An Entity-Based Main Path Analysis Method to Trace Knowledge Evolution at Micro-Level

Chi Yu[1] Weijiao Shang[2]

Xiaozhao Xing[1] Haiyun Xu[3]

Liang Chen[1]*

[1] Institute of Scientific & Technical Information of China, Beijing 100038, P.R. China

[2] Research Institute of Forestry Policy and Information, Chinese Academy of Forestry, Beijing 100091, P.R. China

[3] Business School, Shandong University of Technology, Zibo 255000, P.R. China
E-mail:25565853@qq.com(Liang Chen)
* Corresponding author

Abstract. Main Path Analysis (MAP) method is a significant method for knowledge flow extraction from citation networks. Traditional MPA methods treat documents as network vertices, while neglecting the more granular information within the document, this neglect limits an in-depth understanding of knowledge development. To remedy the weakness, this study leverages deep learning algorithm on MPA method to facilitate an entity-based pathfinding method, thus to improve the interpretability of MPA method. This study introduces a four-step process to implement the proposed method: (1) Data preprocessing to structure the citation network for analysis. (2) Knowledge entity extraction using BERT-BiLSTM-CRF for identifying significant entities. (3) Main path search at the document level with a cluster-based approach for path identification. (4) Entity relationship identification across documents using a BERT-based model with a three-level masking strategy. This study aims to transform literature-based citation networks into detailed entity-based networks, enabling finer-grained knowledge flow extraction. Finally, to demonstrate the advantages of the new method, extensive experiments are conducted on a patent dataset pertaining to thin film head in computer hardware. Experimental results show that our method is capable of discovering more fine-grained knowledge flows from important sub-fields, and improving the interpretability of candidate paths as well.

Keywords: Main Path Analysis, Patent mining, Entity Extraction, Hard Disk Heads.

1 Introduction

With the rapid growth of scientific and technical documents nearby, it is increasingly crucial to extract valuable information from the overload data efficiently, effectively and archiving in-depth insights thereafter. In case of scientific research and decision-making support, one of such valuable information is the knowledge flow which reflects the development trend of the focal domain. Earlier methods mainly dependent on man-

C. Bai et al. (eds.), *Proceedings of 2023 China Science and Technology Information Resource Management and Service Annual Conference (COINFO2023)*, Advances in Economics, Business and Management Research 293,
https://doi.org/10.2991/978-94-6463-498-3_10

ual reading of numerous documents to serve the purpose. With the advances of information technology and the establishment of large-scale digital libraries, increasing researchers have begun to archive knowledge flow in a semi-automatic or automatic manner. For example, Wang et al. [1-2] facilitated technical roadmap by using information extraction method on patent texts, Nallapati et al. [3] utilized network clustering algorithm to illustrate knowledge flow in citation network. Among them, main path analysis method proposed by Hummon and Dorerian [4] have received a lot of attention as a quick way to acquire evolution path from citation network, which can help to achieve more effective intelligence in scientific and technical information.

But the evolution path consists of milestone documents still suffer from information overload, researchers have to read the texts to retrieve a detailed and comprehensive understanding of evolution path. To remedy this weakness, some researchers [1-2,5-6] have turned to information extraction technologies or topic models to generate fine-grained evolution path, such as keyword-based path or topic-based path. The advances of information extraction technologies have significantly increased the performance of NER (Named Entity Recognition) task, and opened up new possibilities of fine-grained evolution path study [7-9]. By integrating NER and bibliometrics, Ding et al. [10] introduced the concept of entitymetrics, which aims to using entities in the measurement of impact, knowledge usage, and knowledge transfer to facilitate knowledge discovery. But the assumption made by Ding et al. that if one paper cites another paper, then an entity in the citing paper will be considered to cite an entity in the cited paper, is problematic. In fact, when a paper cites another paper, it does not necessarily imply that all entities of the former share a citation link with all entities of the latter. Furthermore, due to the combinatorial explosion problem caused by Ding's assumption, the scale of the resulted entity citation network is dramatically increased, accompanied by a substantial amount of link noise.

To address the above issues, this study proposes a novel main path analysis method to extract fine-grained evolution path from citation network. The new path connects knowledge entities with citation links to demonstrate knowledge flow in a fine-grained perspective. As for the combinatorial explosion problem aforementioned, a new multi-head attention mechanism is proposed to identify entity links across different documents, thus to reduce entity links significantly. By integrating entity links with main path analysis method, a detailed representation of knowledge flow can be achieved through entity main paths.

2 Related work

As a time sequence diagram to reflect the historical evolution and knowledge diffusion in context of science [11], citation network has been widely used to trace the developmental trajectory in a research field since 1964 [12]. The related studies can be divided into two broad groups: those measuring the importance of documents within citation network, and those analyzing the structure of citation networks. A representative method of the former is to identify network among most-cited documents on the basis of citation frequencies [12], namely critical path analysis [13]. Aside from documents,

the latter also concerns the citation links from their patterns to the structure of network which were neglected before. One of the most significant works in this group is MPA which considers both the citations a document receives and the documents it cites [4], thus to reflect the knowledge flow more comprehensively since those documents contained in the flow not only build on prior publications but continue to act as an authority in reference to later works [14]. Due to the advantages, studies related to MPA have increased steadily in recent years and have aggregated into a tightly linked strand of the literature [15].

The basic idea of MPA is that different links were not equivalent in citation network, important links serve as main pathway and their removal will alter the entire process of knowledge dissemination. Conversely, the knowledge flow conveyed by the insignificant links is considerably less and exerts a less impact on the citation network. Therefore, the main pathway contains critical links in knowledge dissemination and represent the skeleton of the citation network. Hummon et al. referred to the main pathway as main path and proposed a pipeline method, namely main path analysis (MPA) for its extraction. The basic procedure of MPA starts by assigning traversal weight to each link in the citation network, then emanates at each source vertex or sink vertex to search a citation sequence (path) with local or global optimal strategy adapted to link weight. By source vertex, we mean the vertex that is cited while referring to no other vertices and the sink vertex is the reverse. The last step of MPA is simply selecting the sequence that is the highest in sum of link weights or multiple sequences meeting certain criteria as the final main path(s). For illustration, the procedure is summarized as in Fig.1.

