IRIS: Learning the Underlying Information of Scientific Research Interests Using Heterogeneous Network Representation

Zihan Feng
Department of Management Science and
Engineering, School of Economics and Management,
University of Science and Technology Beijing,
Beijing, China
fengzh18@hotmail.com

Hongfei Cui
Department of Management Science and
Engineering, School of Economics and Management,
University of Science and Technology Beijing,
Beijing, China
cuihf06@hotmail.com

Abstract

Understanding scientific research fields and finding potential relations between seemingly distinct fields can help researchers rapidly grasp their most interested topics with expertises. In this study, we construct a heterogeneous network which contains authors, keywords, papers and institutions, and built an "Integrated Research Interest Space (IRIS)" which can represent both author and keyword nodes. Similar keywords in the sense of research interest and research manner can obvious aggregate together. Authors that are interested in different keywords distributed in different IRIS areas, with strongly associated with research objectives and methodologies of the keywords. The average similarities between authors and their real used keywords is significantly higher than that of randomly chosen author-keyword pairs. Based on these observations, we propose a simple algorithm which attempts to recommend potential interested keywords for researchers, and got meaningful results. Our study may also give useful hints for understanding research interests and discovering potential cross disciplines.

1. Introduction

At present, with the continuous deepening of scientific research and the development of the times, scientific research activities that rely on the interrelationship between authors, institutions, and subject areas are becoming more frequent and closer. An individual's scientific research behavior may seem simple, but behind it there is a huge amount of information, for example: multiple authors who publish the same article have a cooperative relationship, the author has a affiliation with the institution filled in when publishing the article, and the keywords used in the author's writing reflect the author's research subject area and scientific research interests, etc. Relations between people, people and literature, people and keywords, people and institutions can establish large-scale scientific research relationship networks through abstraction. And these networks can refine the description of scientific research behaviors that have occurred and reflect the research interests of researchers as well as their research direction, which may provide an entry point for us to dig out the potential information of research behavior.

Based on the above idea, many researchers attempted to propose enlightening methods to mine the deep information from research networks. For example, the connectivity[1] and robustness[2] of networks have once been used as an indicator of the stability and openness of scientific research cooperation. In addition to the in-depth exploration of basic statistical information of scientific research networks, more and more authors begin to embed the nodes of scientific research networks using network representation learning algorithms. Some of them chose to use homogeneous algorithms applied to heterogeneous networks, such as LINE[3] and node2vec[4], while others selected heterogeneous algorithms such as metapath2vec[5], AspEM[6], BHIN2vec[7]. These approach represents the semantic information that buried in abstract networks as dense real-valued vector space, which makes relevant research more efficient and scalable as well as allows us to dig out more information about potential scientific research than before.

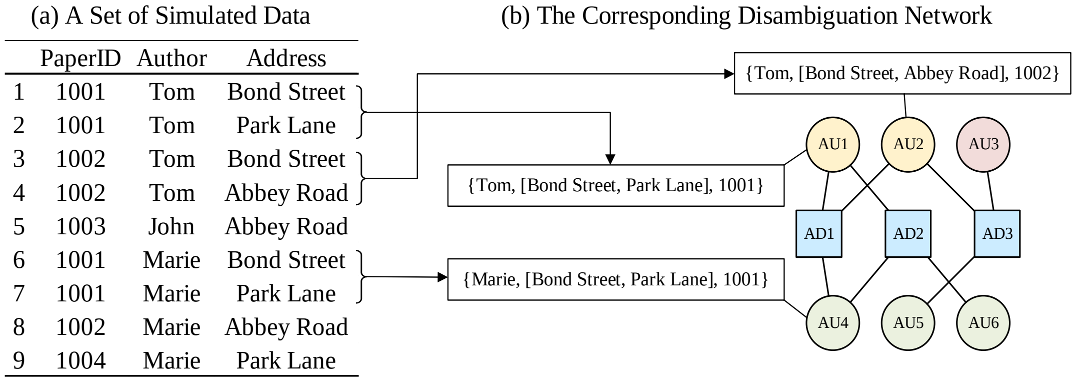
However, one question of the above studies is that most of them didn't pay enough attention to the vector spaces themselves obtained through embedding. In fact, the author vector spaces obtained by network embedding can not only inspires us in the cooperative relationship of researchers, but also providing us with important information such as their research interests and focused areas. Besides, there is a more direct way to express the authors' research interests or focused areas,

URI: https://hdl.handle.net/10125/79678
978-0-9981331-5-7

the keywords. On the one hand, the keywords of an article can reflect the research field of itself or the innovative way of solving problems raised in the article. On the other hand, the keywords describe their users' research interests and characteristics. And the co-occurrence of some keywords which are seemingly irrelevant may illustrate potential cross disciplines to some extends. Therefore, the research about keywords is able to help us understand the authors' interests, the subject itself and the probable intersection of subjects.

As to the existing researches about the "keyword", a large amount of them have studied it from all aspects. Behrouzi et al.[8] and Teklu et al.[9] once used the keyword network to finish the link prediction task in order to explore the evolution trend of a certain academic field. Although this usage had excavated some practical significance and deep meaning from keyword networks, keywords are only used as a mapping of the development of a certain academic field without discussing the relationship between keywords. And it also lacked analysis about researchers' keyword utilization from a intuitive point of view. In another research, Lu et al.[10] proposed an author-defined keyword frequency prediction (AKFP) method considering both authors and keywords content based on deep learning to detect research topics. AKFP seems to establish a relationship between the authors and keywords, but such relationship is only used to word frequency statistics and discussion about the author's personal research interests is still not involved. In addition, the relationships between authors and keywords have been studied using various methods in some papers [11,12]. Based on such relationships, related discussions about authors' research interests were also be conducted. But unfortunately, their research methods can neither support the discussion of keyword spaces, nor be used to analyze the relationship between different keywords.

