descriptive rather than fully reproducible statistics.

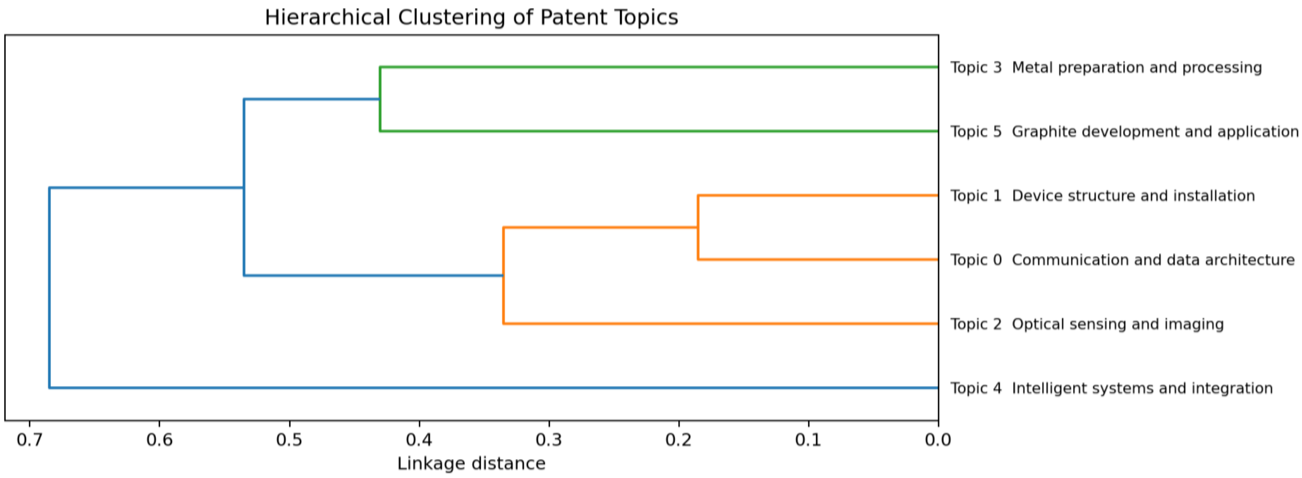

Figure 4. Hierarchical clustering of the six patent topics

5. Discussion

5.1. Main Findings and Contribution

The analysis reveals a layered regional patent portfolio. Communication and data architecture form the largest semantic foundation, device structure provides a hardware-oriented interface, and optical sensing supplies perception and measurement functions. Intelligent systems form a differentiated integration topic, while metal and graphite technologies constitute a materials branch. This interpretation is consistent with the systems perspective of smart manufacturing, which emphasizes the interdependence of connectivity, physical equipment, sensing, data, and intelligent decision making [3-7]. It remains an interpretation of patent-text structure rather than direct evidence of an operating regional innovation system.

The strongest descriptive result is the 0.96 similarity between communication networks and device technologies, together with their close placement in the dendrogram. The shared vocabulary suggests that regional information-technology inventions are frequently expressed through equipment, connection, installation, and configuration concepts. Optical sensing extends this branch toward measurement and image-related functions. The result supports a functional reading from infrastructure to equipment and perception, but it does not by itself demonstrate coordinated R&D or industrial integration.

The materials branch adds a second dimension to the portfolio. Material technologies are less similar to most information-system topics, yet they may be relevant to components, sensors, terminals, thermal management, and communication equipment. This interpretation is compatible with regional diversification research, which argues that new growth paths often emerge from related capabilities while also benefiting from complementary knowledge domains [29, 30]. Smart-specialization research likewise emphasizes the balance between relatedness, knowledge complexity, and strategic diversification [31]. At the same time, the materials topics may signal field-boundary leakage, so their substantive role cannot be confirmed without the search query and representative records.

Topic 4 is particularly informative. Pairwise similarities indicate that intelligent systems share semantic content with communication and device capabilities, whereas the reported dendrogram places Topic 4 at the greatest global distance. One plausible interpretation is that intelligent-system patents reuse infrastructure-related terms while forming a differentiated application and integration profile. Such recombination can be a source of novelty and coordination requirements [32], but the conclusion remains sensitive to the unspecified vector and linkage choices.