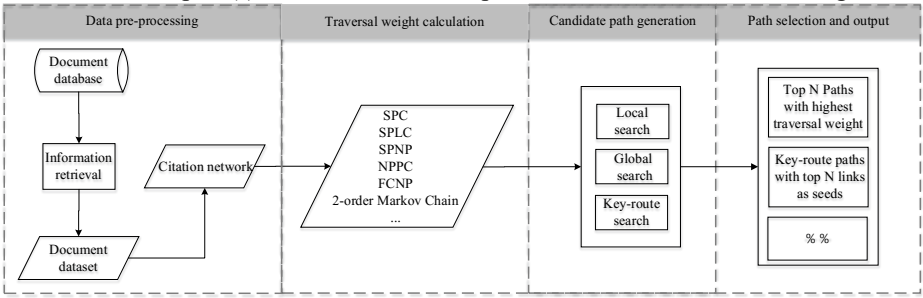


Fig. 1 The basic procedure of MPA

Aside from the general process mentioned above, there are also some studies combining MPAs with other methods to provide more insight into the development of a discipline. Huang et al [16] identified different stages of technology life circle by curving fitting and compared the main paths in different stages to explore technology evolution pathways. Along the direction, they further added co-classification and co-word analysis into MPA to supplement the technical evolution process and extract implicit or unknown pattern and topics [17]. As to the weak interpretability problem of main path caused by the professionalism and complication of the target discipline, Tu & Hsu [18] took advantage of text-mining techniques to add keywords to each vertex on the main path as labels, whereas Kim & Shin [19] leveraged structured experts' judgments

to identify product component-patent linkages, thus to organize knowledge flow in a more structured way to improve its interpretability.

By allowing researchers to concentrate on fewer documents, MPA helps to overcome the issue of information overload and facilitate the acquisition of valuable intelligence efficiently. Especially, the implementation of original MPA method and its variants by social network software Pajek has significantly advanced its development and widespread application. But in the meanwhile, one can see that MPA method still suffers from information overload. The text attached to each vertex of the main path is usually the full text or abstract of the document, which impedes researchers to retrieve a detailed understanding of evolution path directly. For example, the abstract of a paper typically encompasses information like purpose/significance, method/procedure, as well as result/conclusion. Such texts can't facilitate main path to analyze knowledge flow in-depth and unveil the evolutionary mechanism of knowledge entity during knowledge dissemination.

3 Methodology

This study aims to identify the main paths in focal field at a micro-level, thus to receive an in-depth understanding of the mechanism underlying knowledge dissemination. To fulfill this target, a framework combining deep learning algorithms and MPA method is proposed, where deep learning algorithms targets at discerning entities in one documents and interrelations between entities across different documents, and the MPA method is to identify the critical links in citation network. The whole procedure of the methodology is shown in Fig. 2, which consists of four phases: data pre-processing, knowledge entity extraction, main path searching at document-level, entity relationship identification between documents. These modules will be detailed in the following subsections.

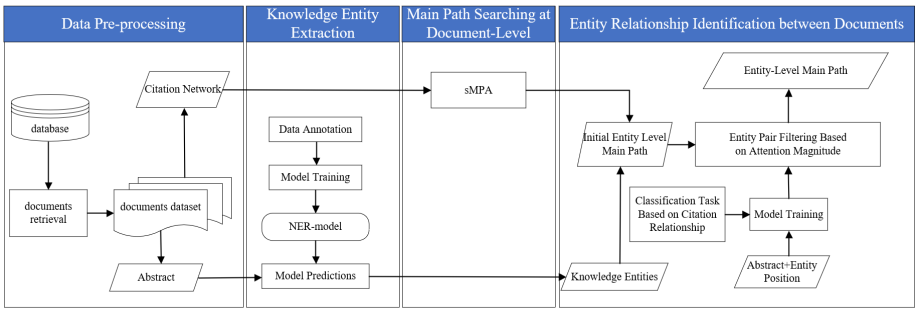


Fig. 2 The Procedure of the Entity-Based MPA Method

3.1 Data Preprocessing

Data preprocessing is the initial data analysis process**,** which comprises a set of methods and functions that clean, organize, and transform raw data into a structured format appropriate for further analysis, modeling, and machine learning[20]. In this study, data

pre-processing refers to a pipeline including obtaining search results from data source, excluding irrelevant and abnormal data, and constructing a citation network from which to take the giant connected component. With the giant connected component and the text attached to each vertex in it, the experimental dataset is achieved.

Since the target network must be directed acyclic graph(DAG) in MPA methods, it is necessary to remove the strongly connected subgraphs that might appear during citation network construction. So far a series of approaches, such as Kosaraju's, Tarjan's or Gabow's algorithm can be utilized to fulfill this purpose. Afterward, these strongly connected subgraphs are shrunk into single vertices in citation network. As for the internal links within the strongly connected subgraphs, they are eliminated while the links between the strongly connected subgraph and external vertices are retained.

3.2 Knowledge Entity Extraction

Entity extraction, also known as named entity recognition, seeks to locate and classify named entity mentions in unstructured text into pre-defined categories. As one of the most fundamental tasks in natural language processing (NLP), entity extraction is widely used to identifying the key information in general texts, such as person, organization, location, time etc. But when it comes to bibliometrics, entity is defined as the individual knowledge units in scientific/technical documents like papers and patents [10], which deviates significantly from its counterpart in NLP. Furthermore, such entity is categorized into macro-level entity (e.g., author, journal, article), mesa-level entity (e.g., keyword) and micro-level entity (e.g., dataset, method, domain entities) to fertilize the analysis of knowledge discovery, among which the micro-entity, namely knowledge entity is closest to the entity in NLP.

