Measuring Technology Diffusion Dynamics Using Patent Full-Text Data and Machine Learning

Alex J. Yang[1,*], Star X. Zhao[2], Yi Bu[3] and Sanhong Deng[1]

[1]School of Information Management, Nanjing University, Nanjing, China
[2]National Institute of Intelligent Evaluation and Governance, Fudan University, Shanghai, China
[3]Department of Information Management, Peking University, Beijing, China

Abstract

Understanding the dynamics of technology diffusion is critical for tracking innovation and its impact across industries. This study leverages full-text patent data and advanced machine learning to develop a novel framework for quantifying technology diffusion. Using the PatentsView dataset from the U.S. Patent and Trademark Office, we analyze 6,913,074 utility patents' full texts with a Doc2Vec model to map semantic relationships in a high-dimensional knowledge space. We propose a dynamic measurement framework based on knowledge distances and temporal evolution, validated through case studies of disruptive patents, including Google's PageRank. Our results reveal patterns of cross-disciplinary diffusion, with metrics like the 5-Year Average Diffusion Index and Diffusion Trend Index offering robust tools for comparative analysis. This approach provides a scalable, data-driven method to study technology diffusion's complex dynamics.

Keywords

Technology Diffusion, Machine Learning, Knowledge Distance, Technological Innovation

1. Introduction

Technological innovation drives economic growth, societal change, and industrial transformation [1, 2]. Yet, quantifying how technologies spread across domains—termed technology diffusion—remains a challenge due to the complexity of knowledge flows [3, 4, 5]. Traditional metrics, such as patent citation counts, capture only a fraction of diffusion dynamics, often missing semantic nuances and cross-disciplinary impacts [6, 7]. Patent documents, rich in technical and contextual detail, offer a unique lens to study diffusion, particularly when analyzed in full text rather than metadata alone [8, 9].

Recent advances in natural language processing (NLP) and machine learning enable deeper semantic analysis of large-scale textual data [10, 11]. Models like Doc2Vec [12] can transform documents into vector representations, capturing latent relationships in knowledge structures [13, 14, 15]. This study harnesses such tools to analyze patent full-text data, aiming to measure technology diffusion dynamically and comprehensively.

2. Patent Full-Text Data and Neural Network Embedding Model

Patents encapsulate technical knowledge, making them ideal for studying technology diffusion. We use the USPTO's PatentsView dataset, which includes metadata for 8,206,179 utility patents granted from 1976 to 2024. Through web scraping, we collected full-text data for 6,913,074 patents granted between 1976 and 2020, covering titles, abstracts, claims, and descriptions. The dataset also includes disambiguated metadata, such as millions of inventors, billions of citations, and Cooperative Patent Classification (CPC) codes, enabling multidimensional analysis. Full-text data reveal semantic and

AII-EEKE 2025 Joint Workshop of the 5th AI + Informetrics (AII) and the 6th Extraction and Evaluation of Knowledge Entities from Scientific Documents (EEKE)
*Corresponding author.
alexjieyang@outlook.com (A. J. Yang); xzhao@fudan.edu.cn (S. X. Zhao); buyi@pku.edu.cn (Y. Bu); sanhong@nju.edu.cn (S. Deng)
0000-0002-5385-5761 (A. J. Yang); 0000-0003-2549-4580 (Y. Bu); 0000-0002-6910-3935 (S. Deng)
© 2025 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

structural details beyond metadata, capturing technical content comprehensively. To ensure quality, we preprocessed texts by removing non-semantic elements (e.g., boilerplate), stop words, and applying stemming and tokenization. This enhances signal-to-noise ratios, making the data suitable for semantic analysis.

To represent patent texts semantically, we employ the Doc2Vec model, an extension of Word2Vec that maps documents to fixed-length vectors. Unlike traditional methods (e.g., TF-IDF), Doc2Vec captures contextual relationships, ideal for long technical texts. We implemented Doc2Vec using Python's Gensim library, selecting the Distributed Bag of Words (DBOW) architecture for its efficiency and semantic accuracy. DBOW predicts context words from document vectors, learning robust representations without relying on word order.

Each patent is assigned a unique 300-dimensional vector, embedded in a knowledge space where cosine distances measure semantic differences. We optimized parameters through experimentation: vector dimension = 300, window size = 10, negative samples = 5, minimum word frequency = 5, and 50 training epochs. This setup balances computational efficiency and semantic quality, enabling precise diffusion analysis.

