Research on the Knowledge Fusion Process in Emerging Fields: A Case Study of Bioinformatics (Postprint)

Authors: Zhou Yuan, Dong Fang, Liu Yufei

Date: 2023-07-26T00:00:00+00:00

Abstract

[Purpose/Significance] Technology convergence is the driving force behind the formation and development of emerging industries, while knowledge fusion is the prerequisite for technology convergence. Research on the knowledge fusion process is of great significance for guiding the formation and development of emerging industries.

[Method/Process] First, a theoretical model is constructed based on existing research to characterize the knowledge fusion process using paper citation networks. Second, verification methods are designed according to the characteristics of paper citation networks at each stage of the knowledge fusion process. Finally, an empirical analysis is conducted using the emerging bioinformatics field as a case study.

[Results/Conclusion] The empirical analysis results from the bioinformatics field demonstrate the effectiveness of the theoretical model and can provide a new approach for studying the knowledge fusion process.

Full Text

Preamble

Volume 63, Issue 8, April 2019

Research on the Process of Knowledge Fusion in Emerging Fields: A Case Study of Bioinformatics

Zhou Yuan[1], Dong Fang[1], Liu Yufei[1],[2]

[1] School of Public Policy and Management, Tsinghua University, Beijing 100084
[2] Center for Strategic Studies, Chinese Academy of Engineering, Beijing 100088

Abstract

[**Purpose/Significance**] Technology fusion is the driving force behind the formation and development of emerging industries, while knowledge fusion is the prerequisite for technology fusion. Studying the knowledge fusion process is of great significance for guiding the formation and development of emerging industries. [**Method/Process**] First, a theoretical model using paper citation networks to characterize the knowledge fusion process was constructed based on existing research. Second, verification methods were designed according to the characteristics of paper citation networks at each stage of the knowledge fusion process. Finally, an empirical analysis was conducted using the emerging field of bioinformatics as an example. [**Result/Conclusion**] The empirical analysis results from the bioinformatics field demonstrate the effectiveness of the theoretical model, providing a new method for studying the knowledge fusion process.

Keywords: knowledge fusion, knowledge flow, citation network, emerging field
Class Number: G254
DOI: 10.13266/j.issn.0252-3116.2019.08.015

Emerging technologies are a critical focus of China's innovation-driven development. Guiding the development of emerging technologies is of great significance for China to seize the high ground in future economy and technology and embark on a path of innovation-driven, sustainable development. The formation and development of emerging technologies follow multiple paths, among which technology fusion, as a new decisive factor in the emergence and development of emerging technologies, has attracted widespread attention and research. Scientific and objective analysis of the technology fusion process is important for guiding the formation and development of emerging technologies.

In F. Hacklin's evolutionary framework of technology fusion, knowledge fusion is the prerequisite for technology fusion and drives the formation and development of emerging technologies. During the development of knowledge fusion, knowledge spillovers different from the original scientific fields emerge. These spillovers often occur between existing scientific fields, and as they deepen, a new research paradigm distinct from those of the original fields gradually forms, signaling the birth of a new scientific domain. For example, the formation of the emerging field of bioinformatics began with biological science research primarily using experimental observation methods. As observational data increased, researchers started applying knowledge and methods from information science to biological science in the 1970s. With deepening knowledge spillovers between biological and information sciences, a new research paradigm emerged, distinct from both fields, indicating the birth of bioinformatics.

Existing research on knowledge fusion and technology fusion mostly focuses on predicting fusion emergence or conducting cross-sectional studies using indicators. Some studies employ qualitative methods, such as Zhao Hongzhou's mining model for scientific discovery to explain cross-fusion in technology evolu-

tion, and E. Leon's expert knowledge-based method for predicting fusion emergence. Other studies use quantitative data-driven approaches, such as R. Kong's similarity measures between fields to assess fusion degree, and H. Park's IPC classification-based indicators for identifying fusion status. Recently, scholars have begun combining clustering methods with patent citation networks. While effective at identifying fusion status from various perspectives, these studies cannot effectively describe the specific evolutionary process of knowledge fusion.