However, scientific/technical documents are domain-specific texts, i.e. entity types in certain domain are specialized on that subject area and not interchangeable across domains. Taking the biomedical domain for example, the typical entity types are usually compound, gene, drug, disease, etc., while in integrated circuit domain the counterparts are component, material, energy flow, measurement and so on. Given that entity types across different domains are not interchangeable, the requirement for data annotation still remains when utilizing current state-of-the-art (SOTA) methods, namely supervised deep learning methods, to conduct knowledge entity extraction. In this regard, a two-step procedure including data annotation and model training is utilized to serve the purpose. Data annotation refers to manual identification of knowledge entities from patent texts, which will not be elaborated here. As for model training, this paper uses BERT-BiLSTM-CRF, one of the state-of-the-art deep neural network models [21] in our framework. This model takes sentences as input and represents every token with its position and segment information as one vector. During training procedure these vectors pass through the layers within BERT-BiLSTM-CRF and output the predicted label for each token in the sentence. With the help of back propagation algorithm, the predicted labels will approximate the true labels and finally enable BERT-BiLSTM-CRF to recognize knowledge entities in new sentences.

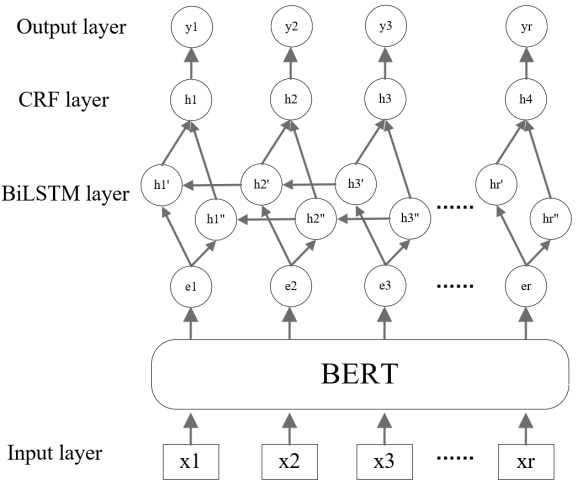


Fig. 3 Architecture of the BERT-BiLSTM-CRF Model for Entity Extraction

3.3 Main Path Search at Document Level

This procedure starts by assigning weight to each link in the citation network, then obtains all the source/sink vertices, and for each source/sink vertex, searches candidate paths with the greedy strategy. By source vertex, we mean the vertex that is cited while referring to no other vertices and the sink vertex is the reverse. The benefit of greedy search strategy is that its time complexity is $O(n)$, where n denote the number of vertices in the citation network, thus can achieves locally optimal path in limited time even for large networks. The pseudo-code of candidate paths search algorithm is shown as follows.

Input: A citation network G and a source vertex s of G.
Output: The path led by s with locally highest sum in link weights.

```
(1)  The procedure greedy-path-search(G, s)
(2)      let P be a list with s in it
(3)      while G has unvisited vertices do
(4)          neighbor_vertices = get_neighbors(G, s)
(5)          If neighbors in empty then
(6)              Break
(7)          end if
(8)          max_weight_neighbor = max(neightbors, link_weights)
(9)          P.append(max_weight_neighbor)
(10)         s = max_weight_neighbor
(11)     end while
(12)     return P
```

Next task is to select paths representative of significant sub-technologies overall in the research field. Previous studies mainly choose paths on basis of their traversal weight which have the limitation of uncovering all significant sub-technologies in the interested field. As a remedy, we propose a cluster-based method for path selection which consists of three steps as follow:

(1) Use Latent Semantic Index (LSI) [22] to represent each document in the citation network as a vector.
(2) For each candidate path, the vectors of all documents on it are added up element-wise and normalized to represent the path.
(3) Cluster all candidate paths in their normalized vector representations and choose the path with the highest sum in link weights in each cluster as output.

Due to its significance, the last step will be described in more detail. Here cosine distance D_c between any two candidate paths, say, u and v is computed by Formula (4):

$$D_c(u, v) = 1 - \frac{U \cdot V}{\|U\| \, \|V\|} \quad (4)$$

Where U, V is the vector representations of path u and v produced by step two, $\frac{U \cdot V}{\|U\| \, \|V\|}$ is the cosine similarity between U and V and $\|\cdot\|$ means Euclidean norm of a vector. Thereafter, a classic algorithm named density peak clustering[23] is used for path clustering. This is a density-based algorithm capable of automatically finding the correct number of clusters by computing two quantities of each data point i consisting of its local density ρ_i and its distance δ_i from points of higher density. As cluster centers are surrounded by neighbors with lower local density and that they are at a relatively large distance from any points with a higher local density, it's easy to identify these cluster centers by drawing a scatter plot with ρ and δ as its axes, namely decision graph.

3.4 Entity Relationship Identification Across Documents

Based on the retrieved knowledge entities and main paths in previous sections, this section utilizes entitymetrics to drill down the main paths from document-level to entity-level, as shown in Fig. 4. Entitymetrics assumes that if two documents have a citation relationship, then any entities within the two documents have a citation relationship as well. This assumption often causes the structure of the entity level main path complicated, as shown in Fig.4(c), and reduces the interpretability of the resulted paths. Therefore, it is necessary to prune the main path to highlight entity pairs across documents with citation relationships.

