Proceedings of the 5th International Conference on Signal Processing and Machine Learning
DOI: 10.54254/2755-2721/108/2025.20802

Identifying Innovation Patterns in Scientific Research from the Perspective of Semantic Change

Zeyu Yang[1,a,*]

[1]*School of Economics and Management, Nanjing University of Science and Technology, Xiao Ling Wei 200, Nanjing, China*
a. yangzeyu20000@163.com
**corresponding author*

Abstract: The degree of semantic change reflects the innovativeness of scientific knowledge, offering a new way to identify research innovation patterns. To achieve this, this paper identified problem and method entities from AI literature in OpenAlex database and applied incremental learning and temporal annotation for word2vec training. Based on diachronic co-occurrence type vector of entity words, this paper conducts clustering analysis and finds three innovation patterns: emerging pattern, disappearing pattern, persisting Pattern. From these patterns, this paper finds the primary driving factors of scientific research innovation are largely twofold: the emergence of new co-occurrence neighbors and the disappearance of old co-occurrence neighbors. Additionally, with the aid of diachronic co-occurrence type vectors, this paper builds an innovation prediction model, which can predict concepts and technologies with innovative potential.

Keywords: innovation, diachronic semantic change, word co-occurrence, incremental learning

1. Introduction

The exploration of knowledge with innovative potential is a common academic pursuit [1]. Previous studies have shown that entity words, as the carrier of knowledge, often reflects the degree of innovation through its semantic change [2]. In other words, analyzing patterns of lexical diachronic semantic change can help predict the trajectory of knowledge evolution [3]. The distributional hypothesis posits that the meaning of a word can be understood through the words it co-occurs with [4], and novel word co-occurrences within a word's context are likely to lead to a higher degree of diachronic semantic change [5]. Similarly, the theory of combinatorial innovation suggests that innovation emerges from knowledge combination [6], which can be reflected in the co-occurrence of words within the same context [7]. Therefore, with the aid of co-word analysis, this paper aims to explore innovation patterns in scientific research from the perspective of lexical diachronic semantic change.

2. Method

This paper adheres to a fundamental hypothesis that the greater the diachronic semantic change of a word, the higher its innovativeness [1]. A smaller cosine similarity indicates a more intense

© 2025 The Authors. This is an open access article distributed under the terms of the Creative Commons Attribution License 4.0 (https://creativecommons.org/licenses/by/4.0/).

DOI: 10.54254/2755-2721/108/2025.20802

diachronic semantic change and, consequently, a higher degree of innovation, and vice versa. By integrating the distributed hypothesis theory with the combinatorial innovation theory, this paper infers that changes in the co-occurring neighboring words of an entity word in its previous stage will affect its future innovativeness. To quantitatively measure this impact, this paper construct a diachronic co-occurrence type vector for every entity words: first classifying co-occurring neighbor words into 6 types, as shown in Figure 1, then quantifying the proportion of each co-occurrence type relative to the total of all co-occurrence types based on their total frequency (specifically for C2, its total frequency is the average of the frequency of t1 and t2). Then this paper generates a six-dimensional vector: (C1A, C1B, C2A, C2B, C3A, C3B), based on the proportions of the total frequencies of co-occurring neighboring words for every entity words.

