Discovering Relationships Between Patent Literature and Academic Papers from a Patent Entity Perspective: A Case Study on the "Data Mining" Theme (Postprint)

Authors: Ning Zichen, Wei Lai

Date: 2023-04-01T16:15:55+00:00

Abstract

[Purpose/Significance] Patent literature and academic papers respectively represent new developments in technological innovation and scientific research. Combining the two through patent entities for technology theme evolution analysis holds certain reference significance for further discovering the relationship between patent technology and scientific research. [Method/Process] Using academic inventors in the data mining field as a link, this study proposes correlation methods and constructs a research framework from three perspectives: patent entity-keyword coupling, IPC coupling, and IPC-keyword co-occurrence. It analyzes the evolution of multi-dimensional correlation relationships among entities, technologies, and themes across different time periods, and explores the entity and theme correlation relationships between patent literature and academic papers in the data mining domain. [Results/Conclusion] Academic inventors are playing an increasingly important role in data mining technology innovation. The technical themes of most entities are similar, with some even demonstrating a high degree of unity. However, there also exist a few cases where technology and themes are not directly related, showing significant differences. Nevertheless, regardless of whether technology and themes are directly related, data mining-related technological inventions and scientific research have already achieved relatively deep mutual penetration.

Full Text

Abstract

[**Purpose/Significance**] Patent documents and academic papers respectively represent new developments in technological innovation and scientific research.

Combining the two through patent subjects for technical theme evolution analysis offers valuable insights into discovering relationships between patented technologies and scientific research. [**Method/Process**] Taking academic inventors in the data mining field as the link, this study proposes an association method and constructs a research framework from three perspectives: patent subject-keyword coupling, IPC coupling, and IPC-keyword co-occurrence. It analyzes the evolution of multi-dimensional relationships among subjects, technologies, and themes across different time periods to explore the subject and thematic connections between patent documents and academic papers in data mining. [**Result/Conclusion**] Academic inventors play an increasingly important role in data mining technology innovation. Most subjects exhibit similar technical themes, with some showing a high degree of unity. However, a minority of technologies are not directly related to themes and show significant differences. Regardless of direct relevance, data mining-related technological inventions and scientific research have achieved deep mutual penetration.

Keywords: patent subject; patent documents; academic papers; association discovery

1 Introduction

A new round of scientific and technological revolution and industrial transformation is emerging, with global scientific and technological innovation presenting new development trends and characteristics. Scientific and technological innovation serves as a direct driver of development. However, in China's current scientific and technological development practice, science and technology have not yet demonstrated a favorable interactive trend. This is mainly manifested in the inability of some scientific research achievements to be promptly applied to technical practice, while many technical problems cannot be effectively solved due to the lack of new scientific achievements, thereby limiting scientific progress and social development to a certain extent. In this context, clarifying the relationship between science and technology is particularly important. Scholars have successively explored the relationship between science and technology and related research methods, aiming to promote mutual penetration and interaction and accelerate the integration of science and technology. Currently, scientific research achievements are primarily produced in the form of academic papers, while patent information serves as the largest available source of technical information, carrying the core content of technological innovation. Therefore, the connection between science and technology is mainly reflected in the association discovery between patent documents and academic papers. The library and information science field has conducted continuous research on the relationship between patent documents and academic papers. Early studies mostly explored the relationship between science and technology from the perspective of inter-subject relationships (between inventors and authors), but research after 2010 has focused primarily on content-based approaches to more deeply investigate the relationship between science and technology.

In recent years, foreign scholars in library and information science have concentrated on exploring co-citation relationships between patent documents and academic papers, manifested in three aspects: (1) analyzing citation patterns by examining how academic papers are cited in patents or how patents are cited in academic papers; (2) investigating how mutual citations facilitate knowledge flow from academia to technology; and (3) constructing citation networks through mutual citation relationships and conducting topological analysis to explore the role of technology and knowledge in domain development. Research from the citation perspective is abundant and has yielded certain results. Additionally, some scholars have explored domain theme evolution through technical theme analysis and proposed recommendation systems to improve recall rates for simple word retrieval, providing research references from a lexical perspective. Moreover, the discovery of relationships between patents and scientific literature is not limited to theoretical research; scholars have also conducted applied research. Through studies of multiple relationships between patent documents and academic papers, they have discovered technological development in specific thematic areas, explored regional economic growth, and conducted comprehensive evaluations of national and institutional research productivity inputs and outputs.

Domestic research on patent documents and academic papers started relatively later than abroad, with co-citation relationships initially forming the research foundation. In recent years, with the development of ontologies and semantic webs, scholars have approached the relationship between patents and academic papers from a lexical perspective using scientometric methods, such as thematic association evolution and similarity calculation methods, providing new research angles. In terms of applications, besides exploring technological and research development in specific fields, frontier hotspot research has also considered the technical contributions of patent documents.