In view of this, we construct a heterogeneous network containing the information of authors, institutions, papers, and keywords. After the network construction, a heterogeneous graph embedding method, metapath2vec, is used to obtain a heterogeneous vector space, which is called Integrated Research Interest Space (IRIS) because of its inclusion of research interests reflected by various related elements through scientific research activities. The IRIS contains high-dimensional vectors corresponding to each individual which are generated based on the relationship between the four types of scientific research elements (authors, institutions, papers and keywords). And in this paper, the author vector space and keyword vector space are selected for analysis. Through separate analysis of the two vector spaces, we find that the keywords show a distribution characteristics which is called "clustering by semantics" by us. The authors with different keywords interests tend to distributed in different IRIS areas and the influencing factors of such distribution feature include research objectives and methodologies of the keywords. Using these characteristics, we can get information about the potential disciplinary links and implicit authors' interest. More importantly, we have also observed that author vectors and keyword vectors have a tendency to "gather around the heterogeneous nodes which are connected to themselves in reality". Taking advantage of this tendency, we propose a simple rank-based keyword recommendation algorithm which can reflect one of the practicality of IRIS. In conclusion, the discovery of these features provides us with new perspectives and methods that can be used to discover the authors' research interest or characteristics as well as help us understand the subject content and the integration of disciplines more deeply. At the same time, it can also be used as a theoretical basis which might provide inspiration for the development of algorithms about scientific research prediction and recommendation.



(a) A Set of Simulated Data

	PaperID	Author	Address
1	1001	Tom	Bond Street
2	1001	Tom	Park Lane
3	1002	Tom	Bond Street
4	1002	Tom	Abbey Road
5	1003	John	Abbey Road
6	1001	Marie	Bond Street
7	1001	Marie	Park Lane
8	1002	Marie	Abbey Road
9	1004	Marie	Park Lane

(b) The Corresponding Disambiguation Network

{Tom, [Bond Street, Abbey Road], 1002}

{Tom, [Bond Street, Park Lane], 1001}

{Marie, [Bond Street, Park Lane], 1001}

Figure 1. Preparation of author disambiguation algorithm

2. Methods

2.1. Dataset used and preprocessing

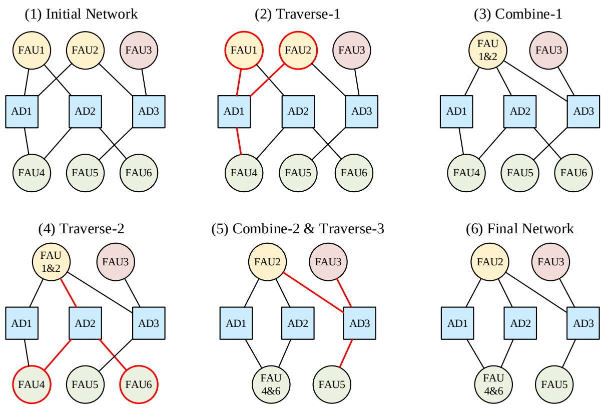


Figure 2. Execution of author disambiguation

2.1.1. Dataset download. To obtain data used to describe scientific research behavior, we downloaded all the papers during 2010.1 to 2021.2 in the domain of the Medical Big Data from PubMed website (https://pubmed.ncbi.nlm.nih.gov/), where can easily get the detail information about published papers, as our study dataset, using a retrieval formula '(Deep learning) OR (Machine learning) OR (Neural network)' in the downloading format 'PubMed', which contains information of PaperID, Full author name, Institution address (called 'Address' in the following paragraph), Medical Subject Headings (called 'Keywords' in the following paragraph), etc. Among those items, the Medical Subject Heading is a kind of artificial language that can express the subject of papers, with the characteristics of conceptualization and standardization which ensure the accuracy of our research about keywords. And this is also a big advantage of using PubMed website as our data source.

2.1.2. Author disambiguation. Consider the situation that different authors may have a same name, we propose an author disambiguation algorithm using the authors' institution addresses. The main idea is that if two authors from two papers have a same full author name with at least one same affiliated institution address among several addresses left when publishing, they will be regarded as the same author and finally represented by a same author ID. This approach is based on an assumption which is reasonable that there are no authors with the same name in an institution.

As a preparation of the author disambiguation algorithm (Fig.1), we first organized the downloaded records into a table as the simulated data shown in Fig.1-a, whose each row contains an address for an author of a specific paper. Then we built an undirected heterogeneous "disambiguation network", which abstracts the relation of authors' names and addresses from all papers in our dataset. The detailed network building method is shown in the corresponding relationship between Fig.1-a and Fig.1-b. During the execution of the algorithm, the AD nodes will be traversed one by one. For each AD node, all the AU nodes connected by the edge to such AD node will be compared in pairs. If one pair of AU nodes have the same full author name, the two AU nodes of this pair will be merged into a single node that linked to all AD nodes which were connected by the two AU nodes before the merging operation. When the traversal of all AD nodes as well as the compare and merge operation are completed, the whole algorithm ends. The schematic diagram of the author disambiguation algorithm is shown in Fig.2.

