Material performance evolution discovery based on entity extraction and social circle theory*

Jinzhu Zhang[1,*], Wenwen Sun

Department of Information Management, School of Economics and Management, Nanjing University of Science and Technology, Nanjing China

Abstract

Topic evolution analysis describes the emerge, develop, and extinct of topics in a field, which can help researchers understand the history and current situation of the research field. However, the material patent text has a certain domain specificity, and the general entity extraction models cannot extract special entities effectively. Moreover, the belief that topics with high similarity have evolution relationship contradicts the rule of "first the change, then the new topic", which cannot clearly present the dynamic changes and accumulation of topics. Therefore, we design a method to extract the material performance entities accurately and construct dynamic evolution path for material performance topics. Firstly, we propose a material entity extraction model BERT-BiLSTM-CRF, which integrates syntactic dependency analysis and attention mechanism, realizing the accurate extraction of material performance entities. Secondly, we design an algorithm for identifying the evolution relationship between performance nodes based on ring boundaries, which can mine the evolution relationship between performance nodes and existing topics, realizing the dynamic accumulation and change of topics. Finally, we construct the dynamic evolution path of material performance, exploring the complex associations of material performance. Experiments in the field of metal materials confirm that the proposed method can effectively construct the dynamic evolution path of material performance topics, which makes the evolution relationships between topics more abundant and interpretable.

Keywords

Entity extraction, material performance evolution, patent entity relationship

1. Introduction

With the emergence of a large number of patents on materials manufacturing and materials innovation, it has become critical to explore the complex associations and evolutionary trends of material performance. Such exploration can help researchers deepen their understanding of material performance and promotes the invention of new materials [1].

Current researches on the evolution of material performance is mainly based on the material microstructure and the manufacture process, considering that changes in the former brings about changes in material performance [2]. This requires readers and researchers to have a certain

Joint Workshop of the 5th Extraction and Evaluation of Knowledge Entities from Scientific Documents (EEKE2024) and the 4th AI + Informetrics (AII2024)
* Corresponding author
zhangjinzhu@njust.edu.cn (J. Zhang)
0000-0001-7581-1850 (J. Zhang); 0009-0007-2773-4679(W. Sun)

level of expertise. And in terms of evolutionary analysis methods, most of the existing studies are based on topic similarity, believing that there is an evolution relationship between two topics whose similarity above a threshold [3, 4]. However, the topic itself is dynamically changing and accumulating, there should be " first the change in material performance, then the new material performance topic ", so the establishment of an evolution relationship based on the similarity between topics is biased.

Therefore, this paper takes the material performance as the research object. Firstly, we desgin an entity extraction model to extract the relevant material performance entities from the material patents. Secondly, we propose a method for constructing the topic evolution path by introducing the social circle theory, realizing the dynamic accumulation of material performance topics and the construction of evolution path. Finally, we explore the complex associations in the evolution of material performance on the basis of the dynamic evolution path.

2. Literature review

This paper conduct the discovery of the material performance evolution based on entity extraction and social circle theory. Therefore, we review the literature about material performance evolution, entity extraction , evolutionary analysis methods.

2.1 Material performance evolution

Through the review of existing studies, we learned that the material performance mainly depends on their microstructure [5], while the manufacture process directly affect the material's microstructure. So the current researches about this are mainly divided into the three perspectives: "performance evolution - microstructure" , "performance evolution - microstructure - manufacture process" , and "performance evolution - manufacture process". The specific relationships are shown in Figure 1:

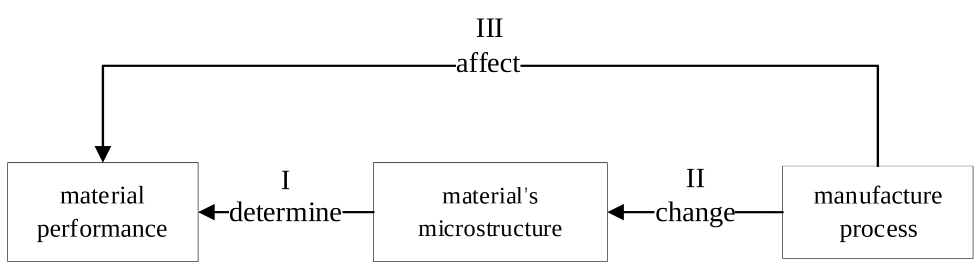


Figure 1: Different Perspectives in the Analysis of The Evolution of Material Performance