Chinese patent documents lack citation information, and the classification systems for patents and academic papers differ, making direct citation analysis or classification number association analysis unsuitable. This paper proposes approaching the issue from the perspective of patent subjects. Patentees are the owners of patents, while patent inventors, as technology developers, may also be academic researchers. Based on these two types of subjects, relationships between innovative patents and scientific research achievements can be established. By analyzing patent subject-literature keyword coupling and patent thematic coupling relationships, networks of patent themes and literature keywords can be constructed. Using the “data mining” domain as an example, this study analyzes and explores the evolution patterns of patent subjects-technical themes-literature keywords in the data mining theme, hoping to provide reference for association discovery between patent documents and academic papers.

2 Concept Definition and Research Status of Patent Subject-Keyword Coupling

2.1 Concept Definition

2.1.1 Patent Subject Concept and Definition Patent subjects are entities related to the patent lifecycle, including formation, application, and utilization. In reality, a patented technology generally involves multiple subjects. For instance, inventors are the core part of patent creation, directly participating in and completing the invention with creative contributions; applicants refer to natural persons, legal persons, or other organizations who legally enjoy an invention and file a patent application with the State Council's patent administration department; and patentees are the owners of patent rights. Additionally, patent subjects include patent assignees, patent agents, and other entities. However, for a patent, patentees and inventors are essential. When the inventor is also the technology owner, the inventor is the applicant and patentee. For institutions, the patentee is generally the institution itself, with relevant development teams serving as inventors.

Academic inventors are special patent inventors. In the article by Wang Gangbo and Guan Jiancheng, "academic inventor" is translated from the term "Academic Inventor" in literature, referring to individuals in universities who engage in both academic research and patent activities, possessing both author and inventor identities. This paper defines academic inventors as patent inventors who have academic publications.

2.1.2 Patent Subject-Keyword Coupling Concept and Definition In library and information science research, scholars generally believe that coupling research was first proposed by American scholar Kessler. As mentioned in the article "On Bibliographic Coupling and Co-citation": Dr. M.M. Kessler first proposed the concept of bibliographic coupling in 1963. Kessler discovered the citation coupling law—if Document A and Document B both cite Document C, then A and B have a coupling relationship and share similarities. This coupling can also be applied to patent subject-keyword coupling.

Patent subject-keyword coupling refers to the coupling between patentees and keywords. Here, keywords are not subject terms from patent documents but rather keywords from scientific literature. Using inventors listed in patent documents as intermediaries, the study retrieves academic papers published by these inventors in CNKI and Wanfang databases and records the paper keywords, thereby forming a coupling network between patentees and keywords. Through patent inventors, relationships between patent subjects (patentees) and keywords are established to identify scientific literature highly relevant to the patent and provide corresponding keywords for analysis.

2.2 Feasibility Analysis

2.2.1 Chinese Patents Lack Citation Information, Creating Blind Spots for Citation-Based Research Although domestic and foreign research methods for association discovery between patent documents and academic papers have focused on citation networks, Chinese patent databases basically contain no cited patents, with only a very small number of citations added by patent examiners in recent years. Additionally, patent documents and scientific literature follow different style norms and content emphases, making scholars rarely cite patent documents in academic writing, which creates certain blind spots in patent citation research.

2.2.2 Themes as Knowledge Units Can Reflect Subjects' Academic Backgrounds In scientific knowledge representation, themes serve as meso-level knowledge units that represent authors' specific research fields or disciplinary backgrounds from a content perspective. In science and technology knowledge networks, although thematic units are implicit scientific molecules representing knowledge content, they are closely associated with other knowledge units and influence the course and direction of scientific development. While themes can be revealed from a classification perspective, the classification systems for patent documents and academic papers differ significantly, making it impossible to comprehensively and reasonably establish mapping relationships. Therefore, it is necessary to start from a lexical perspective, associating similar nodes in terms of content to discover relationships between patent documents and academic papers.

2.2.3 Patent Subjects Serve as Bridges for Building Relationships Between Patent Documents and Academic Papers Patents are classified according to their application fields, while scientific literature is classified according to the Chinese Library Classification system, creating significant differences that prevent direct mapping through subject terms. As carriers of technology and knowledge, subjects play important roles in promoting knowledge flow between science and technology. Inventors participating in patent creation may be academic inventors. By tracking these inventors' academic papers and establishing connections with patent documents, knowledge flow in technological development can be better reflected. Therefore, based on patent subjects, exploring the association and evolution patterns between patent documents and academic papers from a technology-theme perspective can more easily achieve research objectives from the angles of patent subjects' academic backgrounds and technological innovation.

Thus, based on patent subjects (primarily academic inventors), this article takes patent IPCs and academic paper keywords as research objects to explore relationships between patent subjects and literature themes, between patent subjects and technologies, and between technologies and literature themes, thereby discovering patterns of technical theme evolution.