2.1.3. Keyword preprocessing and data selection. When sorting out the keyword information in the data set, we find that some papers did not have and keywords (that is, these papers does not have Medical Subject Headings in original data files). For these papers, we

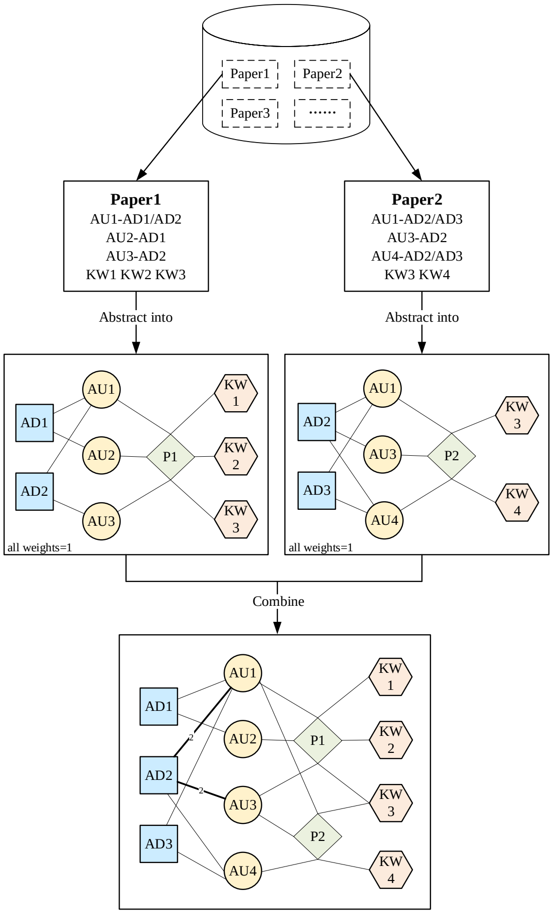


Figure 3. Construction of network

directly removed them because they did not reflect any things about the authors' research interests.

In addition, in order to reflect the authors' partnership and their usage of keywords more clearly, all the authors and keywords that only appear once will be removed. The reserved authors have more collaboration with others and the reserved keywords are used more frequently.

2.2. Construction of the integrated research heterogeneous network

We construct a heterogeneous network according to the method shown in Fig.3. The integrated research heterogeneous network contains four types of nodes: Author (AU) nodes, Address (AD) nodes, Paper (P) nodes and Keyword (KW) nodes. Among the four types of nodes, three types of undirected edges based on their connection in reality are formed: (AD, AU) edges, (AU, P) edges and (P, KW) edges.

It is also worth mentioning that for each edge connected with Paper nodes has the weight of 1, while edges between the AU node and the AD node may have weights other than 1. This is because an edge connected with a same pair of AU node and AD node may appear in different sub-networks more than one times. For example, in Fig.3, the edge (AD2, AU1) appears both in Paper1 and Paper2, that's why the weight of edge (AD2, AU1) is 2 in the final network. When an author repeatedly registers the same address when publishing different papers, the weight will increase. The weight is an objective reflection of the possible situation that an author belongs to multiple institutions in our raw data.

2.3. Construction of IRIS

In order to further explore the inner connection between scientific researchers and keywords, we construct the Integrated Research Interest Space (IRIS) here. The purpose of this step is to embed each network node into a vector through the network representation learning algorithm. And the embedding vector space containing all the vectors is what we call IRIS. The essence of IRIS is a vector space that can reflect the research interests of scientific researchers. The word 'Integrated' means that it contains four types of vectors (Paper, Keywords, Author and Address) rather than just containing simple relationships between nodes with only one type. In order to make the exploration more thorough, we mainly explore the significance of Author vectors and Keyword vectors in this paper, but there is no doubt that, in IRIS, there are still many potential relationships between other types of vectors that can be explored.

At present, there are many algorithms that can be applied to network embedding, such as node2vec[13], LINE[14], SDNE[15], etc. which are suitable for homogeneous networks, and metapath2vec[16], HIN2vec[17], GATNE[18], etc., which are suitable for heterogeneous networks. Considering the heterogeneity and the large scale of our integrated research network, we choose to do network embedding with metapath2vec, an algorithm based on random walks according to meta-path to construct heterogeneous neighborhoods of nodes and then uses heterogeneous skips-gram model to perform node embedding. While metapath2vec was proposed, a similar algorithm metapath2vec++, was also proposed. In the choice of metapath2vec and metapath2vec++, we are inspired by a result of an empirical research applied by the algorithm proponents which was written in the latter part of the corresponding paper. In this empirical study, the author found that, in the vector space, metapath2vec++ often separates two different types of nodes into two columns after dimensional reduction and each column distributes one of the types of nodes. Differently, in the vector space produced by metapath2vec, each group of logically connected heterogeneous nodes is distributed in a short distance in a two-dimensional space. Considering that the cosine similarity (which will be explained in detail in Section 2.4) will be used to describe the similarity between vectors, we choose metapath2vec as the building algorithm of IRIS. Using metapath2vec, the spatial distribution of heterogeneous vectors can reflect the distance between different vectors more directly. What's more, the distribution that reflects the relevance of things with 'adjacent form' is in line with our intuitive perception of relevant objective things more.

When using metapath2vec, we should specify the meta-path scheme in order to guide the random walk. We finally define 'O-A-P-K-P-A-O' as our meta-path scheme referring to an effective meta-path that is often applied to classic DBIS dataset. Among such schema, 'O' represents the Address nodes (the first letter of the synonym 'Organization'), 'A' represents the Author nodes, 'P' represents the Paper nodes and 'K' represents the Keyword nodes. This meta-path includes all types of nodes in our network. And it can directly reflect the three kinds of important information: the affiliation between the author and the organization (reflected by 'O-A' and 'A-O'), the relationship between a couple of authors due to their keyword-interests (reflected by 'A-P-K-P-A') and the authors' usage of keywords (reflected by 'A-P-K' and 'K-P-A'). With these information and other hidden meaning in this meta-path, more semantics can be integrated into IRIS and the vector space will be more informative. Additionally, when performing the random walk step, the walk between 'O' and 'A' (that is to say when passing the 'O-

A' or 'A-O' edges) will be affected by the weight of the (AD, AU) edges. The edge with a higher weight has a higher probability of being passed, and this probability is linearly distributed depending on the weight of this edge.