Perspectives I and II are usually based on the microstructural level of materials, providing a deeper understanding of the performance evolution [2]. For example, Liu et al. [6] analyzed the pore structure characteristics of different parts of the concrete core samples, and then explores the performance evolution regulations of dam concrete; Li et al. [7] prepared Ti(C,N)cermet by using nitrogen atmosphere sintering, and investigated the effects of different sintering temperatures and holding times on the performance of alloy hardness and fracture toughness in Ti(C,N)-based cermet cores. Perspectives III aims to regulate the evolution of material performance by analyzing and optimizing the manufacture process, revealing the influence of different manufacture process on it. Liu et al. [3] controlled the addition of granulated zirconia micro-powder or hollow zirconia microbeads with different mass fractions, and then investigated their effects on the room-temperature performance and high-temperature performance of MgO-Al-C materials.

However, these three methods not only require high levels of expertise and experimental equipment from researchers and readers, but usually proceed in the form of controlling variables, such as varying the temperature of a certain process, then, exploring the influence of different temperatures on the microstructure of the material and the evolution of the material performace, which to some extents restricts the comprehensive understanding and description of the material performance evolution.

2.2 Entity extraction

In recent years, the accuracy of entity extraction tasks in common fields has reached relatively high levels both at home and abroad. Such as: Juae et al. [8] proposed two methods for weakly labeled data generation, which can extract named entities from social media texts more effectively; Chen et al. [9] proposed a BERT model combined with entity vectors for knowledge graph entity extraction, which has good results in entity extraction in CoNLL-2003 corpus.

The accuracy of general-purpose domain entity extraction tasks has now reached a relatively high level. Therefore, more and more scholars begin to turn their attention to entity extraction research in specialized domains. For example, Ren et al [10] proposed a fishery standard named entity extraction model, which integrates the attention mechanism and BiLSTM+CRF, utilize contextual structural features effectively, and have better recognition performance for fishery standard named entity extraction; Cai et al [11] proposed a named entity extraction method for COVID-19 clinical text based on MPNet and BiLSTM, and the model has better recognition effect for COVID-19 clinical trial registration entities.

However, the current entity extraction studies in the field of materials are relatively small, and since material patents contain a great amount of information about technical terms, material compositions, and complex manufacture process. If we simply apply or stack the models, it may lead to inaccurate or incomplete results of entity extraction [12].

2.3 Evolutionary analysis methods

In terms of evolution analysis, researchers usually use the following methods:(1) topic identification; (2) topic evolutionary analysis [13, 14]. Among them, topic evolution analysis is divided into topic intensity evolution and topic content evolution [15].

Regarding topic identification methods, the predominant approach is the Latent Dirichlet Allocation (LDA) topic model and its extended versions [16]. However, it is difficult for such methods to effectively reveal document semantics, as well as to distinguish the meaning of polysemous words in different contexts, leading to the cluster of words or phrases within the same topic lacks semantic relevance, and resulting in poor interpretability of the topics [17].

Furthermore, topic intensity evolution analysis [18] only reflects the relative importance or attention of a topic at a specific time point. It can't directly reflect the overall trend of topic evolution over time. Topic content evolution [19] mainly relies on similarity measures, which believes that there is an evolution relationship between topics whose similarity higher than a threshold , but the similarity metric is highly sensitive to data,and tends to ignore the semantic relationships between topics. This leads to biased analysis results during topic evolution.

3.Data and method

This paper takes the material performance as the research object. Firstly, we integrate syntactic dependency analysis and attention mechanism to construct the material entity extraction model (BERT-BiLSTM-CRF), we obtaining the performance nodes of each material. And then divided the performance nodes of all materials into time batches by year. After that, we designed an algorithm based on the initial performance topics, to realize the dynamic accumulation of material performance topics and the construction of evolution path.