3 Association Methods and Framework Construction

Patent subjects participate in both patent invention and academic creation, so the themes of their corresponding patent documents and academic papers are relatively similar. Based on patent documents and academic papers and using patent subjects as intermediaries, this study explores the association between patent documents and academic papers in the data mining field from three perspectives: patent subject-keyword coupling, technology evolution, and technology-theme association. The overall research framework is shown in Figure 1 [Figure 1: see original paper]. A patented technology corresponds to one or more patentees, which correspond to multiple inventors. Some inventors simultaneously engage in academic creation and publish academic papers. Therefore, patent documents and academic papers can establish connections through patent subjects, forming a "patent document-patentee-patent inventor-academic paper" relationship. This research first retrieves patent documents for a given technical theme and records detailed data. It then forms a corresponding technology co-occurrence network through patent information. Simultaneously, it retrieves relevant academic papers through patent subjects, records paper keywords, forms a patent subject-keyword coupling network, and finally constructs a relationship network between patent technology themes and literature keywords to discover relationships between patent documents and academic papers.

3.1 Patent Subject-Keyword Coupling Network

The association between patent subjects and literature keywords uses academic inventors as the link, pointing to both the inventors' patent documents and published academic papers. The study retrieves academic papers published by patent inventors within two years before the patent application date in CNKI and Wanfang databases, records the patentees and keywords from these papers, and constructs a coupling network between patent subjects and literature keywords, as shown in Figure 2 [Figure 2: see original paper]. Since invention patents are generally published 18 months after the application date, this paper selects academic papers published by inventors within two years before the patent application date and records their keywords, which can reflect inventors' technical backgrounds or main technical directions during the patent invention period.

A patent record corresponds to Patentee 1 and Patentee 2. Patentee 1 corresponds to Inventor 11, Inventor 12, and Inventor 13, while Patentee 2 corresponds to Inventor 21 and Inventor 22. Different inventors correspond to different literature keywords a-h, ultimately forming a coupling relationship between patentees and keywords for a patent.

3.2 Technology Co-occurrence Network

Technology coupling is primarily represented through the coupling of patent IPC classification numbers. Patents are classified according to their application fields, including eight main sections with a hierarchical system of section-class-subclass-group-subgroup, as shown in Figure 3 [Figure 3: see original paper] (specific IPC meanings can be found on the National Intellectual Property Administration's China Patent Publication Gazette website).

A patent may have multiple IPC classification numbers, involving multiple technical fields. Co-occurrence processing of patent IPC classification numbers can reflect core technical fields, peripheral technical fields, and relationships among various technical fields in a specific theme. PageRank is then used for node statistical processing, calculating edges to reflect the importance of technical fields represented by nodes in the network.

Gephi has modified the PageRank algorithm to reduce the impact of "selfish reciprocal relationship nodes" and "nodes with no in-degree or out-degree" on network relationships. This can reflect the importance of patent IPCs while presenting patent IPC coupling relationships, thereby analyzing technology associations from the patent perspective.

3.3 Patent Technology-Literature Keyword Relationship Network

Technology is primarily represented by patent IPCs, while keywords are from academic papers. The technology-theme network is essentially the result of IPC classification number-keyword co-occurrence. Since the number of inventors participating in a patent innovation varies, a patent record's technical background may involve multiple aspects. Additionally, these inventors' publication records during the corresponding period are uncertain, leading to potentially high dispersion of keywords corresponding to a patent. Therefore, it is necessary to first determine the scope of keywords before constructing the technology-theme network.

For keyword scope selection, Gephi's K-core algorithm is used to filter data from the patent subject-keyword coupling network, determining whether keywords are core or peripheral based on modularity. Core keywords can effectively cover the overall distribution of knowledge points in the research field and reflect a patent's main technical direction or discipline affiliation, making core keyword selection essential. Peripheral keywords, although having low connectivity, are diverse in content and can reflect implicit relationships between weakly related themes and technologies to some extent, so changes in such words also need consideration.

After keyword determination, a patent IPC-keyword co-occurrence network is constructed based on patent subjects. Core and peripheral keywords from the same time period are aggregated with the same patent IPC, with duplicate keywords undergoing frequency accumulation processing, as shown in Figure 4

[Figure 4: see original paper]. This ultimately yields a correspondence between all core/peripheral keywords and IPC classification numbers, forming an IPC classification number-keyword matrix. The visualization results undergo modularization processing, with low coupling between modules and high aggregation within modules, enabling exploration of relationships between technologies and themes.