3. Technology Diffusion Measurement Framework

3.1. Dynamic Knowledge Distance

To systematically investigate the dynamic evolution of technology diffusion, this study develops a measurement framework based on dynamic knowledge distance evolution. First, we use the Doc2Vec model to embed preprocessed patent full-text data, representing each patent as a 300-dimensional vector. The collection of all patent vectors forms a high-dimensional knowledge space, enabling precise quantification of semantic similarity and knowledge distances between patents.

We illustrate the measurement approach using the Google PageRank patent as a case study. Filed by Larry Page and others in 1998, the PageRank algorithm calculates webpage importance based on hyperlink structures, forming the foundation of modern search engine technology. As of 2025, it has been cited 1,078 times. However, relying solely on citation counts fails to fully capture the dynamic evolution and cross-disciplinary impact of technology diffusion. We propose a dynamic measurement method integrating temporal dimensions and knowledge distances, outlined as follows:

1. **Temporal Segmentation:** The citing patents of the PageRank patent are divided into annual subsets based on their grant years, forming a time series of citation sets. For year t, the set of citing patents for a core patent i is denoted as $\text{Citations}(i, t)$. This time-series analysis dynamically tracks the evolution of knowledge flows.
2. **Knowledge Distance Calculation:** To quantify knowledge differences between patents, we use cosine distance as the metric. For any two patents i and j with vector representations $\mathbf{v}_i$ and $\mathbf{v}_j$, the cosine similarity is defined as:

$$\text{Cosine Similarity}(i, j) = \frac{\mathbf{v}_i \cdot \mathbf{v}_j}{\| \mathbf{v}_i \| \| \mathbf{v}_j \|}, \tag{1}$$

   The corresponding knowledge distance is:

$$\text{Distance}(i, j) = 1 - \text{Cosine Similarity}(i, j), \tag{2}$$

   This distance ranges from 0 to 1, where smaller values indicate greater similarity in knowledge content, and larger values indicate greater divergence.
3. **Technology Diffusion Quantification:** The diffusion strength of patent i in year t is defined as the average knowledge distance between the core patent and all its citing patents in that year:

$$\text{Diffusion}(i, t) = \frac{1}{|\text{Citations}(i, t)|} \sum_{j \in \text{Citations}(i,t)} \text{Distance}(i, j), \tag{3}$$

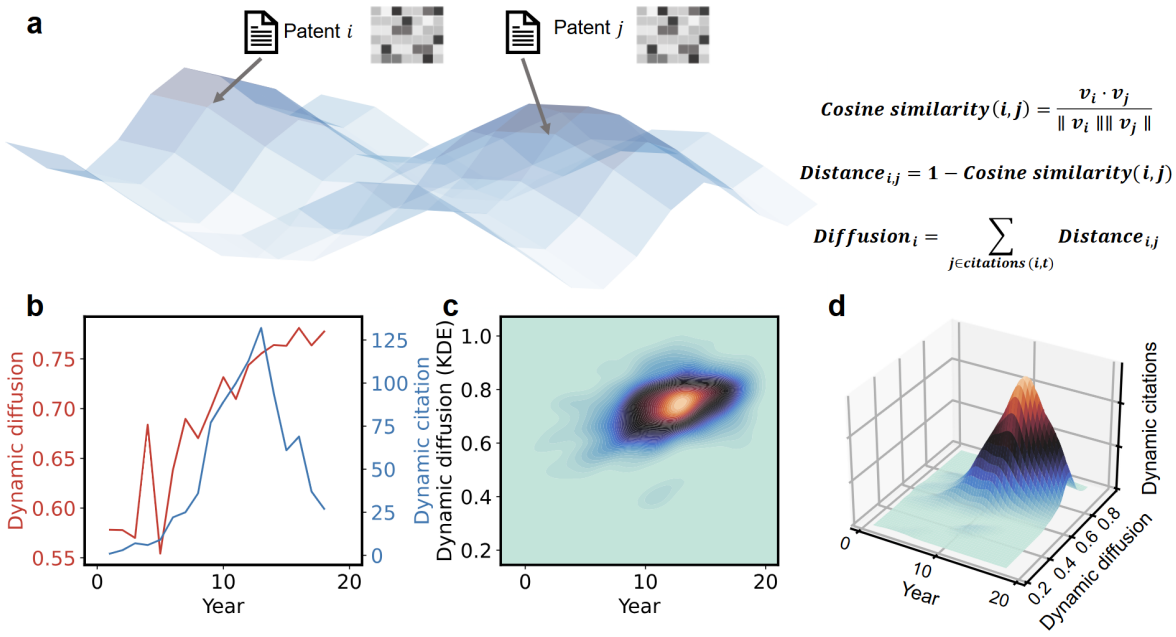


Figure 1: Technology Diffusion Measurement Framework.