Finally, we also need to determine some common parameters, these parameters are listed below:

(1) number of walks per node *w*: 500;
(2) walk length *l*: 50;
(3) vector dimension *d*: 128;
(4) neighborhood size *k*: 7;
(5) size of negative samples *s*: 5.

Among them, the parameters *d*, *k* and *s* are set according to the default value of the original paper; the other two parameter *w* and *l* are appropriately reduced on the basis of the default value.

2.4. Similarity calculation between authors and keywords

Cosine similarity is an index that is widely used in machine learning to measure the similarity of two vector objects. In many existing studies, using cosine similarity to describe the similarity between two vectors has been favored by many researchers. More importantly, the use of this indicator can often lead to good research results [19–21].

In this paper, we also select cosine similarity as the indicator to measure the similarity between the two different vectors (the author vector and the keyword vector) and apply it to analyze the heterogeneous structures in IRIS. The similarity between the n-dimensional author vector $\boldsymbol{A}$ and the keyword vector $\boldsymbol{K}$ is calculated as follows:

$$sim(\boldsymbol{A},\boldsymbol{K}) = \frac{\sum_{i=1}^{n}(A_i \cdot K_i)}{\sqrt{\sum_{i=1}^{n} A_i^2} \cdot \sqrt{\sum_{i=1}^{n} K_i^2}}$$

where A_i and K_i represent the i-th dimension value of n-dimensional vectors $\boldsymbol{A}$ and $\boldsymbol{K}$. In IRIS, both $\boldsymbol{A}$ and $\boldsymbol{K}$ are 128-dimensional vectors, so the value of n is 128. According to the characteristics of cosine similarity, the closer the value is to 1, the higher the similarity between the author corresponding to vector $\boldsymbol{A}$ and the keyword corresponding to vector $\boldsymbol{K}$ is. It also means that the keyword is more relevant to the author's scientific research field, and the author is more likely to form interest in such keyword.

3. Results

3.1. Basic statistics on integrated research heterogeneous network

When processing our dataset downloaded from PubMed, we completed the processing in the order of author disambiguation (Section 2.1.2 in Methods), removal of single-occurring authors, removal of papers without keywords and removal of single-occurring keywords (Section 2.1.3 in Methods). During these operations, the number of papers, authors, addresses and keywords will decrease in each step, which corresponds to the reduction of the number of nodes in the integrated research heterogeneous network. The specific numbers of nodes when finishing each step are shown in Table 1.

After completing all steps of data preprocessing, the network has 14,143 Paper nodes, 17,894 Author nodes, 15,676 Address nodes and 7,140 Keywords nodes (as shown in the last row of the Table 1).

3.2. Analysis of keyword vector space in IRIS

3.2.1. Visualization of keyword vector space. In order to study IRIS clearly, we choose to only study the keyword vector space of IRIS and explore the distribution feature of these keyword nodes at the beginning. To visually show the distribution of each keyword in such vector space, we use a dimensional reduction algorithms, t-SNE, on some vectors of IRIS to facilitate visualization. The t-SNE is a machine learning algorithm, which is specially used for dimensional reduction. With its non-linear feature in the algorithm's principle, it is suitable for reducing high-dimensional vectors to low level (2D or 3D). When t-SNE is executed, the two types of vectors in IRIS, keyword vectors and author vectors, are simultaneously put into this algorithm and then get their two-dimensional vector representation in the output. What needs to be explained here is that using keyword and author vectors as the input of t-SNE at the same time is to ensure that the

Table 1 Number of nodes in the network

Step	Processing	Paper	Author	Address	Keyword
1	Origin	131,691	626,909	264,770	63,416
2	After Author Disambiguation	131,691	591,224	264,770	63,416
3	After removing single-occurring authors	22,856	23,614	22,398	20,284
4	After removing papers without keywords	14,146	17,894	15,677	20,284
5	After removing single-occurring keywords	14,143	17,894	15,676	7,140

internal connection between them will not lose due to the sampling during execution of t-SNE.

After t-SNE, we draw all the 2-dimensional keyword vectors on the coordinate plane, and get the keyword vector distribution map shown in the center area of Fig.4 below. According to the principle of the t-SNE algorithm, the distribution of keyword vectors in this plane can intuitively reflect the distribution of keyword vectors in IRIS.

3.2.2. Keyword distribution features in IRIS. In Fig.4, each blue dot represents a keyword. Overall speaking, keywords dispersed throughout the two-dimensional plane, which means that keywords are also widely distributed in IRIS. However, besides widely distributed, there is a obvious local clustering phenomenon of them. In other words, the keyword vectors are not absolutely uniformly dispersed throughout IRIS, some keywords will gather together to form a keyword cluster. Based on this phenomenon, we conduct a semantic analysis on some keyword clusters with obvious aggregation located in the marginal area and find that the keywords with similar semantics or logical-related relationships tend to gather together and form clusters in IRIS. The semantic analysis results of some keyword clusters are shown in Fig.4.

Taking the green keyword cluster on the left side of Fig.4 as an example, we find that these dots represent the keywords such as 'Conflict', 'Psychological', 'Depressive Disorder', 'Citalopram' and 'Hippocampus'. From the semantic level, we can easily find the relationship between them. For instance, the word 'Depressive Disorder' is a kind of mental illness, which can be easily caused by 'Psychological Conflicts'. And the word 'Citalopram' mentioned in these keywords is a kind of prescription drug that can produce antidepressant effects which is often used in the treatment of depression. The term 'Hippocampus' that is an important part of our brain. From a medical point of view, if the glucocorticoids (a kind of hormone) continuously release simulated by realistic pressure, the hippocampus will be affected and finally lead to the depressive disorder. This mechanism is a common cause of depression. Based on the above analysis, we can find that the keywords in this cluster contain many kinds of semantic information related to depression, such as the depression treatment methods and some causes of depression.