4 Empirical Analysis of Patent Document and Academic Paper Association Based on Subjects

4.1 Data Selection

Big data has attracted attention in recent years due to the development of the internet and information industries. Effectively collecting, analyzing, preserving, and sharing data can provide effective help for scientific research and solve practical problems for users. This study takes "data mining" as the theme, retrieves data from the China Patent Office's "Patent Search and Analysis" database, uses Bradford's Law to determine 22 core patentees (institutions) and 248 relevant patent data records. Detailed patent data is crawled, including patent titles, application numbers, applicants, publication numbers, publication dates, IPC classification numbers, applicants, and inventors.

To facilitate analysis of science and technology evolution, the 248 patent data records are divided into three time periods: 2004-2008, 2009-2013, and 2014-2018, numbered as "1001, 1002…2001, 2002…3001, 3002, 3003…3150." The first digit represents the time period, where "1xxx" indicates patents from 2004-2008.

Academic papers published by inventors from these 248 patents are then retrieved. In CNKI and Wanfang databases, searches are conducted using the inventor's name as "author," the patentee's name as "institution," and the two years before the patent application date as "publication date." All bibliographic information is downloaded, merged, organized, and deduplicated, with the same numbering system applied.

Patent information and academic paper data are integrated through numbering and imported into Gephi 0.9.2 for individual and holistic analysis across the three time periods.

4.2 Results

4.2.1 IPC Co-occurrence Results Using BibExcel for data processing and Gephi's PageRank for visualization, IPC coupling for data mining-related patents is presented for 2004-2008, 2009-2013, and 2014-2018, as shown in Figures 5 [Figure 5: see original paper]-7 [Figure 7: see original paper]. Relevant network parameters are shown in Table 1 .

Since IPC classification numbers are coupled, "isolated points" in the network are removed, leaving only nodes with relationships. From network parame-

ters: (1) The number of nodes, edges, and average weighted degree increase across the three periods, consistent with technological development featuring multi-technology fusion for innovation. (2) Connected components refer to subgraphs where nodes are connected by edges but no relationships exist between subgraphs. The three periods show small but increasing connected component parameters, with node/connected component values also trending upward, indicating growing coupling networks and expanding subgraphs accommodating more technical nodes. (3) Graph density is very small across all three periods and decreases over time, suggesting mismatched growth between coupling network relationships and nodes. Although nodes and edges increase, the actual coupling network becomes increasingly sparse. Additionally, increasing average path length also indicates growing network sparsity.

In 2004-2008, IPCs established relationships through the central node G06F17/30, forming a subgroup with consistent coupling degrees among nodes. In 2009-2013, relationships became more complex, with two new subgroups emerging beyond the central node G06F17/30 subgroup. Regarding IPC coupling degrees, G06F17/30 showed strong coupling with H04L29/06(08), G06N3/12, and G06F17/50(27). After 2014, patent IPC coupling relationships became even more complex, with central nodes becoming G06F17/30 and G06Q50/06, and the number of subgroups and internal nodes increasing compared to 2009-2013. During this stage, G06Q50/06 had the highest coupling degree with G06Q10/06, followed by G06Q50/06 with G06F17/30 and G06Q50/06 with G06Q10/04. High node coupling degrees indicate that a patented technology simultaneously involves multiple fields represented by these nodes, with broad application scope. The increase in patent IPCs and network complexity shows that "data mining" is increasingly applied across multiple fields and that multi-domain technologies and knowledge are being utilized in data mining, representing disciplinary integration.

4.2.2 Patent Subject-Keyword Coupling Results Gephi is used to present patent subject-keyword coupling for data mining-related patents in 2004-2008, 2009-2013, and 2014-2018, as shown in Figures 8 [Figure 8: see original paper]-10 [Figure 10: see original paper]. Relevant network parameters are shown in Table 2 .

From network parameters: (1) The number of nodes and edges grows rapidly across the three periods, with average weighted degree also increasing, indicating that over time, patent subjects increase and patents involve more research themes trending toward integration. (2) With network module resolution at the default value of 1.0, the number of modules grows from the initial 3 to 18, with node label size representing node importance in the network and higher correlation within subgroups. (3) Graph density is very small across all three periods and decreases over time, indicating that although network nodes and edges increase, the actual coupling network becomes increasingly sparse.

In 2004-2008, the three core patent subjects were Tsinghua University, Shanghai

Jiao Tong University, and Zhejiang University, with few shared keywords among them: only "data fusion," "feature selection," and "data mining." In 2009-2013, the number of modules increased to 11, with core nodes adding State Grid, Chongqing University, and Nanjing University of Posts and Telecommunications, though Tsinghua University became less prominent in this stage. Inter-module relationships also became more complex, with more relationships and lower independence among subjects compared to Figure 8 [Figure 8: see original paper]. Shared keywords increased to include "grid computing," "support vector machine," "distributed computing," "monitoring system," "IEC61850," "integrated application server," "harmonic control," "smart substation," and "cognitive radio." After 2014, the patent subject-keyword coupling network became significantly more complex, with State Grid Corporation occupying the most central position. Additionally, various universities of posts and telecommunications gradually emerged as core nodes. The frequency of shared keywords increased, with high-frequency shared keywords including "big data," "neural network," "multi-objective optimization," "support vector machine," "association rule," "energy storage system," and "analytic hierarchy process." As the network becomes more complex, the growth of patent subjects reflects changes in institutions within the field, while keyword node growth reflects technological domain development. Keywords with high coupling degrees have greater importance in the network and play more significant roles in promoting technological development.