Papers serve as carriers and communication media for scientific knowledge, and the citation process among papers represents knowledge selection, evolution, dissemination, and application built upon previous knowledge. Citation relationships reflect knowledge flow processes. A paper citation network, with papers as nodes and citation relationships as links, can reflect complex knowledge flows at large scales and serves as an effective tool for studying knowledge flow during fusion. I. Sakata noted that citation networks are more effective than semantic analysis for describing scientific fields. According to Y. Kajikawa and M. Newman, topological clustering can divide citation networks into clusters based on paper aggregation, with each cluster corresponding to a scientific field. Combined with network visualization, this approach can intuitively display field distribution and evolution, making it effective for describing scientific fields and their changes. Therefore, this paper employs citation network visualization combined with topological clustering to analyze the entire knowledge fusion process.

This paper first reviews research on knowledge flow and field description to construct a theoretical model using citation networks to characterize knowledge fusion. It then designs verification methods and conducts empirical analysis in the emerging field of bioinformatics. Finally, the theoretical model is validated and refined based on empirical results. The proposed model provides a new method for studying knowledge fusion processes and can help researchers investigate the formation and development of emerging fusion fields.

2 Theoretical Model of Knowledge Fusion Process Based on Citation Networks

2.1 Overview of Knowledge Fusion Theory

N. Rosenberg first proposed the concept of technology fusion, which is not an isolated technical phenomenon but requires support from basic and applied sciences and is closely related to socio-economic and legal factors. The technology fusion process comprises four temporal stages: knowledge fusion, technology fusion, market fusion, and industrial fusion, with knowledge fusion as the prerequisite. In an ideal process, knowledge fusion leads to technology fusion, which drives product-market integration, market fusion, and ultimately industry consolidation.

Knowledge fusion has two forms. As shown in Figure 1 [Figure 1: see origi-

nal paper], the first involves two scientific fields (A and B) fusing to form a new field (C) while the original fields persist (A+B=A+B+C). The second involves two fields merging complementarily to form a new field that replaces the originals (AB) (A+B=AB). This paper's theoretical model focuses only on the A+B=A+B+C form.

2.2 Theoretical Model of Knowledge Fusion Process

The knowledge fusion process involves active knowledge flow and new field formation. Research indicates it begins with knowledge flow between two distinct scientific fields, deepening until a new research paradigm emerges, signifying fusion field formation. Specifically, it evolves from independent development of existing fields, to inter-field knowledge flow, and finally to new fusion field formation. This paper uses cross-field citations in citation networks to represent inter-field knowledge flow and cluster changes from topological clustering to represent field evolution, constructing a citation network-based theoretical model shown in Figure 2 [Figure 2: see original paper].

The model divides knowledge fusion into three stages:

Stage 1: Fields A and B develop independently according to their own paradigms, with knowledge flow occurring only within each field. Citations are primarily intra-field, and topological clustering yields only clusters corresponding to fields A and B.

Stage 2: Fields A and B begin cross-paradigm penetration with inter-field knowledge flow, but no new fusion field has formed. Clustering still produces only clusters for fields A and B, with no fusion cluster emerging.

Stage 3: Fields A and B undergo knowledge fusion, forming a new research paradigm that marks the emergence of a fusion field. Clustering now reveals clusters for fields A and B plus a new fusion cluster representing the merged field.

3 Verification Method for the Knowledge Fusion Process Model

3.1 Overall Verification Framework

The verification method addresses two questions: whether the theoretical model matches actual knowledge fusion processes and whether it is complete. The framework consists of three steps, as shown in Figure 3 [Figure 3: see original paper].

Step 1: Citation Network Construction. Develop search queries for the target fields, retrieve paper datasets from databases, extract direct citation information, and construct citation networks.

Step 2: Citation Network Analysis. Use visualization to identify cross-field citations, topological clustering to identify fusion clusters, and LDA topic modeling to extract cluster keywords, analyzing networks from these three perspectives.

Step 3: Knowledge Fusion State Identification. Based on annual network characteristics, identify the knowledge fusion stage for each year, identify fusion cluster research topics, and validate and refine the theoretical model.

3.2 Citation Network Construction

This study uses Thomson Reuters' Web of Science (WOS) database, specifically SCI-EXPANDED, SSCI, and A&HCI, which contain thousands of journals with complete citation information. Search queries are developed using field-specific keywords. To accurately analyze the fusion process, datasets for fields A and B must have the same end date and maximal overlapping time span. Since direct citations better reflect knowledge flow, Java is used to parse direct citation information from the complete datasets, constructing citation networks with papers as nodes and direct citations as links.