3.1 Data sources and preprocessing

We considered that to do material evolution, if we collect data at random intervals, it may lead to a lack of completeness and accuracy in the final analysis of the evolution results.So we takes the concept of Germany's "Industrie 4.0" as the background, and selects metal material as an example, which is one of the key foundational materials closely related to this concept. Then we use the Derwent Innovations Index database as the patent data retrieval platform and the data is retrieved on December 2023. The patent search expression is "TS=('Metal materials' OR 'Metallic materials' OR 'Metal alloys' OR 'Metal compositions' OR 'Metal-based materials') AND WC=('Materials Science')", with a time interval from 2011 to 2023, where 2011 is the year when the concept of "Industry 4.0" was first introduced. Then, the search results were ranked in terms of relevance, and we select the top 10,000 patent texts among them as the dataset. Moreover, considering the number of patents in each period, we divide it into different time batches by year for material performance evolution.

3.2 Method for extracting the material entities

In this part, we first define two types of entity and the relationship between them, and then, we integrate syntactic dependency analysis and attention mechanism to construct an entity extraction model (BERT-BiLSTM-CRF), realizing the extraction of the material entities.

3.2.1 Definition of entity and relationship

Under the background of the continuous improvement of the manufacture process, the processing method of materials is constantly progressing. Slight changes in the manufacture process can bring about corresponding changes in the internal structure of materials, which in turn causes changes of material performance [20]. That is, the change in the manufacture process of the material will bring about changes in the material performance.

Therefore, we defined the performance entity and manufacture process entity of mental material. And then, we analyze the content of mental material patents, finding that the ADVANTAGE part and the NOVELTY part in the abstract are mainly used to describe the performance and manufacture process of that material, so we use them as the basis for extracting the performance entities and the manufacture process entities of each material, and then merged them into performance node and manufacture process node of each material. In addition, we refer to the relationship shown in Fig. 1, establishing the causal relationship between the two for subsequent analysis, as shown in Fig. 2. (Among them, the manufacture process entity and causal relationship will be used in our next step of exploring the reasons for performance evolution, so it is rarely involved in this paper.)

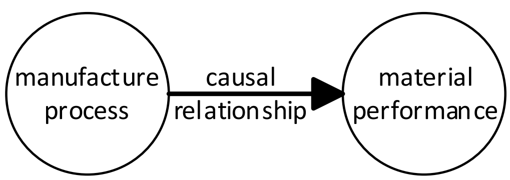


Figure 2: Definition of Entity And Relationship

3.2.2 Construction of material entity extraction model

Since the content of material patents contains a large number of technical terms, material components, we constructed an entity extraction model (BERT-BiLSTM-CRF) by combining syntactic dependency analysis and attention mechanism , which realizes the extraction of the material performance entities.

The model first uses syntactic dependency analysis to analyze the dependency relationships and the syntactic links between words in a sentence, which can help better understand the semantic structure of the sentence and identify the information about material performance and manufacture process. Secondly, by introducing attention mechanism, the weights of extraction are automatically adjusted according to the importance of the context, which can better capture the contexts that contain the information about the key entities of the material performance and manufacture process. The combined use of these methods provides a more accurate extraction of the material performance entities and manufacture process entities from patent contents, providing a basis for subsequent material analysis and research.

3.3 Method for constructing dynamic evolution path for material performance topics

In this part, we first define six evolution types, then we designed an algorithm for identifying the evolution relationship of performance nodes based on ring boundaries. Finally, we present the detailed process of the method for constructing dynamic evolutionary path of material performance topics.

3.3.1 Social Circle Theory

Social circle theory suggests that the social circle formed around a person reflects the closeness of his or her social relationships. In other words, a person's core social circle, i.e., the intimate friends circle, usually consists of relationships with a high degree of relevance; and from there, the normal friends circle and the strangers circle in that order. In addition, considering that there may exist such a part of people in the sea of people: they are temporarily outside your normal friends circle, but there are certain similarities between each other, and they may become your friends or even intimate friends in the future, so this paper defines those who are outside of the normal friends circle and within the strangers circle as potential friends. Therefore, centered on the individual, their affinity rank order is: intimate friends, normal friends, potential friends, strangers, and the position belongs to: within the intimate friends circle, outside the intimate friends circle within the normal friends circle, outside strangers circle within the normal friends circle, outside the strangers circle, specifically as shown in Figure 3.