4.2.3 IPC-Keyword Co-occurrence Results Gephi is used to present IPC-core keyword networks and IPC-peripheral keyword networks, as shown in Figures 11 [Figure 11: see original paper]-13 [Figure 13: see original paper]. Node size represents importance in the network, while node grayscale indicates modularity.

IPC-core keyword co-occurrence results show: 2 modules formed in 2004-2008, 3 modules in 2009-2013, and 7 modules in 2014-2018. IPC-peripheral keyword co-occurrence results show: 2 modules in 2004-2008, 5 modules in 2009-2013, and 13 modules in 2014-2018. Peripheral co-occurrence networks are more dispersed than core co-occurrence networks, with fewer connections between subgroups, which is also reflected in the number of modules.

IPC and keyword distributions are as follows:

(1) IPC-Core Keyword Co-occurrence Network

In 2004-2008, IPCs were mainly G06Q50/00, G06Q10/00, and G06F17/30, corresponding to technical fields at the subclass level: Physics section-Computing/Calculating/Counting class-Electric digital data processing subclass and Data processing systems or methods subclass. Keywords mainly related to network development and service discovery. In 2009-2013, IPCs were mainly G06F17/30, H04L29/06, and G06F21/55, corresponding to Physics section-Computing/Calculating/Counting class-Electric digital data

processing subclass and Electricity section-Electric communication technology class-Transmission of digital information subclass. Keywords mainly involved network security and server pressure capacity control. In 2014-2018, IPCs were mainly G06F17/30, G06Q50/06, G05B23/02, and G01R31/12, with G01R being Physics section-Measuring/Testing class-Measuring electric or magnetic variables subclass. Keywords involved machine learning, deep learning, computer vision, and application-related technical vocabulary in various fields (such as finance, electrical engineering, smart cities, transportation, etc.).

(2) IPC-Peripheral Keyword Co-occurrence Network

In 2004-2008, IPCs were mainly G06F19/00 and G06F17/30, involving keywords mainly related to computer vision, application directions (power, energy), server pressure capacity control, and service discovery terminology. In 2009-2013, IPCs were mainly G06F19/00, G06F17/30, G05B13/04, and G05B19/418, with technical fields adding Physics section-Control/Regulation class-General control or regulation systems subclass. Keywords mainly related to testing, network security, and other essential technologies, as well as artificial intelligence, Web services, and application-specific terminology. In 2014-2018, IPCs were mainly G06F17/30, G06Q50/06, G06Q10/06, G06K9/62, and H04N19, newly adding Electricity section-Electric communication technology class-Pictorial communication subclass and Physics section-Computing/Calculating/Counting class-Data recognition subclass. Keywords were more application-oriented, such as transportation, mobile communication, power, energy, and information retrieval.

4.3 Association Analysis Between Patent Documents and Academic Papers Based on Subjects

4.3.1 Technology Evolution Analysis Based on the resulting networks, further analysis of association relationships is conducted.

Content-wise, in terms of patent classification number types, the three periods have 4, 15, and 44 types respectively, belonging to different sections. In 2004-2008, distribution was mainly concentrated in the Physics section under "Computing; Calculating; Counting." In 2009-2013, inventions added the Physics section's "Measuring; Testing" class and the Electricity section's "Electric communication technology" related patents. In 2014-2018, expansion reached the Physics section's "Signaling devices" and "Control; Regulation," the Electricity section's "Generation, conversion, or distribution of electric power," the Performing Operations/Transporting section's "Grinding; Polishing," and the Mechanical Engineering section's "Lighting; Heating; Weapons; Blasting - Storage or distribution of gas or liquid" related patented technology innovations. This shows that more and more application fields are emphasizing the utilization and innovation of data mining.

Regarding core nodes, in the two periods before 2014, G06F17/30 was at the core position as the most important node. After 2014, the core nodes added

G06Q50/06 based on G06F17/30. Although both belong to the G06 Physics section's "Computing; Calculating; Counting" class, the former mainly relates to "Information retrieval; database structures" in electric digital data processing, while the latter relates to "Systems or methods specially adapted for administrative, commercial, financial, managerial, supervisory or forecasting purposes" for electricity, gas, or water supply. G06F17/30 remains the mainstream technology for data mining, such as computer network technology and mathematical algorithms, while the emergence of G06Q50/06 as a core indicates that data mining-related technological innovation is no longer concentrated only in computer fields but also extensively involves power and water conservancy detection.