3.3 Citation Network Analysis

According to the theoretical model, different fusion stages exhibit distinct characteristics in cross-field citation count, fusion cluster presence, and cluster keywords. This analysis is prerequisite for state identification and model verification.

3.3.1 Cross-Field Citation Identification Cross-field citations reflect inter-field knowledge flow. In network visualization, nodes and links are color-coded: papers from different fields have different colors, and links inherit the color of intra-field connections while cross-field links appear white. The developed Citation Network Data Analyzer (CDA) software, shown in Figure 4 [Figure 4: see original paper], detects white links to identify cross-field citations.

3.3.2 Fusion Cluster Identification I. Rafols proposed that technology diversity uniformity reflects interdisciplinary fusion degree. Building on this, fusion clusters are identified by calculating the proportion of papers from each field within clusters. Clusters uniformly containing papers from both fields are identified as fusion clusters. The process uses Newman clustering algorithm, which automatically determines optimal cluster numbers based on network structure without manual specification, as shown in Figure 5 [Figure 5: see original paper].

3.3.3 Cluster Keyword Extraction Fusion clusters correspond to new fusion fields. LDA topic modeling, a common unsupervised method, extracts keywords from paper titles to identify research topics. Using K=1 topic, α=50,

β=0.01, and 1000 iterations, the model extracts keywords describing each cluster's research focus.

3.4 Knowledge Fusion State Identification

The theoretical model divides knowledge fusion into three stages. State identification determines each year's stage based on cross-field citation presence, fusion cluster existence, and cluster keywords. Table 1 summarizes the characteristics for each stage.

Stage 1: No cross-field citations; only original field clusters exist.
Stage 2: Cross-field citations present but no fusion clusters; only original field clusters exist.
Stage 3: Cross-field citations present; fusion clusters emerge alongside original field clusters.

By analyzing annual state characteristics and stage divisions, the theoretical model is validated. Since the model is derived from existing theory and actual processes are more complex, empirical observations further refine it.

4 Empirical Analysis of the Knowledge Fusion Process Model

Bioinformatics, formed through knowledge fusion between biological and information sciences, represents a strategic emerging field with significant characteristics. It serves as an ideal case for validating the citation network-based theoretical model.

4.1 Citation Network Construction

Following the verification framework, domain experts identified keywords for biological and information sciences. The search period was set to 1995-2016, limited to the USA to reduce noise. The retrieval queries were: "TS=((electrical computing) OR (information technology systems) OR (engineering electrical electronic) OR (engineering industrial) OR (software engineering) OR (computer artificial intelligence) OR (telecommunications) OR (computer hardware architecture) OR (information technology)) AND CU=USA" for information science, and "TS=((biochemistry molecular biology) OR (pharmacology pharmacy) OR (biochemical research methods) OR (genetics biology) OR (biochemistry molecular biology) OR (biotechnology) OR (chemistry medicinal) OR (microbiology)) AND CU=USA" for biological science. The search yielded 24,319 biological papers and 52,254 information science papers. Direct citation information was extracted to construct annual citation network time series.

4.2 Empirical Results

According to the theoretical model's stage characteristics, the bioinformatics fusion process was analyzed and matched to the model.

In visualizations, biological papers appear green and information science papers red. CDA software analysis revealed that before 1998, all citations were intra-field with no cross-field citations, and clustering produced only separate clusters for each field, indicating independent knowledge flow. In 1998, 少量 cross-field citations emerged but no fusion cluster formed, marking the transition from Stage 1 to Stage 2. The 1997 and 1998 visualizations are shown in Figure 6 [Figure 6: see original paper].

Between 1998 and 2003, cross-field citations increased but no fusion clusters appeared, indicating deepening inter-field knowledge flow without new field formation. By 2003, cross-field citations reached a critical mass, and clustering revealed a fusion cluster containing papers from both fields, signifying new paradigm formation. The 1999 and 2003 visualizations are shown in Figure 7 [Figure 7: see original paper], with the 2003 fusion cluster circled. In 2003, four clusters were identified: the fusion cluster contained 61.3% biological papers, while the other three clusters contained 99.9%, 0.9%, and 1.8% biological papers, respectively, demonstrating the fusion cluster's high diversity uniformity.