This study proposes a BERT-based method to serve the purpose. The basic idea is to identify entity citation relationships by leveraging the ability of self-attention mechanism in BERT in capturing internal relationships among texts. However, BERT only supports input length up to 512 tokens, which can be easily surpassed by the concatenation of citing and cited documents, and it does not distinguish between entities and non-entities in texts as well. To remedy the problems, a new model is designed in Fig. 5. Specifically:

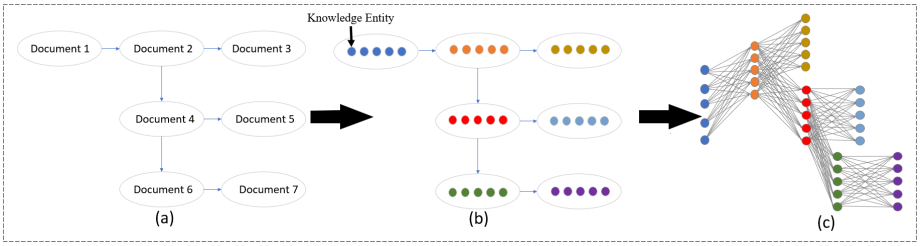


Fig.4 Entity Citation Network and Generation of Entity Main Paths

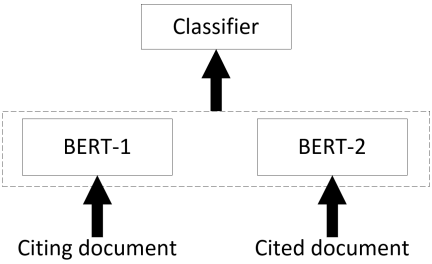


Fig.5 The Joint Model for Identifying Entity Citation Relationships

(1) By combining two BERTs, a joint model is built in which BERT-1 is for citing document and BERT-2 is for cited document, thus solving the problem of long texts when multiple documents are input into the joint model.

(2) To enable the joint model to identify entity citation relationships across documents, a three-layer masking mechanism is proposed in the joint model, as shown in Fig.6, in which the length alignment mask aims to keep the input text of each BERT in the same length, the entity attention mask is to facilitate self-attention mechanism to focus on knowledge entity while ignoring non-entity words, and the inter-document entity relationship mask confines the self-attention mechanism to entity relationships between documents while disregarding those within the same document.

(3) To further improve the performance of the joint model, a fine-tuning task, namely, classification of whether a citation relationship exists between two documents is proposed as well. Given that citation data can be directly sourced from bibliographic databases, it is feasible to generate substantial annotated data at low cost. In the meanwhile, the fine-tuning task is closely related to identification of entity citation relationships, which enables the two tasks to be mutually improved during fine-tuning and achieves better performance for identifying entity citation relationships.

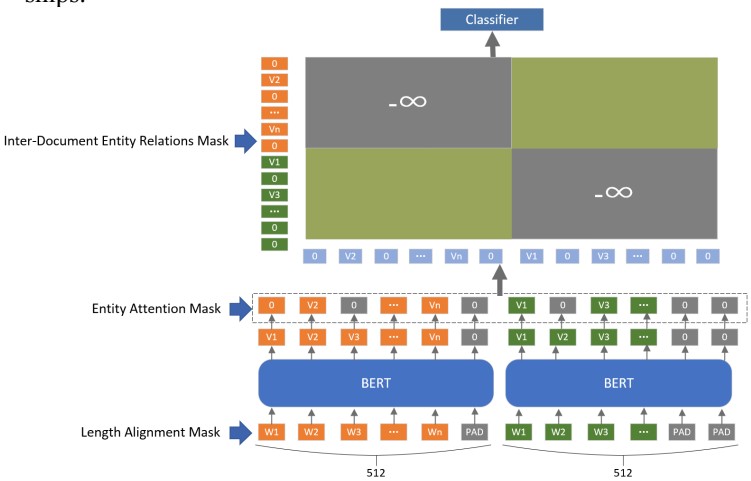


Fig. 6 Three-Level Masking Strategy to Identify Entity Citation Relationships

4 Empirical Study

4.1 Dataset Overview

To demonstrate the advantages of the new method, extensive experiments are conducted on a patent dataset related to thin-film heads in computer hardware, namely TFH-2020 dataset [24]. These patents are collected from the United States Patent and Trademark Office (USPTO) database with a search strategy combining keywords, application time and patent citations. Specifically, at the first stage 137 seed patents are retrieved with the search statement of "ABST/'thin film head' AND APD/1/1/1976->31/12/2003". Through forward and backward citation to these seed patents at the next stage, the patents dataset is extended to 2,048. After removing irrelevant patents, 1,010 patents are kept and their abstracts are used for knowledge entity and semantic relation annotation. The annotated dataset can be accessed public freely through https://github.com/awesome-patent-mining/TFH_Annotated_Dataset. But the citation network constructed by TFH-1010 dataset is abundant with small connected components and isolated vertices, which doesn't fulfill the prerequisites of the proposed method that the network should be a connected component with certain scale. Hence, the forward and backward citation method is employed once more to extend the TFH-1010 dataset. This results in a connected component comprising 3505 patents and serves as the experimental data in this study.

4.2 Knowledge Entity Extraction

The next step is to train BERT-BiLSTM-CRF model with the TFH-1010 annotated dataset, and then extract knowledge entities from other patents in the experimental data. As original BERT [25] is pre-trained on general corpus like BookCorpus and English Wikipedia, the significant distinction between general texts and patent texts will severely jeopardize its performance in knowledge entity extraction from patent texts. To address the problem, this study utilizes BERT-for-patents [26], a specialized version of BERT pre-trained on more than 100 million full-texts of patent documents, which turned out to improve the performance of knowledge entity extraction in patent texts significantly.To measure the performance of BERT-BiLSTM-CRF in the experimental data, the TFH-1010 is randomly divided into training set and testing set with a 4:1 ratio. After 40 epochs of model training, the loss curve is shown in Fig. 7, and the micro-average of precision, recall, F1-value for the test set are 86%, 85%, 86%. Fig. 8 shows precision, recall and F1-value for each type of entity denoted by its first 3 letters.