Technology coupling evolution: In 2009-2013, G06F17/30 showed strong coupling with H04L29/06(08), G06N3/12, and G06F17/50(27), meaning "Information retrieval and database structures" and "Communication control and processing" often appeared in the same patent. "Computer systems based on genetic models" and "Computer-aided design (natural language processing technology for dynamic analysis)" frequently co-occurred in patent inventions. In 2014-2018, G06Q50/06 formed stable coupling relationships with G06Q10/06 and G06F17/30, indicating high correlation among "Information retrieval and database structures," "Data processing systems or methods related to administrative management," and "Systems or methods suitable for electricity, gas, or water supply operation departments." The 2009-2013 technology coupling focused mainly on data mining-related technologies, while 2014-2018 emphasized the combination of technology and domain applications. During this process, the focus of technological invention and innovation shifted from optimizing and improving related technologies to specific applications within particular fields, achieving certain results.

4.3.2 One-Dimensional Relationship Discovery Based on Subjects

From the evolution shown in Figures 8-10, in 2004-2008, core subjects were Tsinghua University, Shanghai Jiao Tong University, and Zhejiang University. In 2009-2013, core subjects were Shanghai Jiao Tong University, Chongqing University, and Nanjing University of Posts and Telecommunications, with power company clusters appearing as important nodes. In 2014-2018, State Grid became the most central subject, occupying nearly half of Figure 7, with strongly connected modules dominated by power companies.

Since patent subject-keyword coupling focuses on academic innovative talents (technical personnel with dual identities), it ignores some inventors who only engage in technological innovation. Therefore, this core subject evolution cannot directly indicate that data mining technology first developed in universities. On the other hand, calculations on empirical data show that the proportion of academic inventors increased from 30% in 2004-2008 to 62.26% in 2009-2013 and 77.33% in 2014-2018, indicating that academic inventors account for an increasingly large proportion of technology inventors.

In the initial stage (2004-2008) of data mining-related patent applications, in-

novators who also focused on academic innovation were mainly university innovators, accounting for 30% of all subjects in this stage. In the preliminary development stage (2009-2013), enterprises (such as power companies) began to value the academic backgrounds of patent subjects and appeared as a subgroup in the coupling network. Inventors with academic publications accounted for 62.26% of subjects in this stage. In the rapid development stage of data mining-related technologies (2014-2018), enterprises with academic innovative talents could compete with universities, with inventors having published papers accounting for 77.33% of subjects. Throughout the development and innovation of data mining-related technologies, academic innovative talents grew from less than one-third to nearly four-fifths, showing that the association between data mining-related scientific research and technological invention is continuously strengthening. Researchers with scientific capabilities have had a significant impact on patent activities. Meanwhile, power industry enterprises emerged from nothing, reflecting the gradual formation of high knowledge capacity complementarity between industry and academia.

4.3.3 Two-Dimensional Relationship Discovery of Patent Subject-Keywords In keyword evolution, 2004-2008 mainly involved computer networks, computer vision, and server-related keywords. In 2009-2013, server optimization, network security, and analysis-related keywords were added based on the previous stage. In 2014-2018, the variety of core keywords greatly increased, focusing on data classification, algorithms, artificial intelligence (machine learning, deep learning, neural networks), and power field-related keywords (see Figures 8-10). Similar terms became more diverse and detailed in content.

In subject-keyword coupling, given keywords may be similar or related, forming co-keyword phenomena. Co-keywords among different subjects show certain biases: co-occurring keywords among universities tend to be theoretical research-oriented, while co-keywords between enterprises and universities, besides technical terms, tend to be application-oriented. Co-keywords among enterprises are mainly application-oriented. Co-keywords in different periods also have distinct characteristics: in 2004-2008, co-keywords were “data fusion,” “feature selection,” “urban planning,” and “spatial object integration”; in 2009-2013, co-keywords were “support vector machine,” “distributed computing,” “integrated monitoring system,” “IEC61850,” “integrated application server,” “harmonic control,” “smart substation,” and “cognitive radio”; in 2014-2018, co-keywords were “big data,” “neural network,” “multi-objective optimization,” “support vector machine,” “association rule,” “energy storage system,” and “analytic hierarchy process.” As time develops, more associations between themes and subjects emerge, and the similarity of themes valued by patent subjects also increases, indicating that these academic themes may be related to the core technologies of inventions during that stage. Meanwhile, besides enterprise application-related keywords, most other co-keywords are popular technical themes of the period, which are quickly noticed and applied by multiple subjects, showing that data

mining technology actively absorbs new scientific and technological knowledge and develops rapidly.