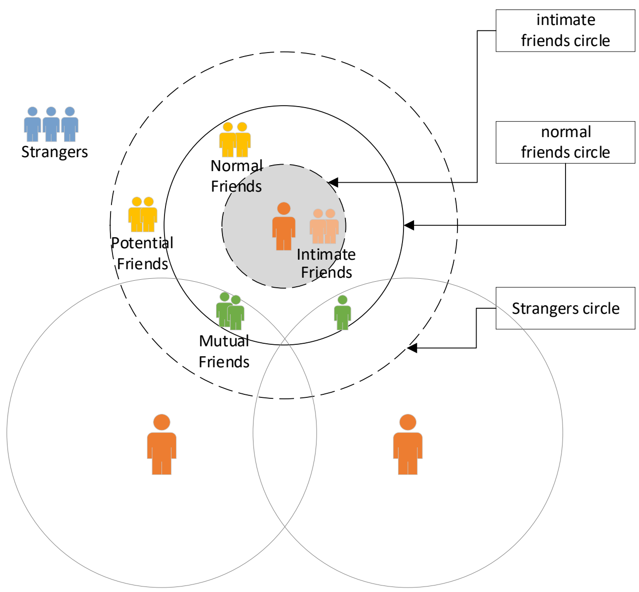


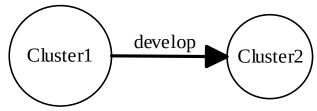
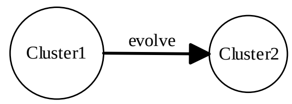
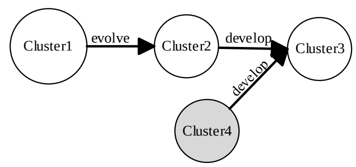
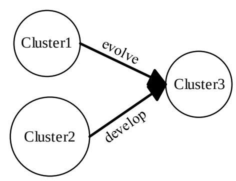
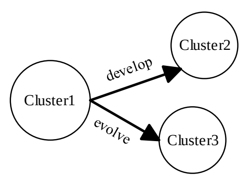
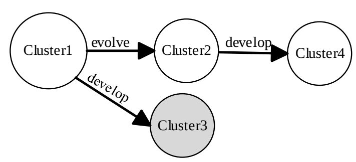
Figure 3: Social Circle Theory

We refer to this theory and combine it with existing research to define six evolution types. And propose an algorithm for identifying the evolution relationship of performance nodes, the details are shown in parts 3.3.2 and 3.3.3 respectively.

3.3.2 Definition of evolution types

This paper defines six evolution types based on existing studies. Among them, the four types of develop, evolve, emerge, and fuse are derived from four different social relationships in the social circle theory, and the two types of extinct and split refer to the existing studies to ensure the diversity of evolution types. In addition, this paper also improves the fuse and split types, by further refining the different contributions of each theme in them, which helps to consider the dynamic interactions between themes in more detail. The details are shown in Table 1.

Table 1
Six patterns of topic evolution types (where Cluster stands for topic)

<table>
<tr><th>social relationships</th><th>Evolution types</th><th>Expression and illustration</th><th>Explanation</th></tr>
<tr><td>intimate friends</td><td>develop</td><td>
</td><td>Cluster2 only develops from Cluster1</td></tr>
<tr><td>normal friends</td><td>evolve</td><td>
</td><td>Cluster2 only evolves from Cluster1</td></tr>
<tr><td>strangers</td><td>emerge</td><td>
</td><td>Cluster4 is treated as an emerge Cluster which has no evolutionary relationship with previous Clusters.</td></tr>
<tr><td>mutual friends</td><td>fuse</td><td>
</td><td>cluster3 comes from cluster1 and cluster2 together.</td></tr>
<tr><td>/</td><td>split</td><td>
</td><td>Cluster1 is divided two or more Clusters</td></tr>
<tr><td>/</td><td>extinct</td><td>
</td><td>Cluster3 is extinct which has no evolution relationship with the following Clusters</td></tr>
</table>

3.3.3 Identifying the evolution relationship of performance nodes based on ring boundaries

We refer to social circle theory and improve the model proposed by Zhang et al. [21], proposing algorithm for identifying the evolution relationship of performance nodes based on ring boundaries, the specific algorithm and its correspondence are shown in Figure 4. Firstly, for the existing performance topics, the centroid of each topic is calculated, and the maximum Euclidean distance between each topic's patent and its centroid is taken as the topic boundary. The topic boundary is extended outward by a ratio less than 1 to obtain the outer ring boundary and shrunk inward by the same ratio to obtain the inner ring boundary. After several comparison tests, we finally set the ratio in this study to 0.2.