Take the brown area at the bottom of Fig.4 as an example again, this area contains many ophthalmic diseases related terms such as 'Glaucoma' and 'Optic Nerve Diseases', as well as 'Visual Fields', 'Intraocular Pressure', 'Retinal Ganglion Cells' and other biological terms related to human eyes and sight. So we summarize this cluster as keywords related to eye diseases.

Using the same analysis methods, it can be easily found that the purple area on the lower right side of Fig.4

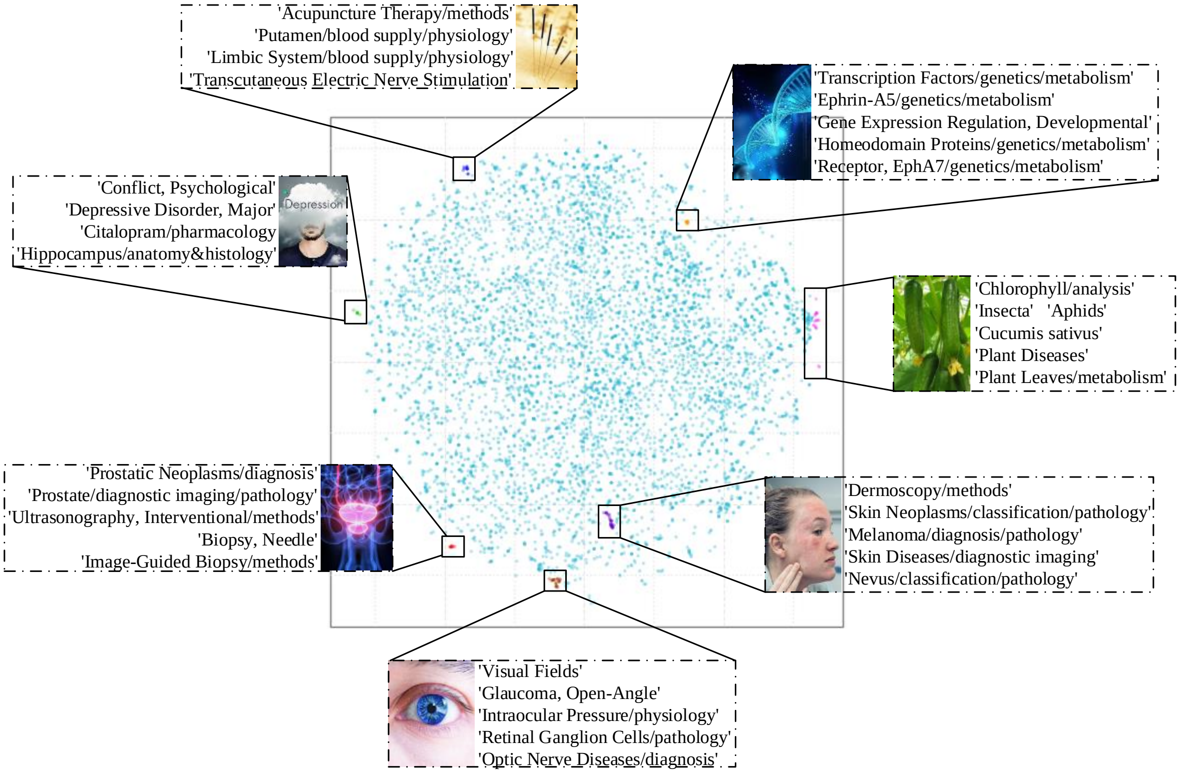


Figure 4. Distribution diagram of the keyword vector space

is related to ‘skin diseases’, and the red area on the lower left side is related to the diagnosis of ‘prostatic neoplasms’. As to the remaining areas, they are related to the topics of ‘plant diseases’, ‘gene expression’ and ‘acupuncture and blood supply’ respectively.

Through the analysis of these seven regions, we find that the keywords in IRIS show a clear pattern of ‘clustering by semantics’. And this pattern is not only applicable to these seven regions, semantic relevance of other keyword clusters in IRIS can also be observed.

3.3. Analysis of author vector space in IRIS

Through the analysis of the keyword vector space, we observe that the keyword vector has the distribution characteristics we call ‘clustering by semantics’, which leads us to explore the author vector space in the same way. Considering that keywords can well reflect the authors’ research interests and their research directions, we select the keywords that are frequently used in the top 30, and then dye the author vector space 30 times with the criterion of ‘whether he/she used the keyword’.

For a certain highly popular keyword, if the author have used the keyword in his published paper in IRIS, the dot corresponding with this author will be dyed red and the other authors who have not used the keyword will be represented by blue dots. Through this method, we find that in IRIS, author vectors distribute strongly associated with research objectives and methodologies of the keywords. What’s more, by comparing the thirty vector dying maps, some novel difference or overlap of authors’ research interests can be exposed. Two examples of relevant analysis are as follows.

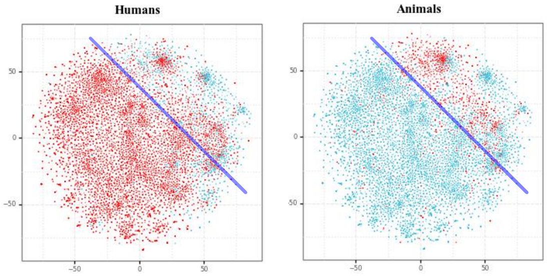


Figure 5. Author distribution - ‘Humans’

As shown in Fig.5, by using the hot keywords ‘Humans’ and ‘Animals’ to dye, we can clearly see that authors whose research object are ‘Humans’ tend to distribute on the bottom left of the two-dimensional plane and authors studying ‘Animals’ tend to distribute on the upper right side of the plane. This may indicate that the two keywords represent two different research directions, and most authors will only choose one of them to devote their efforts and that is what we call ‘ the difference of authors’ research interests ’. But this dedication is not absolute, besides the blue line, it can be seen that there are still many authors who study Humans and Animals together, that is what we call ‘ the overlap of authors’ research interests’.