To identify the fusion field, fusion clusters were identified annually from 2003-2016, with visualizations for 2008 and 2016 shown in Figure 8 [Figure 8: see original paper]. LDA extracted keywords from each year's fusion cluster, presented in Table 2 . Bioinformatics research topics include genomics, proteomics, and biochips. The field emerged in the 1970s with analysis methods and databases. The Human Genome Project (HGP), launched in 1990 and reaching draft completion in June 2000, propelled bioinformatics development. Table 2 shows keywords evolving from gene chip-related terms (gene, expression, microarray) in 2003, to database terms (database, gene expression protein data) in 2004, to human genome sequencing terms (gene human database sequencing) after 2008, matching bioinformatics' evolution. This confirms the fusion cluster represents bioinformatics. Thus, the biological-information science fusion process transitioned from Stage 2 to Stage 3 in 2003.

The empirical analysis shows the fusion process spans six years in Stage 2 (1998-2003), from initial inter-field knowledge flow to new field formation. This temporal dimension, not analyzed in the original theoretical model, provides a refinement based on empirical observation.

Conclusion

This paper constructed a citation network-based theoretical model of knowledge fusion and validated it using bioinformatics. The findings demonstrate: (1) the model's validity in describing actual knowledge fusion processes; (2) knowledge fusion can be divided into three stages with distinct characteristics; (3) citation network visualization combined with topological clustering effectively describes fusion processes. The study focuses on science-driven fusion, though economic, market, and policy factors also influence field formation—their impacts require further investigation.

References

[1] LEE S, SEO L H, PARK Y. Using patent information for designing new product and technology: keyword-based technology roadmapping[J]. R&D management, 2008(38):169-188.
[2] KIM E, CHO Y, KIM W. Dynamic patterns of technological convergence in printed electronics technologies: patent citation network[J]. Scientometrics, 2014, 98(2):975-998.
[3] HACKLIN F. Management of convergence in innovation[M]. Heidelberg: Physica-Verlag HD, 2008.
[4] CURRAN C, LEKER J. Patent indicators for monitoring convergence?examples from NFF and ICT[J]. Technological forecasting & social change, 2011, 78(2):256-273.
[5] CHO Y, KIM E, KIM W. Strategy transformation under technological convergence: evidence from the printed electronics industry[J]. Social science electronic publishing, 2015, 674(67):106-114.
[6] ZHAO H Z. On the mining model of scientific discovery (Part 1)[J]. Science of science and management of S&T, 1981(2):3-5.
[7] ZHAO H Z. On the mining model of scientific discovery (Part 2)[J]. Science of science and management of S&T, 1981(3):34-38.
[8] ELEON, GUILD P. Using expert knowledge to envision future converging technologies[C]//Management of engineering and technology. Portland: Portland International Center for IEEE, 2007:878-882.
[9] KONG R. Patterns and processes of contemporary technology fusion: the case of intelligent robots[J]. Asian journal of technology innovation, 2007, 15(2):45-65.
[10] PARK H, YOON J. Assessing coreness and intermediarity of technology sectors using patent co-classification analysis: the case of Korean national R&D[J]. Scientometrics, 2014, 98(2):853-890.
[11] KIM D, LEE H, KWAK J. Standards as a driving force that influences emerging technological trajectories in the converging world of the internet and things: an investigation of the M2M/IoT patent network[J]. Research policy, 2017, 46(7):1234-1254.
[12] LIU X T, ZHU L, HE S, et al. A review of knowledge flow research[EB/OL]. Beijing: Sciencepaper Online [2018-06-11]. http://www.paper.edu.cn/releasepaper/content/201406-161.
[13] ZHOU Q J, YANG L Y, YUE T, et al. Research on discipline structure and knowledge flow based on journal co-citation and inter-citation networks[J]. Journal of intelligence, 2014, 33(8):84-91.
[14] LIANG Y X, LIU Z Y, YANG Z K. Analysis of knowledge flow theory in citation analysis[J]. Studies in science of science, 2010, 28(5):668-674.
[15] WANG L, ZHANG Q P. Research on knowledge flow process and mechanism based on citation network[J]. Journal of Harbin Institute of Technology (social sciences edition), 2014(1):110-116.
[16] SAKATA I, SASAKI H, AKIYAMA M, et al. Bibliometric analysis of service innovation research: identifying knowledge domain and global network of