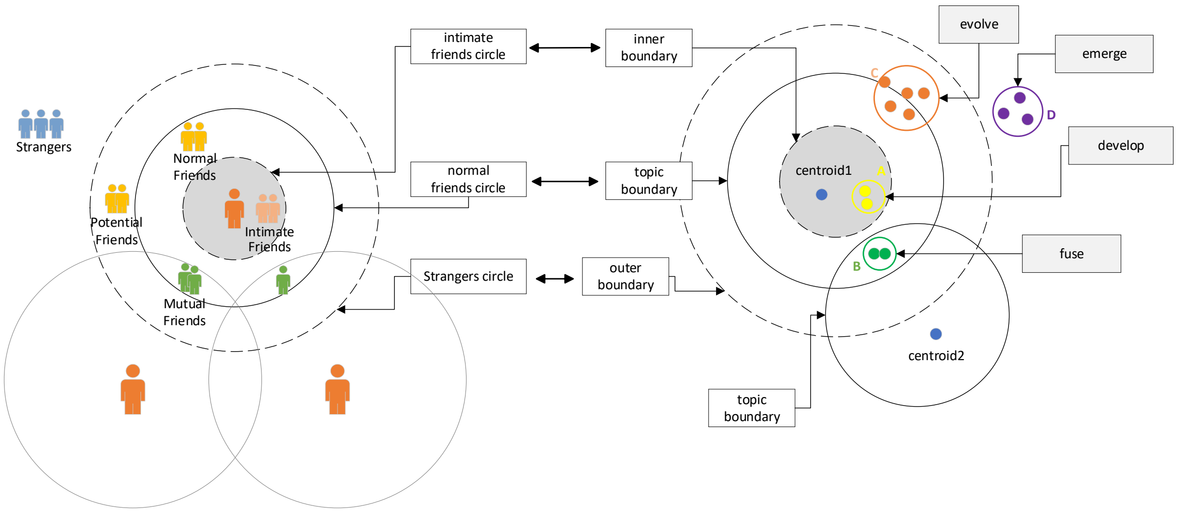


Figure 4: Algorithm for Identifying the Evolution Relationship of Performance Nodes based on Ring Boundaries (See Appendix A for the Entire Picture)

Subsequently, for material performance nodes in subsequent batches, the Euclidean distance between each performance node and the centroids of existing performance topics is calculated separately to identity the evolution relationship between them. The specific process is as follows: If the performance node does not exceed the inner ring boundary of that performance topic, it is labeled as a "strong tie," indicating that the performance node has a develop relationship with the performance topic. If it exceeds the inner ring boundary but not the topic boundary, it is checked if the performance node is within the overlapping boundary region of the performance topic. If so, it is labeled as a fuse relationship; if not, it is checked if the performance node exceeds the outer ring boundary. If not, it is labeled as an evolve node; if yes, it is labeled as a "weak tie," indicating that the node is an emerge node. The specific positional relationships are shown in Table 2.

Table 2
Specific position for each evolution type

Evolution types	Position
develop	inside the inner ring boundary
evolve	outside the inner ring boundary and inside the outer ring boundary
emerge	outside the outer ring boundary
fuse	boundary intersection

Then, hierarchical clustering is introduced to obtain the different types of performance topics in the batch, and merge similar topics that exceed a threshold (we set it to 0.8 in this paper). For a topic in the previous batch, if its number of topics obtained more than two after clustering in this batch, the evolution type is considered as split. Furthermore, in the construction of the evolutionary path, if a topic has no evolution relationship with the following topics, we consider it as extinct type.

3.3.4 Construction of the dynamic evolution path for metal material performance topics

The construction of the dynamic evolution path of metal material performance mainly includes the following steps, which are shown in Figure 5. Firstly, after the extraction of performance entities of each material, we get the performance node of each material, and then, all material performance nodes are divided into time batches according to the year. Secondly, the K-Means algorithm is used to cluster the first batch of data to obtain the initial performance topics.

Subsequently, for performance nodes in subsequent batches, the algorithm for identifying the evolution relationship of performance nodes (see Section 3.3.3 for the specific algorithmic process) is used to identify their evolution relationships with each performance topic. Then, hierarchical clustering is introduced to obtain the performance topics of different evolution types in this batch, and merge similar topics that exceed a threshold. Finally, incremental iterations are carried out in the above manner to obtain the material performance topics at different year batches, thereby achieving the dynamic construction of the material performance evolution path.