Here, $|\text{Citations}(i,t)|$ denotes the number of citing patents in year t. By computing $\text{Diffusion}(i,t)$ for each year, we obtain the dynamic trend of technology diffusion.

Applying this framework to the PageRank patent reveals key insights. As shown in Figure 1, the annual citation count rises initially, peaks around 2008, and then declines, indicating rapid early influence followed by a gradual decrease. However, the annual average technology diffusion exhibits a sustained upward trend, suggesting that the PageRank patent's knowledge increasingly permeates diverse technical domains over time, reflecting enhanced cross-disciplinary impact.

To further explore the spatial distribution of diffusion, we employ Kernel Density Estimation (KDE) to generate dynamic density plots of technology diffusion (Figure 1). In these plots, the x-axis represents time, the y-axis represents knowledge distance, and high-density regions in the upper-right indicate greater knowledge distances in later years. This pattern underscores the patent's long-term diffusion across broader domains. To integrate citation counts and diffusion, we construct a three-dimensional co-density plot , with time on the x-axis, knowledge distance on the y-axis, and citation counts on the z-axis. The results show an early citation peak followed by a sustained rise in diffusion, consistent with the Sleeping Beauty phenomenon of disruptive technologies [16], where initial influence is latent but manifests broadly over time.

To validate the framework's robustness and generalizability, we applied it to a set of disruptive technology patents identified by Kelly et al. [17], selecting those with over 500 citations. For each patent, we calculated annual diffusion strength and density, constructing three-dimensional co-distribution plots. As shown in Figure 2, all disruptive patents exhibit significant growth in diffusion strength over time, indicating expanding influence across diverse knowledge domains. Citation peaks typically occur years after grant, with diffusion peaks lagging behind, consistent with the Sleeping Beauty pattern. High-density diffusion regions cluster in later time periods and higher knowledge distances, reflecting long-term cross-disciplinary impact. The citation-diffusion co-distribution further highlights early citation peaks followed by sustained diffusion growth, underscoring the complex dynamics of knowledge flow in disruptive technologies.

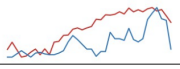
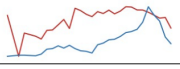
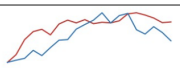
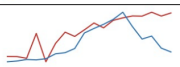
Patent IDs	Patent titles	Inventors	Assignee	Citation	Dynamic diffusion/ Dynamic citations	Diffusion density	Diffusion & citations
US4399216	Processes for inserting DNA into eucaryotic cells and for producing proteinaceous materials	Richard Axel et al.	Columbia University	899			
US4683195	Process for amplifying, detecting, and/or-cloning nucleic acid sequences	Kary B. Mullis et al.	Cetus Corporation	5073			
US4736866	Transgenic non-human mammals	Philip Leder et al.	Harvard University	518			
US4816397	Multichain polypeptides or proteins and processes for their production	Michael A. Boss et al.	Celltech	1028			
US4816567	Recombinant immunoglobin preparations	Shmuel Cabilly et al.	Genentech	4296			
US5225539	Recombinant altered antibodies and methods of making altered antibodies	Gregory P. Winter	Medical Research Council	1734			
US5770429	Transgenic non-human animals capable of producing heterologous antibodies	Nils Lonberg et al.	GenPharm International	831			
US5939598	Method of making transgenic mice lacking endogenous heavy chains	Raju Kucherlapati et al.	Abgenix	951			
US5960411	Method and system for placing a purchase order via a communications network	Jeffrey P. Bezos et al.	Amazon	2318			
US6285999	Method for node ranking in a linked database	Lawrence Page et al.	Stanford University	1078			

Figure 2: Technology Diffusion of Disruptive Patents.

3.2. Higher-Order Diffusion Metrics

Dynamic diffusion strength varies by patent age and domain, complicating comparisons. We introduce two standardized metrics over a 5-year window, reflecting typical diffusion cycles:

1. **5-Year Average Diffusion Index:**

$$\text{Diffusion}_5(i,t) = \frac{1}{5}\sum_{k=0}^{4}\text{Diffusion}(i,t+k), \tag{4}$$

   High values indicate broad, cross-disciplinary impact. For PageRank, this index highlights rapid early diffusion across fields like machine learning.