Methodologically, the study shows the value of connecting transformer-based topic modeling with multiple views of relatedness. BERTopic provides interpretable clusters, c-TF-IDF identifies discriminative vocabulary, cosine similarity summarizes pairwise proximity, UMAP offers an exploratory document map, and hierarchical clustering describes a global branching structure. The principal contribution is therefore an integrated analytical workflow rather than a new causal theory. Its publishability depends on complete reporting of data construction, model parameters, topic validation, and robustness.

5.2. Policy Implications

Strengthen the communication-intelligence interface. The high semantic relatedness between communication, devices, and intelligent systems suggests that regional programs may benefit from jointly supporting industrial communication, interoperable data platforms, edge-cloud architectures, equipment integration, and intelligent control. Policy instruments should require demonstrable cross-layer integration rather than funding isolated technology labels.

Build demonstration chains linking devices, sensing, and intelligent applications. The device and optical-sensing topics form a plausible bridge between infrastructure and application-oriented intelligence. Demonstration projects in industrial vision, intelligent inspection, predictive maintenance, and connected equipment could test whether this semantic relatedness translates into operational integration.

Connect materials research with component and terminal development. The materials branch should not be evaluated only by direct similarity to software-intensive topics. Targeted programs can link advanced materials with sensors, communication components, thermal management, energy storage, and intelligent terminals. Because the present evidence is text-based, such programs should be guided by additional firm, project, and commercialization data.

Promote differentiated but coordinated regional specialization. Regional policy should avoid duplicating the same complete technology chain in every locality. A more efficient approach is to support specialized nodes in communication hardware, intelligent systems, optical sensing, and materials while building shared testing platforms, technology-transfer mechanisms, and cross-city commercialization channels. This approach is consistent with the view that innovation depends on cognitive, organizational, and institutional proximity in addition to geography [18].

Use semantic patent analytics as a policy-feedback tool. Repeating the analysis with a stable search strategy and documented parameters could reveal changes in topic prevalence and relatedness. Such evidence can complement patent counts and citations. Citation-based indicators are informative but reflect technological value, examination practices, field-specific behavior, and exposure time [33, 34], while text-based indicators offer an additional view of novelty and technological change [35]. Neither type of indicator should be used as a stand-alone measure of policy success.

5.3. Limitations and Robustness Requirements

Several limitations should be acknowledged. First, the study relies on a single patent database and public patent texts, which may not capture unpublished firm knowledge or all forms of innovation. Second, the threshold of at least ten forward citations may favor older patents because citation opportunities vary across filing years. Third, the analysis pools records from 2014-2024 and therefore identifies portfolio structure rather than annual topic evolution. Fourth, topic boundaries and labels remain sensitive to the embedding model, preprocessing choices, dimensionality reduction, and density-based clustering. Finally, semantic similarity indicates textual relatedness and cannot by itself establish collaboration, knowledge spillovers, complementarity, or causal effects on manufacturing performance.

Future research can compare alternative sentence encoders, preprocessing configurations, clustering parameters, and random seeds; report topic stability and expert agreement; and evaluate patent-specific language models and technology-entity extraction methods [36, 37]. A temporal extension could estimate annual topic prevalence and align topics across time windows. Linking patent topics with assignee networks, firm outcomes, commercialization records, project data, and policy texts would allow the study to move from semantic structure toward validated mechanisms and performance effects.

6. Conclusions

This study develops a semantic framework for mapping next-generation information technology patents in the Chengdu-Chongqing Economic Circle. The dataset contains 90,232 citation-screened records from 2014-2024, of which 78,511 were assigned to six substantive topics and 11,721 to the Topic -1 outlier/noise class. Communication networks and data architecture dominate the classified set; device structure and optical sensing form hardware- and perception-oriented branches; intelligent systems form a differentiated integration topic; and metal and graphite technologies constitute a materials branch.

The results combine strong pairwise relationships with differentiated global profiles. Communication and device topics show the highest reported cosine similarity (0.96), optical sensing is positioned near the infrastructure-device branch, and intelligent systems share substantial pairwise similarity while remaining globally distinct in the reported dendrogram. Material topics are more peripheral and require additional field-validation evidence. These findings support a layered descriptive interpretation, but not causal claims about regional collaboration or industrial upgrading.

By integrating Sentence-BERT, BERTopic, UMAP, HDBSCAN, c-TF-IDF, cosine similarity, and hierarchical clustering, the study advances regional patent analysis from topic counts toward structural interpretation. The framework provides a structured basis for technology landscaping and evidence-informed regional specialization and can be extended through temporal modeling, classification-code validation, assignee networks, and industry-performance data.