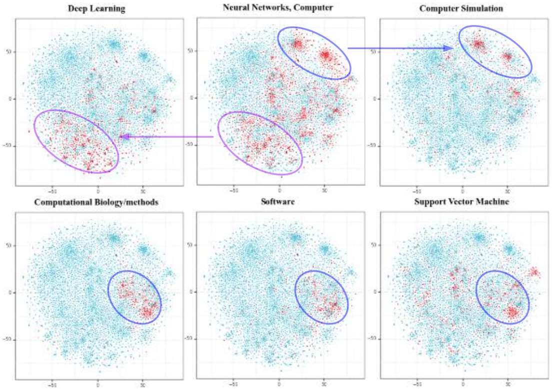


Figure 6. Author distribution - ‘Algorithms’

By comparing the maps dying with some keywords that indicate methodology, we can find multiple ways of using these methodologies in different studies. As shown in Fig.6, we find that among all the authors who use the ‘Neural Networks’ method, those in the first quadrant tend to use the keyword ‘Computer Simulation’, while some authors in the third quadrant tend to use the keyword ‘Deep Learning’. From this phenomenon, we infer that the application of neural networks in the field of life sciences is not single. Some researchers apply neural network along with computer simulations to simulate life activities or physiological structure of living body, while others apply it to deep learning and use relevant methods to process or mine the large amounts of medical data.

At the same time, we find that users of the three keywords ‘Computational Biology/methods’, ‘Software’ and ‘Support Vector Machine’ overlapped widely around the positive half of the x-axis. This overlap can be understood from a realistic perspective. Based on our understanding of Computation Biology, we know that most of the mature research methods in this field are finally presented in the form of software, so authors who use ‘Computation Biology/methods’ is highly overlapping with the authors using ‘Software’. As for the overlap between the ‘Computational Biology/methods’ and ‘Support Vector Machine’ areas, we explain as follows. Comparing with other fields, sample collection in the field of computational biology is more difficult. Therefore, when using algorithms about machine learning, researchers in this field tend to use SVM, a traditional machine learning method that does not require large amounts of data. The low overlap between the authors of ‘Computational Biology’ and ‘Deep Learning’ can verify our analysis as well. When using ‘Deep Learning’, the large-capacity dataset are often needed, so it does not often appear in the application of Computational Biology.

In this part, we perform a simple analysis of the author vector space in IRIS relying on keywords. Then we find that the distribution of authors is affected by two factors, the research objectives and methodologies of the keywords. These two factors are found through our limited experiments, and perhaps more factors can be discovered through more similar experiments.

3.4. The relationship between author vector space and keyword vector space

In order to study the distribution of similarities between each author and the keywords he/she used, we defined an indicator called Keyword Concentration Index (KCI) to measure the aggregation degree of an author with the keywords in IRIS he used in reality. Denote the corresponding vector of a certain author *a* in IRIS as *A*, and the set of his published papers is defined as *P* = {P_1, P_2, …, P_i, …}, the *i*-th element P_i in the paper set *P* represents the *i*-th paper published by that author. For each element in *P*, P_i = {$\boldsymbol{K_{i1}}$, $\boldsymbol{K_{i2}}$, …, $\boldsymbol{K_{ij}}$, …}, where $\boldsymbol{K_{ij}}$ represents the vector corresponding to the *j*-th keyword in the paper P_i in IRIS, and the author's KCI is calculated as follows:

$$KCI_a = \frac{\sum_{P_i \in P} \sum_{K_{ij} \in P_i} sim\left(\boldsymbol{A}, \boldsymbol{K_{ij}}\right)}{\sum_{P_i \in P} |P_i|}$$

To test whether the KCI similarities are significantly higher than random cases, we also calculate KCI_a^{pt} as follows, which is the average similarity between an author to random selected keywords. We used the KCI_a^{pt} values to do a permutation test.

$$KCI_a^{pt} = \frac{\sum_{P_i \in P} \sum_{K_{ij} \in P_i} sim(\boldsymbol{A}, \boldsymbol{K_{random}})}{\sum_{P_i \in P} |P_i|}$$

The probability density distributions of KCI_a and KCI_a^{pt} are shown in Fig.7-A. We can observe clearly that for the author's real used keywords, almost all the KCI values are higher than 0.25. In sharp contrast, for the randomly chosen author-keyword pairs, most KCI values are lower than 0.25. It means that in IRIS, the aggregation degree of an author with his keywords used in real scientific research activities is higher than the keywords selected randomly in our permutation test.