2. **5-Year Diffusion Trend Index:**

$$\text{Diffusion trend}_5(i,t) = \frac{\sum_{k=0}^{4}\left(\text{Diffusion}(i,t+k)-\bar{D}\right)\left(k-\bar{T}\right)}{\sqrt{\sum_{k=0}^{4}\left(\text{Diffusion}(i,t+k)-\bar{D}\right)^2}\sqrt{\sum_{k=0}^{4}\left(k-\bar{T}\right)^2}}, \tag{5}$$

   Positive trends signal growing diffusion; negative trends suggest decline. These metrics enable robust cross-patent comparisons, accounting for temporal and domain variations.

4. Conclusion

This study presents a novel framework for measuring technology diffusion using patent full-text data and machine learning. By embedding patents in a semantic knowledge space, we quantify diffusion through knowledge distances and temporal trends, validated with disruptive technologies. These tools offer researchers and policymakers a data-driven approach to track innovation's spread, informing strategies for technology development and adoption. As NLP and patent databases evolve, this framework can adapt to new data sources, deepening our understanding of technological progress.

Acknowledgments

This research was supported by the National Natural Science Foundation of China Young Student (Doctoral) Project (#724B2015).

Declaration on Generative AI

The author(s) have not employed any Generative AI tools.

References

[1] E. M. Rogers, A. Singhal, M. M. Quinlan, Diffusion of innovations, in: An integrated approach to communication theory and research, Routledge, 2014, pp. 432–448.
[2] M. S. Granovetter, The strength of weak ties, American Journal of Sociology 78 (1973) 1360–1380. doi:10.1086/225469.
[3] A. J. Yang, On the temporal diversity of knowledge in science, Journal of Informetrics 18 (2024) 101594. doi:10.1016/j.joi.2024.101594.
[4] P. Ji, et al., Signal propagation in complex networks, Physics Reports 1017 (2023) 1–96. doi:10.1016/j.physrep.2023.03.005.
[5] D. Wang, A.-L. Barabási, The science of science, Cambridge University Press, 2021.
[6] A. J. Yang, et al., Are disruptive papers more likely to impact technology and society?, Journal of the Association for Information Science and Technology 76 (2025) 563–579. doi:10.1002/asi.24947.
[7] A. J. Yang, Unraveling topic switching and innovation in science, Information Processing & Management 62 (2025) 104171. doi:10.1016/j.ipm.2025.104171.
[8] M. P. Polak, D. Morgan, Extracting accurate materials data from research papers with conversational language models and prompt engineering, Nature Communications 15 (2024) 1569. doi:10.1038/s41467-024-45914-8.
[9] S. Huang, et al., Fine-grained citation count prediction via a transformer-based model with among-attention mechanism, Information Processing & Management 59 (2022) 102799. doi:10.1016/j.ipm.2021.102799.
[10] H. Hu, A. J. Yang, S. Deng, D. Wang, M. Song, Cotel-d3x: A chain-of-thought enhanced large language model for drug–drug interaction triplet extraction, Expert Systems with Applications 273 (2025) 126953. doi:10.1016/j.eswa.2025.126953.
[11] A. J. Yang, Y. Bu, Y. Ding, M. Liu, Knowledge entity extraction from scientific documents using deep learning, in: Proceedings of Joint Workshop of the 5th Extraction and Evaluation of Knowledge Entities from Scientific Document and the 4th AI + Informetrics, iConference, 2021, pp. 145–149.
[12] Q. Le, T. Mikolov, Distributed representations of sentences and documents, in: International Conference on Machine Learning, PMLR, 2014, pp. 1188–1196.
[13] V. Tshitoyan, et al., Unsupervised word embeddings capture latent knowledge from materials science literature, Nature 571 (2019) 95–98. doi:10.1038/s41586-019-1335-8.
[14] J. Sourati, J. A. Evans, Accelerating science with human-aware artificial intelligence, Nature Human Behaviour 7 (2023) 1682–1696. doi:10.1038/s41562-023-01648-z.
[15] D. Kang, R. S. Danziger, J. Rehman, J. A. Evans, Limited diffusion of scientific knowledge forecasts collapse, Nature Human Behaviour 9 (2025) 268–276. doi:10.1038/s41562-024-02041-0.
[16] A. J. Yang, S. X. Zhao, S. Deng, Revisiting delayed recognition in science: A large-scale and comprehensive study, Journal of Information Science (2024). doi:10.1177/01655515241244462.
[17] B. Kelly, D. Papanikolaou, A. Seru, M. Taddy, Measuring technological innovation over the long run, American Economic Review: Insights 3 (2021) 303–320. doi:10.1257/aeri.20190499.