4.3.4 Technology-Keyword Network Relationship Evolution Analysis Based on Subjects Core themes in different stages show obvious differences. As shown in Figures 11-13, in 2004-2008, the technical terms involved were not the main data mining technologies but rather the same or similar development languages that could be used in the technical implementation process, providing support for data mining technology innovation. Service discovery-related themes may have been proposed to improve services and attract audiences, thereby generating demand for data mining. In 2009-2013, themes mainly focused on network security and server pressure capacity control, when data mining technology was basically formed and required in-depth consideration of optimization directions. In 2013-2018, themes were mostly application-oriented, with power system-related technical terms being the most numerous, while also involving specialized technical terms such as image recognition.

Peripheral themes in different stages show high similarity, all involving essential basic technical terms such as network security and server maintenance, as well as demand and application field-related terms. The difference lies in the fact that in 2004-2008, when data mining was just emerging, peripheral themes proposed artificial intelligence-related algorithms such as machine learning and deep learning, which became core themes after 2014. This shows that technological innovation and improvement require certain accumulation and preparation. Peripheral themes in 2014-2018 may similarly become core themes in a few years, emerging as core data mining technologies.

Due to these characteristics, selecting the IPC-core keyword co-occurrence network based on patent subjects can better reflect the evolution of the technology-theme association network and reveal the relationship between patent core technologies and disciplinary themes.

(1) 2004-2008: Data mining mainly involved technical fields of electric digital data processing, data processing systems or methods for administration and finance, and information retrieval and database structure technologies. These technologies were associated with network development and service discovery themes in scientific development—scientific themes that actually proposed some demand themes during the formation of data mining technology, matching with administration, management, finance, business, and other fields.

(2) 2009-2013: G06F17/30 and H04L29/06 were associated with network security themes, involving technical fields of information retrieval and database structure technology and communication control and processing technology. In network communications, data is transparent to developers, so emphasizing network security and preventing information leakage is crucial. G06F21/00, G06K9/62, and G06F21/55 were associated with server pressure capacity control themes, involving technical fields of security devices for protecting comput-

ers and their components, programs or data, and data recognition technology. Server pressure capacity includes maximum concurrent users, throughput, and disaster recovery. Although these scientific themes differ somewhat from technology, they all aim to further optimize data mining and propose optimization directions.

(3) 2014-2018: H04N19/xx was associated with image recognition themes, involving image communication technology. Inventors had disciplinary backgrounds in image recognition, enabling good integration and penetration between technological invention and scientific discovery, forming an important component of data mining. G06F, G06N, G06Q, G01R, and H04L were associated with application field themes such as transportation, education, and mobile communications. G06Q, H02J, G01R, G05B, and G06F were associated with power grid-related application field themes. G06Q, G06F, G06K, G06R, and G08B were associated with energy application-related themes. These technology-theme co-occurrences concentrated in various application fields, particularly prominent in the power and energy industry. Scientific knowledge and technological innovation jointly practiced in application fields, promoting knowledge transfer in science and technology. F17D5/00 was associated with physical material-related themes such as optical fibers, optical power effects, and equipment. This technology is protective or observation devices, both focusing on physical hardware protection and improvement, showing extremely high similarity between technology and theme. B24B51/00 was associated with G05B19/418 and G08G1/xx, relating to information retrieval (natural language processing, association mapping, etc.) and industrial production themes. The related technologies are machine tool devices or processes for grinding or polishing and traffic control systems for road vehicles. The similarity between this technology and theme is not high, but academic inventors with such scientific backgrounds conducted innovations in the performing operations/transporting field, showing that such scientific knowledge plays an important role in data mining innovation in the performing operations/transporting field.

During the 15 years of data mining, G06F17/30 (Physics section-Computing/Calculating/Counting class-Electric digital data processing subclass-Information retrieval and database structure group) has always been the core innovation field. The disciplinary backgrounds of related academic inventors change with the core themes of data mining. Some themes are directly related, while others are not directly related but all solve problems, promote technological innovation development, and facilitate deep mutual penetration between data mining-related technological inventions and scientific research. The integration of multidisciplinary and multi-domain scientific discoveries with technology will inevitably promote innovation and application of data mining or other technologies and drive development of related knowledge theories.

References

[1] Dong Kun, Xu Haiyun, Luo Rui, et al. Review on the analysis of the re-

lationship between science and technology[J]. Journal of the China Society for Scientific and Technical Information, 2018, 37(6): 642-652.

[2] Wang Gangbo, Guan Jiancheng. The link between nano science and technology: An analysis based on academic inventors[J]. China Soft Science, 2009(12): 71-79.

[3] KESSLER M M. Bibliographic coupling between scientific papers[J]. American documentation, 1963, 14(1): 10-25.

[4] BIKARD M. Made in academia: The effect of institutional origin on inventors' attention to science[J]. Organization science, 2018, 29(5): 818-836.