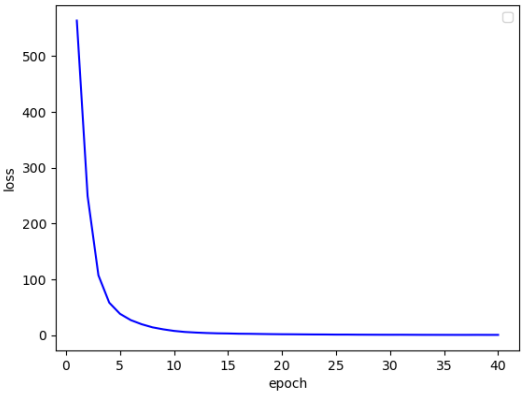


Fig. 7 Knowledge Entity Extraction Loss Curve

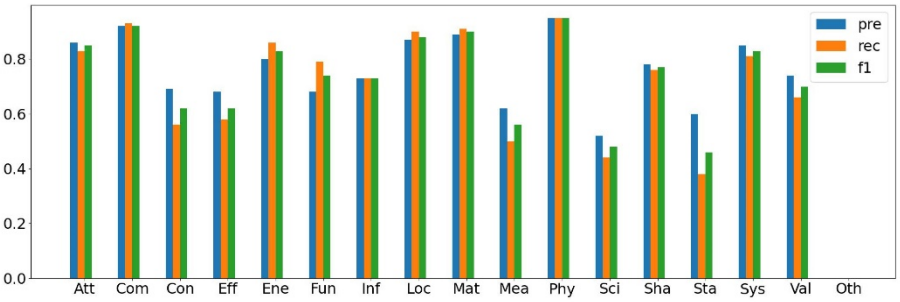
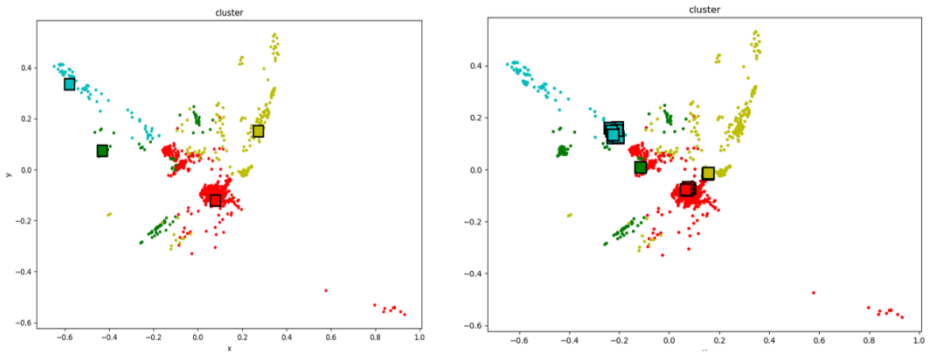


Fig. 8 Result of Knowledge Entity Extraction for different entity type

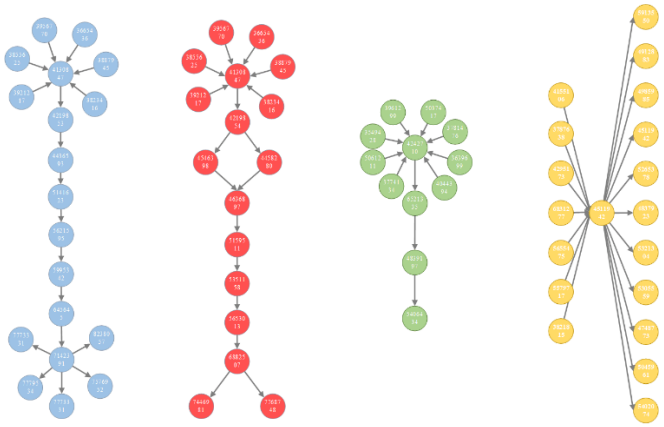
4.3 Generation of Main Path at Document-Level

With the greedy search strategy (cf. Subsection 3.3), candidate paths can be obtained from our citation network. After each document in the citation network is vectorized with the LSI model, the vectors of all documents along a candidate path are added up element-wise and normalized to represent the path. Then, the candidate paths are clustered into different sub-fields with the density peak clustering method [27]. This density-based method can automatically find the correct number of clusters [28]. Intuitively, the paths located at the cluster centers, as shown in Fig.9 (a), can best represent the topics in each cluster. However, such paths are not suitable to be main paths due to their low topological weights. To address the problem, the path with the largest topological weight in each cluster is chosen to represent the resulting cluster, as depicted in Fig. 9 (b), and the details of each main path are shown in Fig. 10.



(a) (b)

Fig. 9 The semantic distribution of candidate path



(a) path 1 (b) path 2 (c) path 3 (d) path 4

Fig. 10 The detailed information of main paths in different sub-fields

4.4 Generation of Main Paths at Entity-Level

Given that long paths are more suitable for reflecting knowledge dissemination in citation networks, path 1 and path 2 are chosen to construct the entity-level main path. However, when entities are incorporated into these main paths, the issue of combination explosion caused by entitymetrics would result in significantly complicate paths, as illustrated in Fig. 11. Hence, it is imperative to prune the main paths with the joint model (cf. subsection 3.4) to explicitly delineate the knowledge flow in micro-level. To fine-tune this joint model, a dataset comprising 9,657 samples were constructed where each sample is composed of three fields, i.e., the abstract of the citing patent, the abstract of the cited patent and a signal indicating whether a citation relationship exists in-between. Considering that the fine-tuning task belongs to binary classification, the dataset is balanced with a 1:1 ratio of positive to negative samples. After fine-tuning, this model demonstrated remarkable performance in classifying the citation relationships by achieving a score of 0.89 in F1-value.