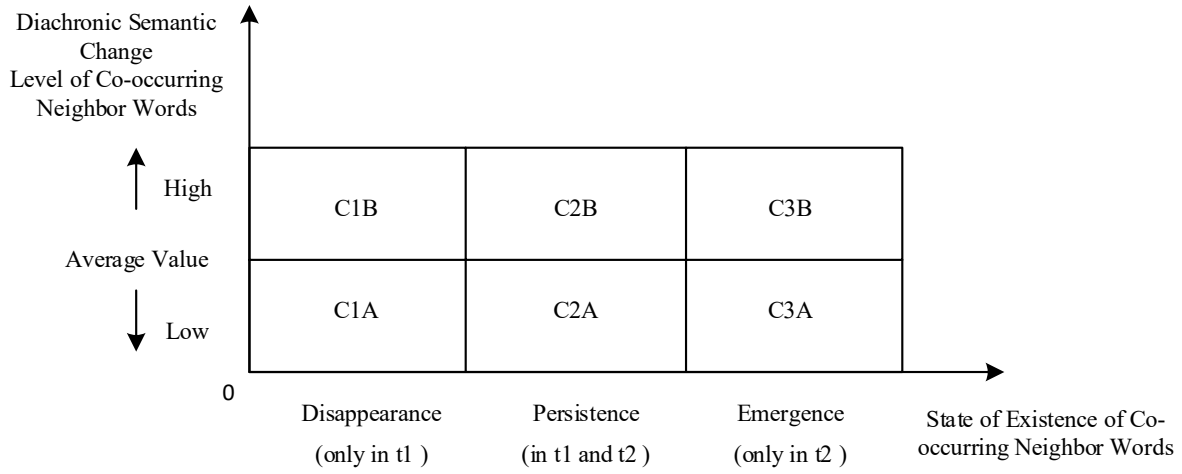

Figure 1: Six types of co-occurring neighbor words based on two criteria.

The technical path of this research is shown in Figure 2. To ensure the reliability and accuracy of the experimental results, this study focuses on artificial intelligence, a rapidly growing field of artificial intelligence, where research output has surged in recent years. Therefore, this paper retrieves papers of AI-related literature from OpenAlex published between 1996 and 2023. This paper divides the data into four stages: 2008-2011 (T1), 2012-2015 (T2), 2016-2019 (T3), and 2020-2023 (T4), considering both 2008 and 2016 are key milestones in AI development [8]. To identify entities, this paper applied the domain glossary developed by Chen Guo et al. to match the title, abstract, and keywords of each document [9]. With the incrementally trained word2vec model, this paper calculates diachronic semantic change between entity terms and their corresponding temporally annotated copies for adjacent stages. Then this paper constructs diachronic co-occurrence type vectors for all entity words, using them to analyze innovation pattern. This paper also uses these vectors to construct innovation prediction model, aimed at demonstrating the effectiveness and application prospects of the innovation pattern. Specifically, this paper use the Root Mean Square Error (RMSE) as evaluation metric and train innovation prediction models based on Linear Regression (LR), Decision Trees (DT), Support Vector Regression (SVR), Random Forest (RF), and K-Nearest Neighbor Classification (KNN).

DOI: 10.54254/2755-2721/108/2025.20802

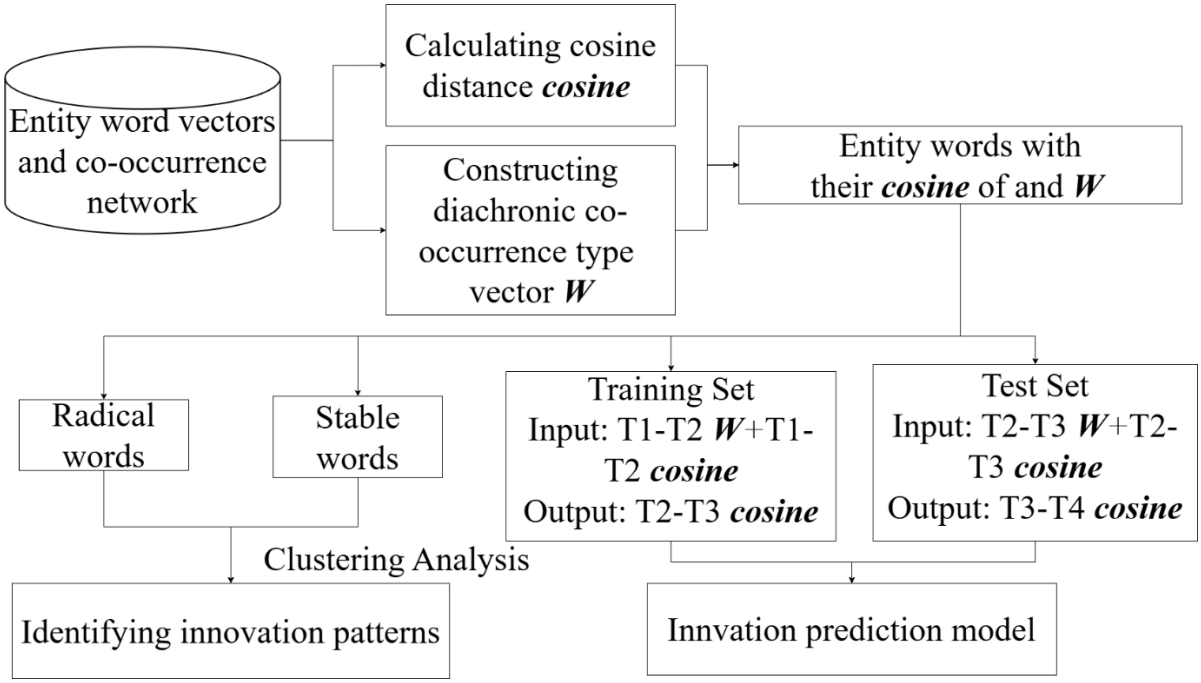

Figure 2: Schematic of the technical path.