References

[1] Vial, G. (2019). Understanding digital transformation: A review and a research agenda. The Journal of Strategic Information Systems, 28(2), 118–144. https://doi.org/10.1016/j.jsis.2019.01.003

[2] Verhoef, P. C., Broekhuizen, T., Bart, Y., Bhattacharya, A., Dong, J. Q., Faber, N., & Hoekstra, J. C. (2021). Digital transformation: A multidisciplinary reflection and research agenda. Journal of Business Research, 122, 889–901. https://doi.org/10.1016/j.jbusres.2019.09.022

[3] Lu, Y. (2017). Industry 4.0: A survey on technologies, applications and open research issues. Journal of Industrial Information Integration, 6, 1–10. https://doi.org/10.1016/j.jii.2017.04.005

[4] Xu, L. D., Xu, E. L., & Li, L. (2018). Industry 4.0: State of the art and future trends. International Journal of Production Research, 56(8), 2941–2962. https://doi.org/10.1080/00207543.2018.1444806

[5] Zhong, R. Y., Xu, X., Klotz, E., & Newman, S. T. (2017). Intelligent manufacturing in the context of Industry 4.0: A review. Engineering, 3(5), 616–630. https://doi.org/10.1016/j.eng.2017.05.015

[6] Frank, A. G., Dalenogare, L. S., & Ayala, N. F. (2019). Industry 4.0 technologies: Implementation patterns in manufacturing companies. International Journal of Production Economics, 210, 15–26. https://doi.org/10.1016/j.ijpe.2019.01.004

[7] Kusiak, A. (2018). Smart manufacturing. International Journal of Production Research, 56(1-2), 508–517. https://doi.org/10.1080/00207543.2017.1351644

[8] The State Council of the People's Republic of China. (2021). China issues master plan for Chengdu-Chongqing economic circle. https://english.www.gov.cn/policies/latestreleases/202110/20/content_WS6170174dc6d0df57f98e392f.html

[9] Ernst, H. (2003). Patent information for strategic technology management. World Patent Information, 25(3), 233–242. https://doi.org/10.1016/S0172-2190(03)00077-2

[10] Daim, T. U., Rueda, G., Martin, H., & Gerdsri, P. (2006). Forecasting emerging technologies: Use of bibliometrics and patent analysis. Technological Forecasting and Social Change, 73(8), 981–1012. https://doi.org/10.1016/j.techfore.2006.04.004

[11] Abbas, A., Zhang, L., & Khan, S. U. (2014). A literature review on the state-of-the-art in patent analysis. World Patent Information, 37, 3–13. https://doi.org/10.1016/j.wpi.2013.12.006

[12] Tseng, Y. H., Lin, C. J., & Lin, Y. I. (2007). Text mining techniques for patent analysis. Information Processing and Management, 43(5), 1216–1247. https://doi.org/10.1016/j.ipm.2006.11.011

[13] Yoon, B., & Park, Y. (2004). A text-mining-based patent network: Analytical tool for high-technology trend. The

Journal of High Technology Management Research, 15(1), 37–50. https://doi.org/10.1016/j.hitech.2003.09.003

[14] Arts, S., Cassiman, B., & Gomez, J. C. (2018). Text matching to measure patent similarity. Strategic Management Journal, 39(1), 62–84. https://doi.org/10.1002/smj.2699

[15] Hain, D. S., Jurowetzki, R., Buchmann, T., & Wolf, S. (2022). A text-embedding-based approach to measuring patent-to-patent technological similarity. Technological Forecasting and Social Change, 177, 121559. https://doi.org/10.1016/j.techfore.2022.121559

[16] Krestel, R., Chikkamath, R., Hewel, C., & Rauterberg, J. (2021). A survey on deep learning for patent analysis. World Patent Information, 65, 102035. https://doi.org/10.1016/j.wpi.2021.102035

[17] Lee, J. S., & Hsiang, J. (2020). Patent classification by fine-tuning BERT language model. World Patent Information, 61, 101965. https://doi.org/10.1016/j.wpi.2020.101965

[18] Boschma, R. (2005). Proximity and innovation: A critical assessment. Regional Studies, 39(1), 61–74. https://doi.org/10.1080/0034340052000320887

[19] Jaffe, A. B., Trajtenberg, M., & Henderson, R. (1993). Geographic localization of knowledge spillovers as evidenced by patent citations. The Quarterly Journal of Economics, 108(3), 577–598. https://doi.org/10.2307/2118401

[20] Kogler, D. F., Rigby, D. L., & Tucker, I. (2013). Mapping knowledge space and technological relatedness in US cities. European Planning Studies, 21(9), 1374–1391. https://doi.org/10.1080/09654313.2012.755832

[21] Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet Allocation. Journal of Machine Learning Research, 3, 993–1022.