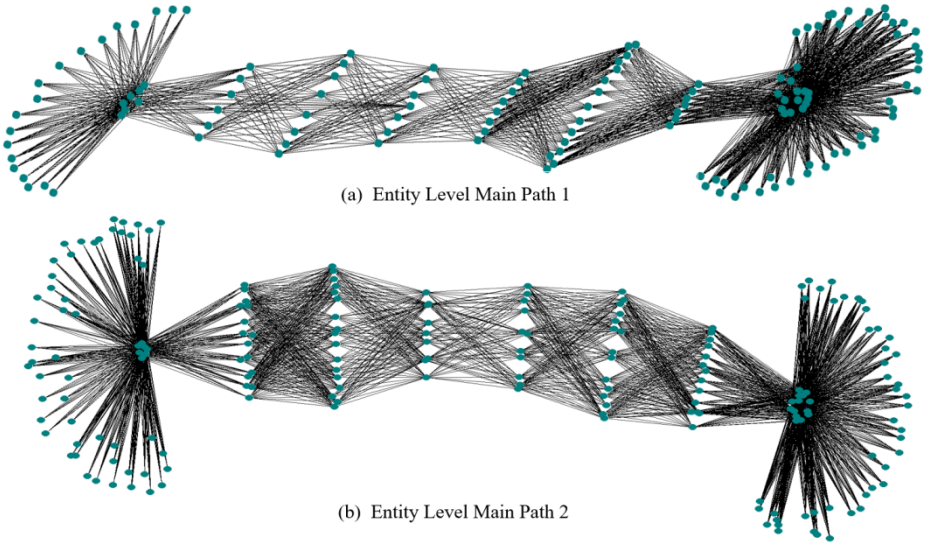


Fig. 11. Initial Entity Level Main Path

By fine-tuning the above classification task, the joint model optimizes the self-attention mechanism within to identify the entity pairs across documents with citation relationships. To gain an in-depth understanding of the joint model in this task, the weights of different entity pairs on the self-attention layers of the joint model are plotted in Fig. 12, where x-axis is the weight value and y-axis indicates the number of entity pairs, yellow dots indicate entity pairs from two documents with citation relationship, while the blue dots indicate the opposite. It is not difficult to see that the two types of dots differ significantly in terms of weight at self-attention layers, suggesting that the self-attention layers can effectively distinguish between entity pairs in two documents with citation relationship and those without.

Through the fine-tuned joint model, entity pairs with citation relationships path 1 and path 2 are identified and assembled to the entity main paths in Fig. 13 and 14. In these figures, vertices of different colors represent different types of knowledge entities. More detail information about the knowledge entities types can be referred to Table 1.