[5] AHMADPOOR M, JONES B F. The dual frontier: Patented inventions and prior scientific advance[J]. Science, 2017, 357(6351): 583-587.

[6] HUANG M H, YANG H W, CHEN D Z. Increasing science and technology linkage in fuel cells: A cross-citation analysis of papers and patents[J]. Journal of informetrics, 2015, 9(2): 237-249.

[7] DING C G, HUNG W C, LEE M C, et al. Exploring paper characteristics that facilitate the knowledge flow from science to technology[J]. Journal of informetrics, 2017, 1(1): 244-256.

[8] WANG X, ZHAO Y, LIU R, et al. Knowledge-transfer analysis based on co-citation clustering[J]. Scientometrics, 2013, 97(3): 859-869.

[9] LIU G. Visualization of patents and papers in terahertz technology: A comparative study[J]. Scientometrics, 2013, 94(3): 1037-1056.

[10] GAO J P, TENG L, PANG J. Hybrid document co-citation analysis: Making sense of the interaction between science and technology in technology diffusion[J]. Scientometrics, 2012, 93(2): 459-483.

[11] QI Y, ZHU N, ZHAI Y, et al. The mutually beneficial relationship of patents and scientific literature: Topic evolution in nano-science[J]. Scientometrics, 2018, 115(2): 893-911.

[12] RISCH J, KRESTEL R. What should I cite? Cross-collection reference recommendation of patents and papers[C]// KAMPS J, TSAKONAS G, MANOLOPOULOS Y, et al. 21st international conference on theory and practice of digital libraries. Cham: Springer, 2017: 40-46.

[13] GOLNABI H. Carbon nanotube research developments in terms of published papers and patents, synthesis and production[J]. Scientia iranica, 2012, 19(6): 2012-2022.

[14] WONG C Y, GOH K L. The sustainability of functionality development of science and technology: Papers and patents of emerging economies[J]. Journal of informetrics, 2012, 6(1): 55-65.

[15] PRATHAP G. Totalized input-output assessment of research productivity of nations using multi-dimensional input and output[J]. Scientometrics, 2018,

115(1): 577-583.

[16] Du Jian, Sun Yinan, Li Yongjie, et al. Identifying innovation frontiers from the intersection of science and technology[J]. Journal of the China Society for Scientific and Technical Information, 2018, 37(6): 642-652.

[17] Qin Jiahui, Ye Yeqi, Ye Ying. Citation time lag and cycle analysis of scientific papers and technology patents[J]. Information Studies: Theory & Application, 2019, 42(1): 94-99.

[18] Xu Hongjiao, Zeng Wen, Zhang Yunliang. Research on thematic association evolution method between papers and patents based on Word2vec[J]. Journal of Intelligence, 2018, 37(12): 36-42.

[19] Zeng Wen, Xu Hongjiao, Li Ying, et al. Research on similarity calculation methods for scientific journal literature and patent literature based on VSM[J]. Technology Intelligence Engineering, 2016, 2(3): 37-42.

[20] Dong Kun, Wu Hong. Analysis of research hotspots in 3D printing technology based on paper-patent integration[J]. Journal of Intelligence, 2014, 33(11): 73-76, 61.

[21] Luo Yunzhong, Chen Weijie, Xu Xiaolin. Patent intelligence analysis and utilization[M]. Shanghai: East China University of Science and Technology Press, 2007: 14.

[22] Wei Lai, Gao Feifei. Research on the cooperative relationship between patent inventors and applicants[J]. Journal of the China Society for Scientific and Technical Information, 2016, 35(5): 463-471.

[23] Wang Gangbo, Guan Jiancheng. The link between nano science and technology: An analysis based on academic inventors[J]. China Soft Science, 2009(12): 71-79.

[24] Qiu Junping. On bibliographic coupling and co-citation[J]. Library, 1987(3): 13-25.

[25] Ye Chunxia, Yu Xiang, Li Wei. Research on interdisciplinary knowledge networks of inter-firm patent cooperation[J]. Journal of Intelligence, 2013(4): 113-120.

[26] Wang Yuefen, Wang Jinshu, Guan Peng. Construction and evolution analysis of subject knowledge networks based on theme-theme association[J]. Information Science, 2018, 36(9): 9-15, 102.

[27] Liu Yong, Du Yi. Network data visualization and analysis tool: Gephi Chinese tutorial[M]. Beijing: Publishing House of Electronics Industry, 2017: 182-187.

[28] Wei Lai, Gao Xiran. Role positioning of university data librarians under the background of big data[J]. Information and Documentation Services, 2015(5): 90-94.

Author Contributions:

Ning Zichen: Conducted data investigation, proposed research ideas, and wrote the paper;

Wei Lai: Provided overall guidance on research ideas and content writing.

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