knowledge[J]. Technological forecasting & social change, 2013, 80(6):1085-1093.
[17] KAJIKAWA Y, OHNO J, TAKEDA Y, et al. Creating an academic landscape of sustainability science: an analysis of the citation network[J]. Sustainability science, 2007, 2(2):221.
[18] NEWMAN M. The structure and function of complex networks[J]. Siam review, 2003, 45(1/2):40-45.
[19] ROSENBERG N. Technological change in the machine tool industry, 1840-1910[J]. Journal of economic history, 1963, 23(4):414-443.
[20] NELSON R, WINTER S. An evolutionary theory of economic change[M]. Cambridge: Belknap Press of Harvard University Press, 1982.
[21] FAGERBERG J, VERSPAGEN B. Technology-gaps, innovation-diffusion and transformation: an evolutionary interpretation[J]. Working papers, 2002, 31(8/9):1291-1304.
[22] CURRAN C, BRÖRING S, LEKER J. Anticipating converging industries using publicly available data[J]. Technological forecasting & social change, 2010, 77(3):385-395.
[23] ITTIPANUVAT V, FUJITA K, SAKATA I, et al. Finding linkage between technology and social issue: a literature based discovery approach[J]. Journal of engineering & technology management, 2014, 32(2):160-184.
[24] IWAMI S, MORI J, KAJIKAWA Y, et al. Comparison of indicators to detect emerging researches using time transition in quasi-crystals[C]//IEEE international conference on industrial engineering and engineering management. Beijing: IEEE, 2014:48-52.
[25] RAFOLS I, MEYER M. Diversity and network coherence as indicators of interdisciplinarity: case studies in bionanoscience[J]. Scientometrics, 2009, 82(2):263-287.
[26] LI Y Y, ZHAO Y L. Patent-based technology fusion analysis method and its application[J]. Studies in science of science, 2016, 34(2):203-211.
[27] LOU Y, YANG P P, HUANG L C, et al. Patent-based technology fusion measurement method: a case study of IT and electric vehicle technology fusion[J]. Modern information, 2017(8):144-155.
[28] NEWMAN M. Modularity and community structure in networks[C]//APS march meeting. Baltimore: American Physical Society, 2006:8577-8582.
[29] GRIFFITHS T, STEYVERS M. Finding scientific topics[J]. Proceedings of the National Academy of Sciences of the United States of America, 2004, 101:5228-5235.
[30] HOGEWEG P. The roots of bioinformatics in theoretical biology[J]. PLoS computational biology, 2011, 7(3):e1002021.
[31] GRAU J, BENGAL I, POSCH S, et al. VOMBAT: prediction of transcription factor binding sites using variable order Bayesian trees[J]. Nucleic acids research, 2006, 34(Web Server issue):W529-W533.
[32] FLEISCHMANN R, ADAMS M, WHITE O, et al. Whole-genome random sequencing and assembly of haemophilus influenzae rd[J]. Science, 1995, 269(5223):496-512.

Author Contributions

Zhou Yuan: Framework design, paper revision and writing guidance
Dong Fang: Experimental design and paper writing
Liu Yufei: Concept refinement, paper revision

Research on the Process of Knowledge Fusion in Emerging Fields: A Case Study of Bioinformatics

Zhou Yuan[1], Dong Fang[1], Liu Yufei[1],[2]

[1] School of Public Policy and Management, Tsinghua University, Beijing 100084
[2] Center for Strategic Studies, Chinese Academy of Engineering, Beijing 100088

Abstract: [Purpose/significance] Technology fusion is the driving force of new industries' formation and development. While knowledge fusion is the prerequisite of technology convergence, it is of great significance in guiding the formation and development of new industries. [Method/process] Firstly, this paper built a theoretical model which using the citation network to characterize the knowledge fusion process based on the existing research. Then, based on the characteristics of paper citation network in each stage of knowledge fusion, the paper introduced a verification method. Finally, it conducted an empirical analysis in the field of bio-information technology. [Result/conclusion] The results of empirical analysis in the field of biological information show the validity of the theoretical model. Therefore, It can be deemed as a new method on studying the knowledge fusion.

Keywords: knowledge fusion, knowledge flow, citation network, emerging field

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