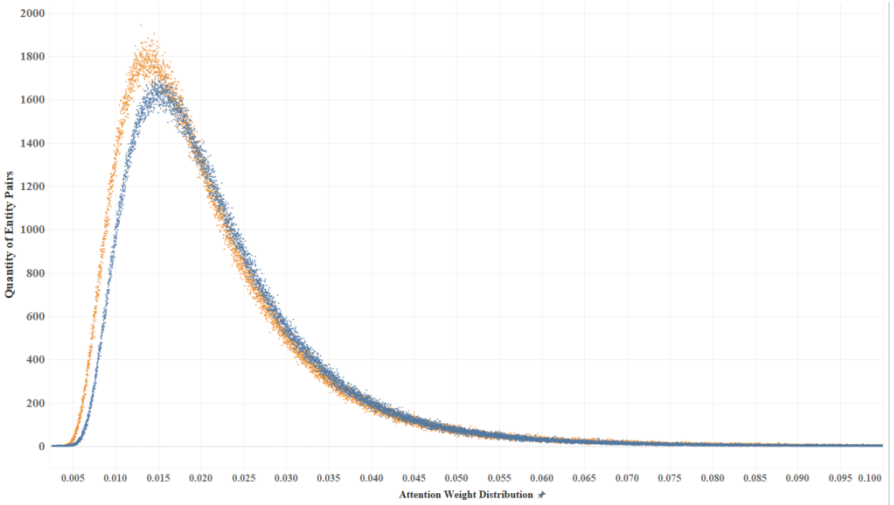


Fig. 12 Distribution of Self-Attention Weights

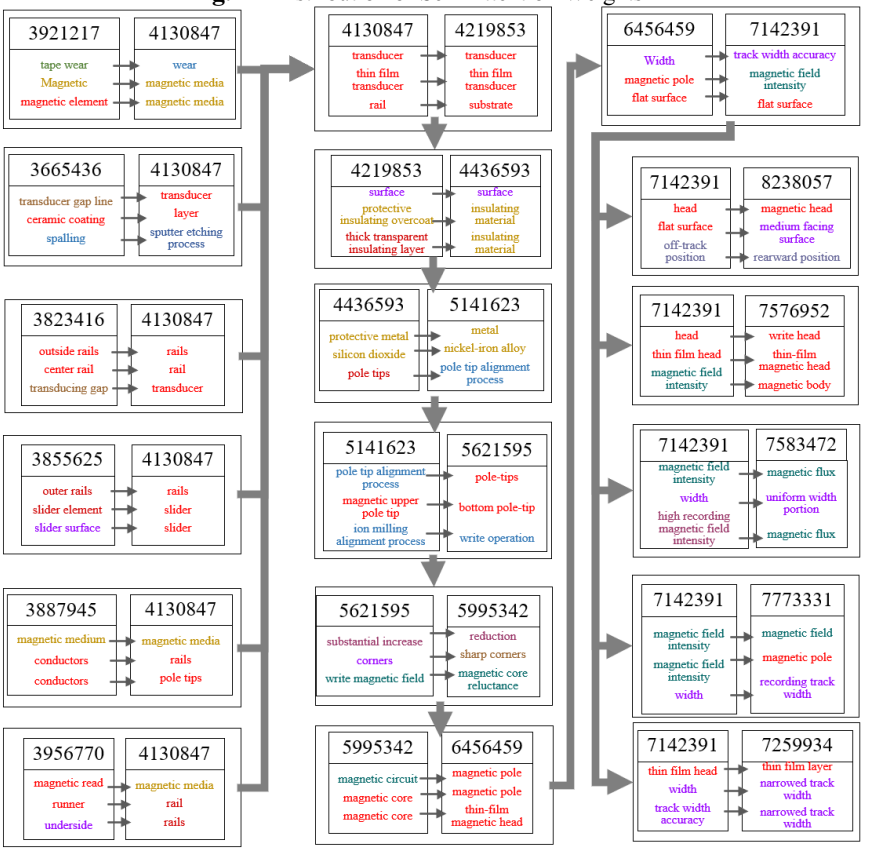


Fig. 13 Entity-based main path 1

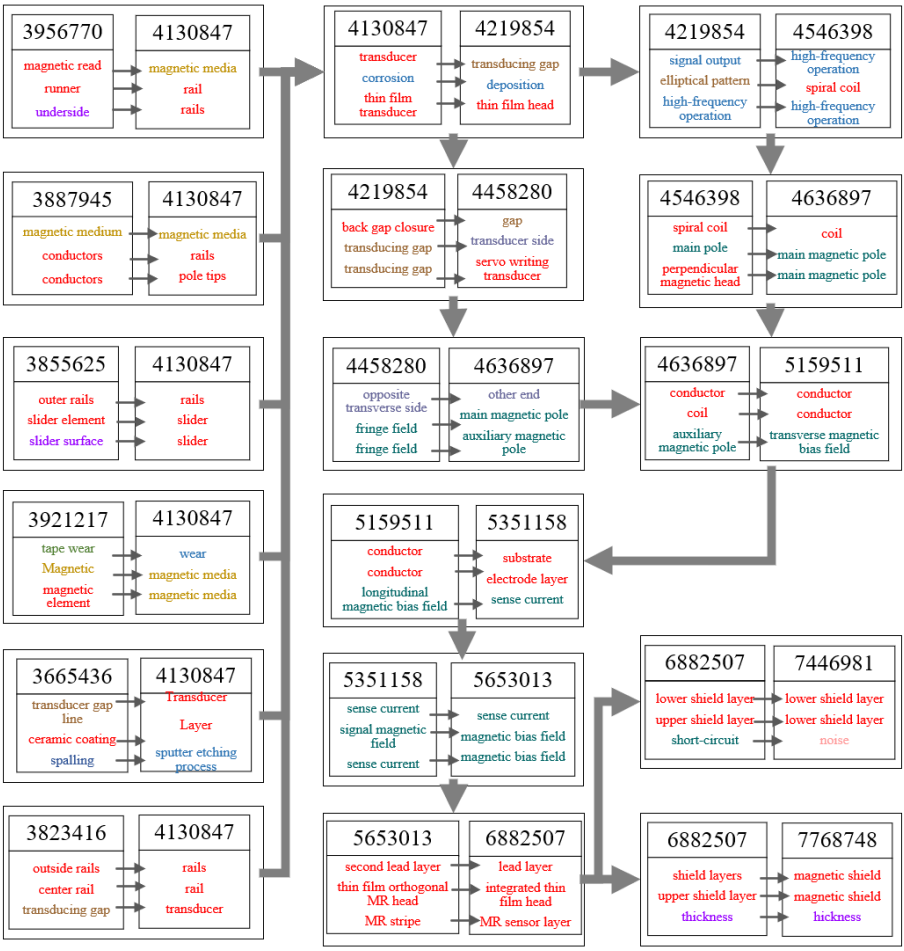


Fig. 14 Entity-Based Main Path 2

Table 1. Entity Types and Color Indications

Entity Type	Description	Color
Energy flow	Entity relevant to energy	
Location	Place or position	
Effect	Change caused an innovation	
Function	Manufacturing technique or activity	
Shape	The external form or outline of something	
Component	A part or element of a machine	
Attribution	A quality or feature of something	
Consequence	The result caused by something or activity	
Material	The matter from which a thing is made	
Scientific concept	Terminology used in scientific theory	

4.5 Experimental Results & Discussions

To obtain an in-depth understanding of the characteristics of the entity-based main path analysis method and achieve valuable insights about the knowledge evolution in micro-level, the experimental results are further analyzed at the entity and path levels respectively. We find that:
(1) Knowledge development exhibits distinct characteristics in different stages. As can be seen from path 1 and path 2, in early stage patents were primarily focused on material science, inventors strived to find suitable materials to serve as magnetic mediums, insulating materials and protective metals to improve the data storage density and read-write efficiency in hard disk drive. But with the advancement of technology, inventors have begun to develop more compact and efficient structures to improve the performance of hard disk drive in terms of energy conservation, portability and high-speed read-write capability. This indicates the mature stage of hard disk drive technology.
(2) The direction of R&D shifts from being singular to diverse along the entity main path. In early stage, inventions were mainly concentrated on system design and material selection, which indicates that the primary objective at that time was to realize the basic functions of hard disk drive and to validate the technological feasibility. But as technology advances, numerous requirements have appeared, such as noise reduction, energy conservation, stability in data transmission, etc. In response to these challenges, an increasing trend of diversity in R&D activities began to emerge in later stages.
(3) As for path 1 and path 2, it is not difficult to see that they share some features in common, that is, the patents in early stage mainly pertain to system design and material selection. But with the advancement of technology, the R&D directions of the two paths diverge: path 1 endeavors to improve the performance of data transmission in hard disk drive by enhancing the track width accuracy, optimize the alignment process of pole tip to magnetic head and so on, while path 2 starts from the fundamental principles of electromagnetism and achieves technological progress through refining the structure of hard disk drive.

5 Conclusion

Due to the ability of tracing the most significant developmental path of a field through a citation network, the MPA method is widely used to find a set of documents that plays an important role in a specific area and identify a main evolutionary pathway in scientific and technological field. Especially when the citation network gets larger and more complicated, it significantly relieves researchers and technology managers from laborious and cumbersome job in literature review. But traditional main path views documents as vertices while disregards that these documents are professional texts and researchers have to achieve valuable insights from the main path via manual reading and summarization, which is not only time consuming but also suffers from subjective bias.

To remedy the weakness, this study leverages the knowledge entities identified from texts to facilitate an entity-based MPA method to trace knowledge flow in citation network at micro-level. For the new type of main path, the vertices indicate knowledge entities from corresponding documents and the links connect knowledge entities with

citation relationships across documents. To identify entity pairs with citation relationships across documents, a joint model incorporating three-level masking attention mechanism is proposed and a fine-tuning task, i.e., classification of whether a citation relationship exists between two documents is suggest to improve the joint model's performance in identifying entity citation relationships between documents. To demonstrate the advantages of our method, extensive experiments are conducted on a patent dataset pertaining to hard disk drive in computer hardware. Experimental results show that our method is capable of discovering more detailed knowledge flows from the focal citation network, and revealing the characteristics of knowledge development in different stages as well.