3. Experimental Analysis

3.1. Dataset Construction

This paper retrieved 4,380,750 bibliographic records from the OpenAlex database on artificial intelligence (AI) publications from 1996 to 2023. The distribution of literature data in the field of artificial intelligence from the OpenAlex database is shown in Figure 3. The number of publications in the field of artificial intelligence has shown a consistent upward trend, with a rapid increase in publication volume especially after 2016. This indicates that artificial intelligence technologies have entered a period of rapid development and have attracted greater attention from researchers and industries alike.

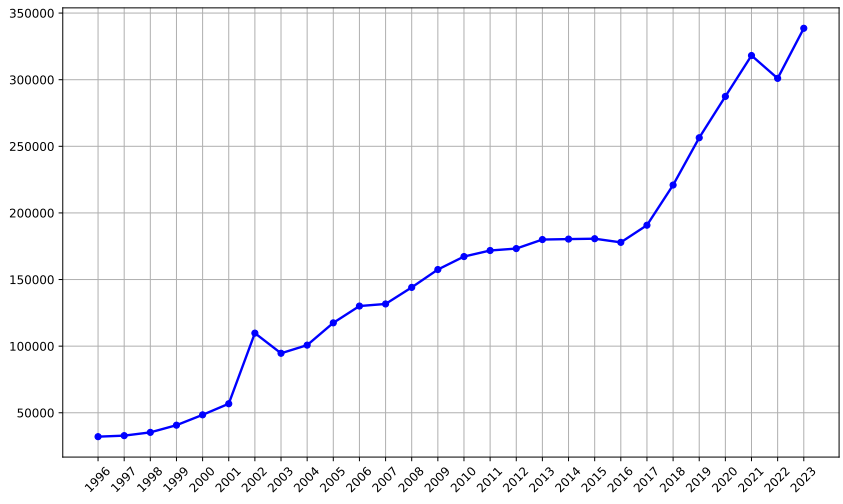

Figure 3: Annual distribution of the amount of literature data

DOI: 10.54254/2755-2721/108/2025.20802

After multiple experiments, the optimal parameter configuration for the study is presented in the Table 1. These parameters were meticulously adjusted and balanced to ensure that the model could adequately capture the semantic features and associations of the words, thereby enhancing the quality and expressiveness of the word embeddings. Finally, if two entity terms appear together in the bibliographic data of the same document, they are considered a pair of co-occurring entities. In this way, co-occurrence relationships are identified, and diachronic co-occurrence type vectors are constructed for all entity terms.

Table 1: Word2Vec parameter settings.

Parameters	Description	Setting Value
vector_size	Dimension of word vectors	100
min_count	Minimum occurrence frequency of words in the corpus	5
window	Context window size	5
sg	sg=0 for CBOW algorithm; sg=1 for Skip-gram algorithm	0
epochs	Number of iterations	100

3.2. Identifying Innovation Patterns

Ranked by the degree of diachronic semantic change, the top 1% are "radical words", while the bottom 1%, are "stable words". This paper uses these words of T2-T3 and T3-T4, with their corresponding diachronic co-occurrence type vector of T1-T2 and T2-T3, to conduct clustering analysis. The clustering results for T1-T2 and T2-T3 are almost consistent, as shown in Figure 4 and 5. A cluster is classified as an innovation pattern if radical words make up over approximately 80%, and as a stable pattern if stable words exceed approximately 80%. Therefore, we can identify the following three basic innovation patterns in scientific research:

Emerging Pattern: This innovation pattern indicates high knowledge novelty, characterized by the emergence of many new co-occurring neighbor words, with minimal disappearance or persistence of existing neighbors. This pattern is represented by Cluster_1, where the majority are radical words, with the highest proportion of Type C3 co-occurrences. The combined proportion of C3A and C3B exceeds 80%.

Disappearing Pattern: This innovation pattern also represents knowledge novelty, with the disappearance of many existing co-occurring neighbor words, accompanied by a small number of new and persistent co-occurring neighbor words. This pattern is represented by Cluster_2, where the majority are radical words, with the highest proportion of Type C1 co-occurrences. The combined proportion of C1A and C1B is approximately 70%.