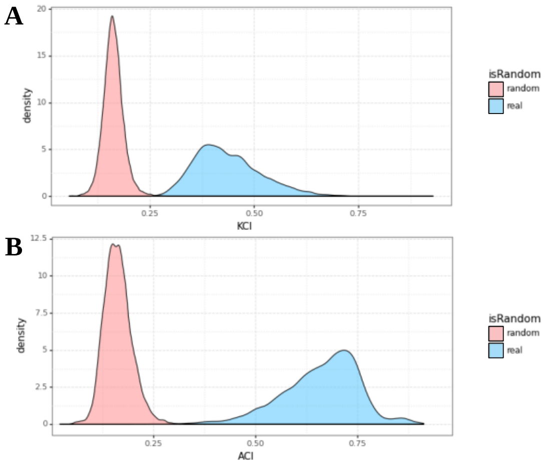


Figure 7. Probability density distribution of KCI and ACI

Similarly, we also define indicator about author called Author Concentration Index (ACI) to measure the aggregation degree of a certain keyword with the authors in IRIS who used it in their papers, and an index, ACI_k^{pt}, that shows the aggregation degree in random cases. The ACI of a keyword *k* can be calculated as follows:

$$ACI_k = \frac{\sum_{P_i \in P} \sum_{A_{ij} \in P_i} sim\left(\boldsymbol{K}, \boldsymbol{A_{ij}}\right)}{\sum_{P_i \in P} |P_i|}$$

and the ACI_k^{pt} can be calculated as follows:

$$ACI_k^{pt} = \frac{\sum_{P_i \in P} \sum_{A_{ij} \in P_i} sim(\boldsymbol{K}, \boldsymbol{A_{random}})}{\sum_{P_i \in P} |P_i|}$$

The probability density distribution of ACI_k with real keyword-authors relations and the same distribution of ACI_k^{pt} with random chosen authors are shown in Fig.7-B. Similar to the KCI distribution, we can clearly see that for the real users of keywords, ACI values are almost all higher than 0.3, and most of them distribute from 0.5 to 0.75. While for authors who randomly select keywords in permutation test, most ACI values are lower than 0.3.

Through studying two indicators, KCI and ACI, we find that no matter the keywords or the authors, they tend to distribute around the heterogeneous node that have a relationship in with them in the scientific research activities that have taken place.

3.5. Inspiration for keyword recommendation

According to the distribution feature of keywords with the authors who used them discussed in Section 3.4 using the indicator KCI, we propose a simple similarity rank-based method for keyword recommendation.

For author *A* who needs keyword recommendation, this method will be performed as the following steps:

(1) Calculate the similarity *sim*(***A***, ***Ki***) between the vector of Author *A* and all the keyword vectors K_i in IRIS in turn;

(2) Filter the keywords which similarity with Author A is too low, with a threshold KCI_0;

(3) Sort all the remaining keywords according to their similarities with Author *A* in descending order;

(4) After sorting, the keywords with the highest similarity ranking are the recommended keywords.

According to Fig.7-A, here, we heuristically set the threshold KCI_0 as 0.3, which is higher than almost all random Author-Keyword similarities and lower than most KCI values of authors with their real used keywords.

Then we randomly select one of the authors in IRIS, Yuki Sakai, and performed keyword recommendation through this rank-based method. Table 2 shows the top-10 recommended keywords and their similarity with Yuki Sakai. The green keywords in the Table 2 represent that the keywords have been used by the author and the red keywords represent that they have not been used. For the red keywords, we find that these results are reliable to understand from the semantic level.

Table 2. Keyword recommendation results

Top-n	Recommended Keywords	Similarity
1	Obsessive-Compulsive Disorder/diagnostic imaging/pathology/physiopathology	0.6556
2	Neural Pathways/diagnostic imaging/pathology/physiopathology	0.5854
3	Brain/diagnostic imaging/pathology/physiopathology	0.5693
4	Gene Knockdown Techniques	0.5293
5	Corpus Striatum/diagnostic imaging/pathology/physiopathology	0.5157
6	Anxiety/physiopathology	0.5083
7	Ventral Tegmental Area/physiopathology	0.4995
8	Disease Models, Animal	0.4540
9	Stress, Psychological/diagnostic imaging/physiopathology	0.4395
10	Mice	0.4126

By reviewing the author's related information from the original dataset, we find that the author corresponds to two published papers: 'A common brain network among state, trait, and pathological anxiety from whole-brain functional connectivity'[22] and 'Diffusion functional MRI reveals global brain network functional abnormalities driven by targeted local activity in a neuropsychiatric disease mouse model'[23]. By reading the title and abstract of the two papers, it can be seen that the papers are relevant with the topic of neurological diseases based on brain network. Focusing on the results of keyword recommendation, the results ranking 5, 7 and 9 are all keywords with the subtopic word 'physiopathology', which is semantically consistent with what we have learned about the author's research topic (in fact, many keywords used by this author contains the subtopic 'physiopathology'). Furthermore, the 'Corpus Striatum' and 'Ventral Tegmental Area' mentioned in the recommended results are both the physiological structure of the brain, which are widely discussed in neurobiological theories. There is no doubt that these two keywords are consistent with the author's research topic about the brain network, so the results may provide inspiration for the author's next research.

Through the above analysis, we find that the recommendation results given by the rank-based method can withstand the semantic scrutiny, which proves that the results given by such methods are reliable at the semantic level. And this success also tells us the research on IRIS can indeed help us solve many practical problems.

4. Discussion and Conclusion

In this study, we established a heterogeneous network to reflect the authors' research interests and then used metapath2vec to embed it and finally get a vector space containing four types of vectors called IRIS. Through the separate analysis and integrate analysis of the author vector space and the keyword vector space in IRIS, we find the keyword and author vectors' distribution characteristics in their own separate vector spaces and the interaction between the two vector spaces. At the end of the paper, a simple rank-based keyword recommendation algorithm is proposed. These patterns we find can not only help us understand the authors' interest, the subject itself and the disciplinary integration from a realistic perspective, but also can be used as a kind of theoretical basis which may provide us with some methodological inspiration for practical applications when analyzing other social systems.

However, there are still many aspects of our research that can be improved. First of all, this article only focuses on the characteristics and connections of the authors and keywords in the vector space. In fact, there are many elements worth studying, such as the authors' addresses and the journal of each paper. Secondly, the embedding algorithm we use is relatively simple, so the overall work can be seen as a pilot study which proves the effectiveness of the vector space analysis method in the study of heterogeneous networks. Thirdly, our results can provide guidance for the keyword recommendation task, but limited by the length of the article, the in-depth research cannot be presented here. And this topic is really worth discussing in the future studies.