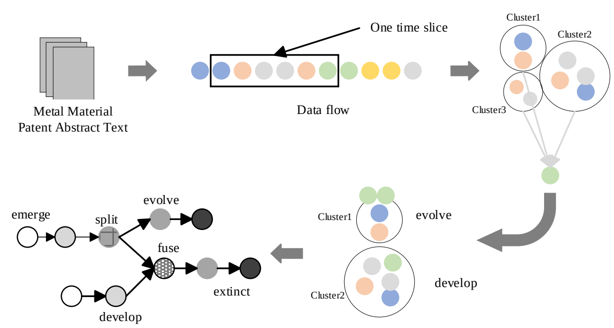


Figure 5: Process of Constructing The Evolution Path of Metal Material Performance

4.Result

In the construction of the evolution path of metal material performance, we use examples from the years 2021-2023 to obtain the results. Specifically, there were 66 performance Clusters in 2021, 55 performance Clusters in 2022, and 60 performance Clusters in 2023. The result of the 2021-2023 evolution path are shown in Figure 6, where yellow represents the year 2021, red represents the year 2022, and green represents the year 2023. (see Appendix B for the entire picture, where Cluster stands for topic)

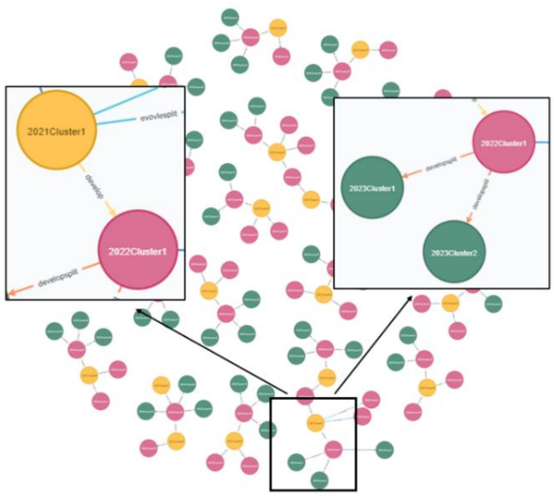


Figure 6: Evolution Path of Metal Material Performance from 2021 to 2023

From the results of the evolution path above, it can be observed that in 2022, Cluster1 developed from the Cluster1 in 2021. The contents of the two Clusters are as follows: ['high alloy excellent low corrosion', 'resistance good powder strength temperature'], ['high excellent alloy low mechanical', 'strength corrosion resistance process good']. It is not difficult to find that both Clusters focus on improving the corrosion resistance and mechanical performance of metal materials, which aligns with the practical application requirements of metal materials [22].

In addition, in 2022, Cluster1 further developed and split into Cluster1 and Cluster2 in 2023. The contents of the three Clusters are as follows: ['high alloy excellent low corrosion', 'resistance good powder strength temperature'], ['good surface heat', 'resistance layer wear low good'], and ['resistance good base layer wear', 'layer wear surface low heat']. As can be seen, the 2023 Cluster1 maintained the original corrosion resistance and good wear resistance performance of Cluster1 in 2022, and further improved the surface heat performance of metals. This may be achieved through improvements in material technology and alloy additions, thus enhancing the performance in practical applications.

5. Conclusion

Regarding current issues in topic evolution, such as poor topic interpretability and neglect of semantic relationship among topics, this paper proposes an algorithm for identifying the evolution relationship of performance nodes based on ring boundaries, which can not only realize the dynamic accumulation and construction of metal material performance evolution path, but enrich and improve the topic evolutionary analysis method. Currently, we are combining the manufacture process entities of each material to further analyze the causes of the evolution of material performance in depth, and to better understand the evolution trends and the changing patterns of material performance.

Acknowledgments

This work is supported by the National Natural Science Foundation of China (No. 72374103, 71974095) and the Postgraduate

Research Practice Innovation Program of Jiangsu Province (No. KYCX23_0632).

References

[1] Kaushal Jha, Suman Neogy, Santosh Kumar, R N Singh, & G K Dey. Correlation between microstructure and mechanical properties in the age-hardenable Cu-Cr-Zr alloy. Journal of Nuclear Materials: Materials Aspects of Fission and Fusion, 2021, 546, 152775. doi: 10.1016/j.jnucmat.2020.152775.