Persisting Pattern: This pattern indicates a lack of innovation, with persistent co-occurring neighbor words dominating, and newly emerging neighbors with lower innovation potential. This pattern is represented by Cluster_3, where the majority are stable words, with the highest proportion of Type C2 co-occurrences. The combined proportion of C2A and C2B is approximately 50%.

DOI: 10.54254/2755-2721/108/2025.20802

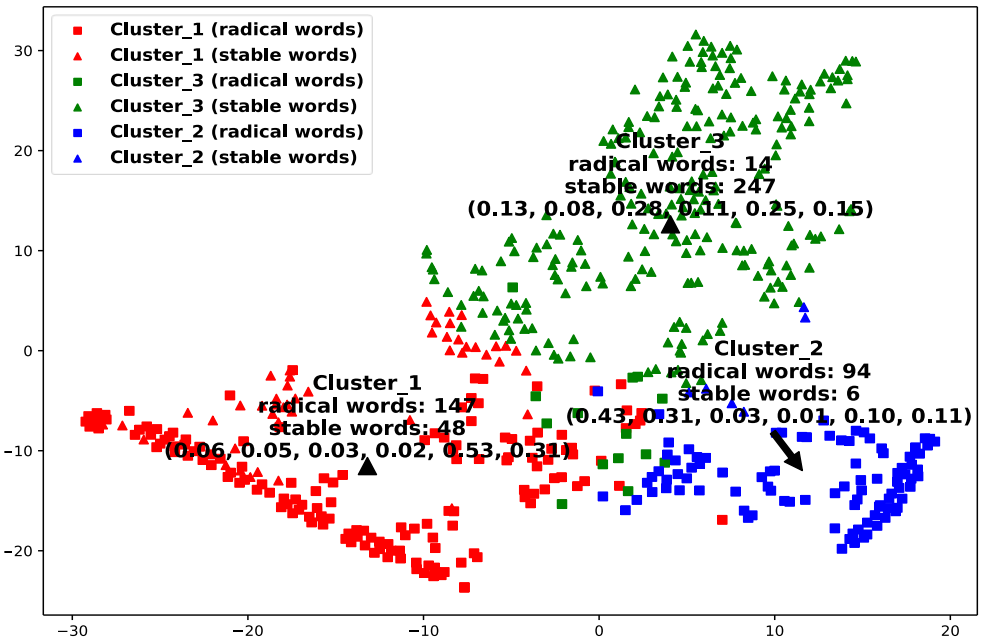

Figure 4: Clustering results with average diachronic co-occurrence type vectors of T1-T2.

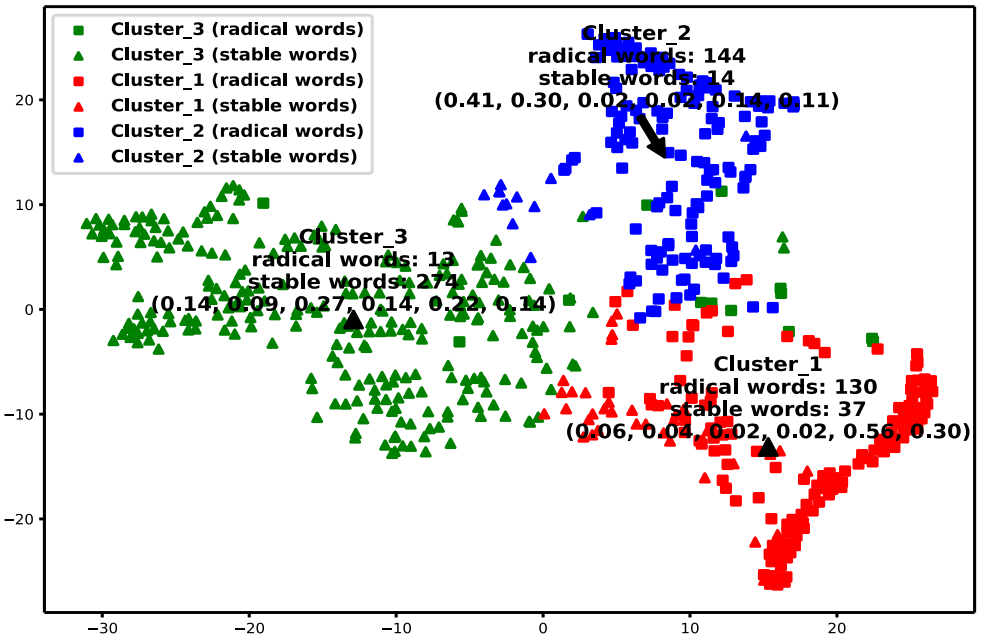

Figure 5: Clustering results with average diachronic co-occurrence type vectors of T2-T3.