[22] Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. Proceedings of NAACL-HLT 2019, 4171–4186. https://doi.org/10.18653/v1/N19-1423

[23] Reimers, N., & Gurevych, I. (2019). Sentence-BERT: Sentence embeddings using Siamese BERT-networks. Proceedings of EMNLP-IJCNLP 2019, 3982–3992. https://doi.org/10.18653/v1/D19-1410

[24] Grootendorst, M. (2022). BERTopic: Neural topic modeling with a class-based TF-IDF procedure. arXiv preprint arXiv:2203.05794. https://doi.org/10.48550/arXiv.2203.05794

[25] McInnes, L., Healy, J., Saul, N., & Großberger, L. (2018). UMAP: Uniform Manifold Approximation and Projection. Journal of Open Source Software, 3(29), 861. https://doi.org/10.21105/joss.00861

[26] Campello, R. J. G. B., Moulavi, D., Zimek, A., & Sander, J. (2015). Hierarchical density estimates for data clustering, visualization, and outlier detection. ACM Transactions on Knowledge Discovery from Data, 10(1), 1–51. https://doi.org/10.1145/2733381

[27] Salton, G., & Buckley, C. (1988). Term-weighting approaches in automatic text retrieval. Information Processing and Management, 24(5), 513–523. https://doi.org/10.1016/0306-4573(88)90021-0

[28] Röder, M., Both, A., & Hinneburg, A. (2015). Exploring the space of topic coherence measures. Proceedings of the Eighth ACM International Conference on Web Search and Data Mining, 399–408. https://doi.org/10.1145/2684822.2685324

[29] Tödtling, F., & Trippl, M. (2005). One size fits all? Towards a differentiated regional innovation policy approach. Research Policy, 34(8), 1203–1219. https://doi.org/10.1016/j.respol.2005.01.018

[30] Neffke, F., Henning, M., & Boschma, R. (2011). How do regions diversify over time? Industry relatedness and the development of new growth paths in regions. Economic Geography, 87(3), 237–265. https://doi.org/10.1111/j.1944-8287.2011.01121.x

[31] Balland, P. A., Boschma, R., Crespo, J., & Rigby, D. L. (2019). Smart specialization policy in the European Union: Relatedness, knowledge complexity and regional diversification. Regional Studies, 53(9), 1252–1268. https://doi.org/10.1080/00343404.2018.1437900

[32] Fleming, L. (2001). Recombinant uncertainty in technological search. Management Science, 47(1), 117–132. https://doi.org/10.1287/mnsc.47.1.117.10671

[33] Hall, B. H., Jaffe, A., & Trajtenberg, M. (2005). Market value and patent citations. The RAND Journal of Economics, 36(1), 16–38.

[34] Jaffe, A. B., & de Rassenfosse, G. (2017). Patent citation data in social science research: Overview and best practices. Journal of the Association for Information Science and Technology, 68(6), 1360–1374. https://doi.org/10.1002/asi.23731

[35] Kelly, B., Papanikolaou, D., Seru, A., & Taddy, M. (2021). Measuring technological innovation over the long run. American Economic Review: Insights, 3(3), 303–320. https://doi.org/10.1257/aeri.20190499

[36] Bekamiri, H., Hain, D. S., & Jurowetzki, R. (2024). PatentSBERTa: A deep NLP based hybrid model for patent distance and classification using augmented SBERT. Technological Forecasting and Social Change, 206, 123536. https://doi.org/10.1016/j.techfore.2024.123536

[37] Puccetti, G., Giordano, V., Spada, I., & Tornincasa, L. (2023). Technology identification from patent texts: A novel named entity recognition method. Technological Forecasting and Social Change, 186, 122160. https://doi.org/10.1016/j.techfore.2022.122160