[2] Enyu Guo, Guohua Fan, & Tongmin Wang. Microstructure Evolution Mechanism of Metallic Materials: Progress on In Situ Studies Using Synchrotron Radiation Source. Failure Analysis and Prevention, 2021, 16(01): 1-14+91. doi: 10.3969/j.issn.1673-6214.2021.01.001.

[3] Yang Liu, Baoliang Liu, Shi Kai, Xiaoyuan Han, Yi Xia, & Jianzhao Shang. Effects of different zirconia-based raw materials on properties of MgO-Al-C materials. REFRACTORIES, 2023, 57(01): 59-64. doi: 10.3969/j.issn.1001-1935.2023.01.013.

[4] Jinxia Liu, Qianqian Hou, Jing Du, Fuhou Chai, & Li Zhang. Hot Topics Evolution Research of Emerging Field Under the Sub Topics and Vocabulary Related Perspective. Journal of Intelligence. 2023,42(3):123-129. doi:10.3969/j.issn.1002-1965.2023.03.017.

[5] Xiaoyan Song, Kai Guo, Hao Lu, Dong Liu, & Fawei Tang. Integrating computational materials science and materials informatics for the modeling of phase stability. Journal of Materials Informatics. 2021; 1(1): 7. doi: 10.20517/jmi.2021.06.

[6] ZhiGuo Liu, Ri You, & Jintao Wu. Study on relationship between dam concrete material performance evolution and its pore structure. WATER RESOURCES AND HYDROPOWER ENGINEERING, 2016, 47(11): 25-28+35.

[7] Yinzhi Li, Gang Liu, Xinzhe Chen, Zhenhua Su, Lianwu Yan, & Yingbiao Peng. Effects of Sintering Temperature and Holding Time on the Microstructure and Properties of Ti(C,N)-Based Cermets[J]. Packaging Journal, 2023, 15(03): 37-45.

[8] Juae K, Yejin K, & Sangwoo K. Weakly labeled data augmentation for social media named entity recognition. Expert Systems with Applications 2022, 209:1188217. doi: 10.1016/j.eswa.2022.118217

[9] Wei Chen, Rui Zhang, & Zhong Yi. Knowledge Graph Entity Extraction Method Based on BERT Model and Entity Vector. Journal of Chinese Computer Systems. 2022,43(8):1577-1582. doi: 10.20009/j.cnki.21-1106/TP.2021-0073.

[10] Ren Yuan, Yu Hong, Yang He, Liu Jusheng, Yang Huining, Sun Zhetao, Zhang Sijia, Liu Mingjian, &Sun Hua. Recognition of quantitative indicator of fishery standard using attention mechanism and the BERT+BiLSTM+CRF model[J]. Transactions of the Chinese Society of Agricultural Engineering (Transactions of the CSAE). 2021, 37(10): 135-141. doi: 10.11975/j.issn.1002-6819.2021.10.016

[11] XiaoQiong Cai, ZengLiang Zheng, QianMin Su& JinLei Guo. Named entity recognition of COVID-19 clinical text based on MPNet and BiLSTM. Intelligent Computer and Applications. 2023, 13(01): 164-170+177. doi: 10.3969/j.issn.2095-2163.2023.01.028.

[12] Anzhi Yuan, Furong Peng, Bo Gu, & Yuhua Qian. Technology Entity Extraction of Patent Literature with Limited Annotated Data. Journal of Zhengzhou University(Natural Science Edition). 2021, 53(4): 61-68. doi: 10.13705/j.issn.1671-6841.2021196

[13] Parlina A, Ramli K, & Murfi H. Theme Mapping and Bibliometrics Analysis of One Decade of Big Data Research in the Scopus Database. Information, 2020, 11(2): 69. doi: 10.3390/info11020069.

[14] Zibiao Li, & Li Zhang. Evolution of Patented Technology Topics of Steel Materials Based on LDA Model. SCIENCE AND TECHNOLOGY MANAGEMENT RESEARCH, 2020, 40(24): 175-183. doi: 10.3969/j.issn.1000-7695.2020.24.023.