3.3. Innovation Prediction Model

Table 2 lists RMSE of five prediction algorithms across seven TOPN values, where "ALL" represents all entities in the test set. It can be observed that SVR consistently delivers the best performance across most TOPN values, which is ranked by the degree of entity semantic change from T3 to T4. Therefore, SVR could be considered as a preferred algorithm for building higher-performance prediction models in the future. Additionally, the proposed prediction model achieves lower RMSE values compared to the baseline model across all TOPN values. Notably, the model demonstrates significantly superior performance over the baseline for TOP50, TOP100, TOP200, TOP500,

DOI: 10.54254/2755-2721/108/2025.20802

TOP1000, TOP5000 and All. This indicates that the model excels in predicting entities undergoing significant semantic changes in the previous stage, providing valuable insights for researchers to assess whether currently emerging entities are likely to be applied in various scenarios in the future.

Table 2: Comparison of the prediction effect of the proposed model and the baseline model -RMSE.

Model	TOP50	TOP100	TOP200	TOP500	TOP1000	TOP5000	All
DT	0.1716	0.1670	0.1522	0.1269	0.1134	0.1033	0.1233
RF	0.1640	0.1565	0.1410	0.1164	0.1019	**0.0885**	0.0907
SVR	**0.1313**	**0.1273**	**0.1189**	**0.1014**	**0.0954**	0.1019	**0.0884**
LR	0.1581	0.1526	0.1405	0.1171	0.1043	0.0915	0.0893
KNN	0.1775	0.1680	0.1535	0.1261	0.1077	0.0808	0.0956
Baseline	0.2177	0.2121	0.1968	0.1681	0.1504	0.1130	0.0998

4. Conclusion

The three fundamental innovation patterns identified by this study reveal that unless an entity word experiences the disappearance of many co-occurring neighbor words, or the addition of many new co-occurring neighbor words during a certain phase, it will not exhibit significant innovation. With the innovation prediction model built in this study, researchers can identify concepts and technologies with innovative potential and ultimately facilitate breakthroughs at the scientific frontier.

Acknowledgements

This study was supported by the Postgraduate Research & Practice Innovation Program of Jiangsu Province [KYCX23_0633].

References

[1] *Gebhart, T. and Gebhart, R.J. (2023) The Emergence of Higher-Order Structure in Scientific and Technological Knowledge Networks. Academy of Management Proceedings, 2023(1), 12214.*

[2] *Olson, J.A., Nahas, J. and Chmoulevitch, D. (2021) Naming Unrelated Words Predicts Creativity. Proceedings of the National Academy of Sciences, e202340118.*

[3] *Mahanty, S., Boons, F., Handl, J. and Batista-Navarro, R. (2021) Computation of Semantic Change in Scientific Concepts: Case Study of "Circular Economy". In: Tahmasebi, N. (Ed.), Computational Approaches to Semantic Change, Language Science Press, 2nd Edition, 123.*

[4] *Firth, J.R. (1957) A Synopsis of Linguistic Theory 1930-1955. In: Langendoen, D.T. (Ed.), Studies in Linguistic Analysis, Linguistic Society of America Press, 1-10.*

[5] *Hamilton, W. L., Leskovec, J., & Jurafsky, D. (2016). Diachronic word embeddings reveal statistical laws of semantic change. arXiv preprint arXiv:1605.09096.*

[6] *Koestler, A. (1964) The Act of Creation. Hutchinson, London.*

[7] *Arts, S., Hou, J. and Gomez, J.C. (2020) Natural Language Processing to Identify the Creation and Impact of New Technologies in Patent Text: Code, Data, and New Measures. Research Policy, 50(2), 104144.*

[8] *Zhang, C. and Lu, Y. (2021) Study on Artificial Intelligence: The State of the Art and Future Prospects. Journal of Industrial Information Integration, 23(1), 100224.*

[9] *Chen, G., Peng, J., Xu, T., et al. (2023) Extracting Entity Relations for "Problem-Solving" Knowledge Graph of Scientific Domains Using Word Analogy. Aslib Journal of Information Management, 75(3), 481-499.*