Though, this study is subject to the following limitations. (1) Knowledge entity varies in information granularity, some entities represent the entire system or sub-systems, while others represent its components. The proposed method doesn't take the information granularity into consideration, which may increase the difficulties in interpreting entity-based main path. (2) The 3-layer masking mechanism in the joint model is an unsupervised deep learning method. However, it lacks of precise evaluation in identify entity citation relationships across documents, and there is considerable room left for its improvement. (3) Given the significant distinctions between paper and patent in citation network characteristics and textual features, the effectiveness of this proposed method on scientific paper need further validation.

Acknowledgements This research received the financial support from the Basic Scientific Research Business Project of the Central-Level Public Welfare Research Institute under grant number ZD2023-13, and National Natural Science Foundation of China under grant number 72274113, respectively.

References

1. Xue feng, W., Peng jun, Q., Yun, F.: The research on the construction of a new type of technology roadmapping: Based on SAO structure information. Studies in Science of Science 33(8), 1134-1140 (2015).
2. Jun fan, G., Xue feng, W., Peng jun, Q.: The research on construction model for technology roadmapping based on SAO analysis. Studies in Science of Science 32(7), 976-981,1002 (2014).
3. Nallapati, R. M., Ahmed, A., Xing, E. P., Cohen, W. W.: Joint latent topic models for text and citations. In: Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining, pp. 542–550. ACM, New York, NY, USA (2008)
4. Hummon, N. P., Dereian, P.: Connectivity in a citation network: The development of DNA theory. Social Networks 11(1), 39–63 (1989).
5. Fei fei, W. Yaru, Z. Lu cheng, H. Xin, L. Jing jing, L.: Multi-dimension Dynamic Evolution Analysis of Technology Topics Based on AToT by Taking Grapheme Technology as an Example. library and information service 61(5), 95-102 (2017).
6. Yu lin, L. Li rong, J.: The Evolution Analysis of Core Technology Cluster Based on Dynamic Patent Directed Network. journal of intelligence 40(04),101-108 (2021).

7. Liao, F., Ma, L., Pei, J.: Combined self-attention mechanism for Chinese named entity recognition in military. Future Internet 11(8), 180 (2019).
8. Lizhen, Q., Gabriela, F., Liyuan, Z., Weiwei, H., Timothy, B.: Named entity recognition for novel types by transfer learning. arXiv preprint arXiv:1610.09914 (2016).
9. Chen, L., Xu, S., Zhu, L.: A deep learning based method for extracting semantic information from patent documents. Scientometrics 125(1), 289-312 (2020).
10. Ding, Y., Song, M., Han, J., Yu, Q., Yan, E., et al.: Measuring the Impact of Entities. Entitymetrics 8(8), e71416 (2013).
11. Xiao, Y., Lu, L. Y., Liu, J. S., Zhou, Z., Knowledge diffusion path analysis of data quality literature: A main path analysis. Journal of Informetrics 8(3), 594-605 (2014).
12. Garfield, E.: Citation indexing for studying science. Nature 227.(5259), 669-671 (1970).
13. Garner, R.: Computer-oriented graph theoretic analysis of citation index structures (Doctoral dissertation, Drexel Institute of Technology). (1967).
14. Lucio Arias, D., Leydesdorff, L., Main path analysis and path dependent transitions in HistCite™ based historiograms. Journal of the American Society for Information Science and Technology 59(12), 1948-1962 (2008).
15. Liu, J. S., Lu, L. Y., Ho, M. H. C.: A few notes on main path analysis. Scientometrics 119(1), 379-391 (2019).
16. Huang, Y., Zhu, F., Porter, A. L., Zhang, Y., Zhu, D., Guo, Y.: Exploring technology evolution pathways to facilitate technology management: From a technology life cycle perspective. IEEE Transactions on Engineering Management, 68(5), 1347-1359 (2020).
17. Huang, Y., Zhu, D., Qian, Y., Zhang, Y., Porter, A. L., Liu, Y., Guo, Y.: A hybrid method to trace technology evolution pathways: a case study of 3D printing. Scientometrics 111(1), 185-204. (2017).
18. Tu, Y. N., Hsu, S. L.: Constructing conceptual trajectory maps to trace the development of research fields. Journal of the Association for Information Science and Technology 67(8), 2016-2031 (2016).
19. Kim, J., Shin, J.: Mapping extended technological trajectories: integration of main path, derivative paths, and technology junctures. Scientometrics 116(3), 1439-1459 (2018).
20. Maharana, K, Mondal, S, Nemade, B.: A review: Data pre-processing and data augmentation techniques. Global Transitions Proceedings 3(1), 91-99 (2022).
21. Dai, Z., Wang, X., Ni, P., Li, Y., Li, G., Bai, X.: Named Entity Recognition Using BERT BiLSTM CRF for Chinese Electronic Health Records. In: 12th International Congress on Image and Signal Processing, pp. 1–5. BioMedical Engineering and Informatics (CISP-BMEI), Suzhou China (2019).
22. Foltz, P. W.: Using latent semantic indexing for information filtering. ACM SIGOIS Bulletin, 11(2-3), 40-47 (1990).
23. Rodriguez, A., Laio, A.: Clustering by fast search and find of density peaks. science, 344(6191), 1492-1496 (2014).
24. Chen, L., Xu, S., Zhu, L., Zhang, J., Lei, X., Yang, G.: A deep learning based method for extracting semantic information from patent documents. Scientometrics 125(1), 289-312 (2020).
25. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In: 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL HLT 2019), Vol. 1, pp. 4171-4186. Association for Computational Linguistics, Minneapolis, MN (2019).
26. Google. BERT for Patents. http://github.com/google/patents-public-data/blob/master/models/BERT%20for%20Patents.md, last accessed 2023/08/26.

27. Rodriguez, A., Laio, A.:Clustering by fast search and find of density peaks. Science 344(6191), 1492-1496 (2014).
28. Xu, S., Qiao, X., Zhu, L., Zhang, Y., Xue, C., Li, L.: Reviews on determining the number of clusters. Applied Mathematics & Information Sciences 10(4), 1493-1520 (2016).