[15] Geller, M., Vasconcelos, V.V., & Pinheiro, F.L. Toxicity in Evolving Twitter Topics. Lecture Notes in Computer Science, 2023, 10476,40-54. doi: 10.1007/978-3-031-36027-5_4.

[16] Lee J W, & Han D H. Data Analysis of Psychological Approaches to Soccer Research: Using LDA Topic Modeling. Behavioral Sciences, 2023,13(10), 43-47. doi: 10.3390/bs13100787.

[17] Behrouzi S, Sarmoor Z S, Hajsadeghi K, & Kavousi K. Predicting scientific research trends based on link prediction in keyword networks. Journal of Informetrics, 2020, 14(4): 101079. doi: 10.1016/j.joi.2020.101079.

[18] Wu Z, Xie P, Zhang J, Zhan B, & He Q. Tracing the Trends of General Construction and Demolition Waste Research Using LDA Modeling Combined With Topic Intensity. Frontiers in Public Health, 2022, 10(3), 22-26. doi: 10.3389/fpubh.2022.899705.

[19] Reza T H S, Sadegh A, & Soroush T. An embedding approach for analyzing the evolution of research topics with a case study on computer science subdomains. Scientometrics, 2023,128(3): 11-16. doi: 10.1007/s11192-023-04642-4.

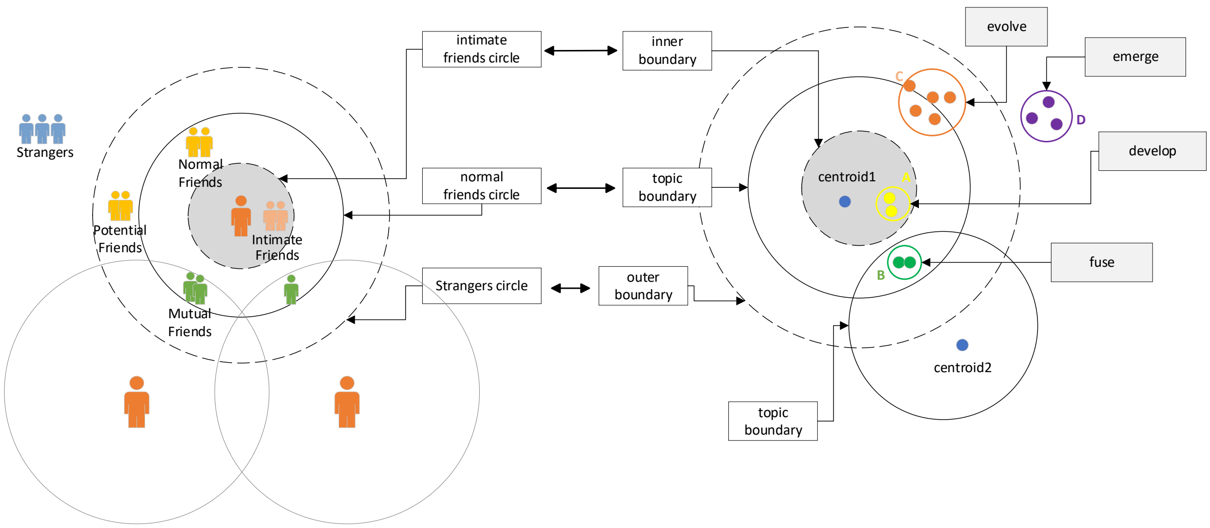
[20] Qing-Miao Hu, & Rui Yang. The endless search for better alloys. Science, 2022, 378: 26-27. doi:10.1126/science.ade5503

[21] Zhang Y, Zhang G, Zhu D, & Liu J. Scientific evolutionary pathways: Identifying and visualizing relationships for scientific topics. Journal of the Association for Information Science and Technology , 2017, 68(8): 1925-1939. doi: 10.1002/asi.23814.

[22] Y.B. Lei, Z.B. Wang, B. Zhang, Z.P. Luo, J. Lu, & K. Lu. Enhanced mechanical properties and corrosion resistance of 316L stainless steel by pre-forming a gradient nanostructured surface layer and annealing. Acta Materialia. 2021, 208, 116773. doi: 10.1016/j.actamat.2021.116773.

Appendices A

Algorithm for Identifying the Evolution Relationship of Performance Nodes based on Ring Boundaries



Appendices B

Evolution path of metal material performance from 2021 to 2023(See annex 1 for details)