In conclusion, we have drawn many practical conclusions that can help people better understand the research behavior through our research about IRIS. But as a pilot study based on heterogeneous graph embedding methods, our experimental process is relatively concise. Many meaningful excavations about

IRIS are worthy of further supplementation and expansion in our future work.

Acknowledgement

This work was supported by the National Natural Science Foundation of China (No. 72101029), Beijing Municipal Natural Science Foundation (No. 4204104), and the Fundamental Research Funds for the Central Universities (No. FRF-TP-18-013B1). We thank Mr. Xinglai Zhang, Mr. Jingyu Zhang and Miss Yuena Zhao for their useful suggestions.

References

[1] Newman M E J 2001 Scientific collaboration networks. I. Network construction and fundamental results *Phys. Rev. E* **64** 016131

[2] Oh P 2015 THE EVOLUTION OF SCIENTIFIC COLLABORATION NETWORKS 232

[3] Zhang J, Yu W, Liu J and Wang Y 2018 Predicting Research Collaborations Based on Network Embedding *J. China Soc. Sci. Tech. Inf.* **37** 132–9

[4] Lin Y, Wang K, Liu H, Xu K, Ding K and Sun X 2020 Application of Network Representation Learning in the Prediction of Scholar Academic Cooperation *J. China Soc. Sci. Tech. Inf.* **39** 367–73

[5] Dong Y, Chawla N V and Swami A 2017 metapath2vec: Scalable Representation Learning for Heterogeneous Networks *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* KDD '17: The 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (Halifax NS Canada: ACM) pp 135–44

[6] Shi Y, Gui H, Zhu Q, Kaplan L and Han J 2018 AspEm: Embedding Learning by Aspects in Heterogeneous Information Networks *ArXiv180301848 Cs*

[7] Lee S, Park C and Yu H 2019 BHIN2vec: Balancing the Type of Relation in Heterogeneous Information Network *Proc. 28th ACM Int. Conf. Inf. Knowl. Manag.* 619–28

[8] Behrouzi S, Shafaeipour Sarmoor Z, Hajsadeghi K and Kavousi K 2020 Predicting scientific research trends based on link prediction in keyword networks *J. Informetr.* **14** 101079

[9] Teklu) (Urgessa, Joon) (Kim Min and Seek) (Lee Joong 2017 Exploring Scientific Knowledge Landscape in User Experience/Human Computer Interaction using Author Defined Keyword Network Analysis. **16** 77–101

[10] Lu W, Huang S, Yang J, Bu Y, Cheng Q and Huang Y 2021 Detecting research topic trends by author-defined keyword frequency *Inf. Process. Manag.* **58** 102594

[11] Lu W, Liu Z, Huang Y, Bu Y, Li X and Cheng Q 2020 How do authors select keywords? A preliminary study of author keyword selection behavior *J. Informetr.* **14** 101066

[12] Wei X and Lin C 2014 The Research of Science Collaboration Behavior Based on Author-Year-Keyword Network ——The Case of the Library and Information Science *J. Intell.* **33** 117–23

[13] Grover A and Leskovec J 2016 node2vec: Scalable Feature Learning for Networks *ArXiv160700653 Cs Stat*

[14] Tang J, Qu M, Wang M, Zhang M, Yan J and Mei Q 2015 LINE: Large-scale Information Network Embedding *Proc. 24th Int. Conf. World Wide Web* 1067–77

[15] Wang D, Cui P and Zhu W 2016 Structural Deep Network Embedding *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* KDD '16: The 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (San Francisco California USA: ACM) pp 1225–34

[16] Dong Y, Chawla N V and Swami A 2017 metapath2vec: Scalable Representation Learning for Heterogeneous Networks *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* KDD '17: The 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (Halifax NS Canada: ACM) pp 135–44

[17] Fu T, Lee W-C and Lei Z 2017 HIN2Vec: Explore Meta-paths in Heterogeneous Information Networks for Representation Learning *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management* CIKM '17: ACM Conference on Information and Knowledge Management (Singapore Singapore: ACM) pp 1797–806

[18] Cen Y, Zou X, Zhang J, Yang H, Zhou J and Tang J 2019 Representation Learning for Attributed Multiplex Heterogeneous Network *Proc. 25th ACM SIGKDD Int. Conf. Knowl. Discov. Data Min.* 1358–68

[19] Roy P K, Chowdhary S S and Bhatia R 2020 A Machine Learning approach for automation of Resume Recommendation system *Procedia Comput. Sci.* **167** 2318–27

[20] Bhalse N and Thakur R 2021 Algorithm for movie recommendation system using collaborative filtering *Mater. Today Proc.*

[21] Sejal D, Ganeshsingh T, Venugopal K R, Iyengar S S and Patnaik L M 2016 Image Recommendation Based on ANOVA Cosine Similarity *Procedia Comput. Sci.* **89** 562–7

[22] Takagi Y, Sakai Y, Abe Y, Nishida S, Harrison B J, Martínez-Zalacaín I, Soriano-Mas C, Narumoto J and Tanaka S C 2018 A common brain network among state, trait, and pathological anxiety from whole-brain functional connectivity *NeuroImage* **172** 506–16

[23] Abe Y, Takata N, Sakai Y, Hamada H T, Hiraoka Y, Aida T, Tanaka K, Bihan D L, Doya K and Tanaka K F 2020 Diffusion functional MRI reveals global brain network functional abnormalities driven by targeted local activity in a neuropsychiatric disease mouse model *NeuroImage* **223** 117318