Construction Method and Feature Analysis of Correlation Community Model for Scientific Researchers

Jin Lian[1,a], Ming Gao[2,b*], Junfeng Chen[1,c]

[1]School of Artificial Intelligence, Jianghan University, Wuhan, Hubei, China
[2]Office of Network and Digital Construction, Jianghan University, Wuhan, Hubei, China

[a]lianjin@jhun.edu.cn, [b*]gaoming@jhun.edu.cn
[c]1693231891@qq.com

Abstract. This article analyzes some scientific research papers and author data of a certain organization, and uses the reachability matrix to construct a research community correlation model. The research community correlation model is a data model used to analyze the correlation and cooperation of a group of scientific researchers within an organization. From the establishment process and structure of the model, it is possible to obtain the cooperation between scientific researchers within the organization and its external cooperation. Based on the analysis of data in the past three years, the article obtains information such as the research team, research achievements, and internal correlation strength of each department, and analyzes the trend of changes and the problems that the data model can reflect. At the same time, the data structure of this model supports more dimensional statistics and calculations, effectively carrying out statistical analysis functions of traditional personal data portraits.

Keywords: reachable matrix, association, community.

1 Introduction

The level of university research is an important indicator for evaluating the quality of universities[1]. How to use data, analyze the overall situation of university scientific research, predict the development trend, identify problems in time, and promote the development of university scientific research in a reasonable and effective way is a very worthwhile research topic. The application of big data can effectively improve productivity for university scientific research, especially interdisciplinary scientific research[2]. In the analysis of university scientific research data, the daily summarized data, statistical data, and categorical data are studied more, while the research on scientific research authors and the scientific research process itself is relatively less.

Scientific research activities, the essence of which is still human activities, is the activities of groups and organizations, starting from the cluster of scientific research and extracting the basic characteristics for analysis is an entry point for the study of scientific research activities itself [3]. The study of group behavior has important theoretical and applied value[4]. Group activities have general characteristics, the

Y. Feng et al. (eds.), *Proceedings of the 4th International Conference on Internet, Education and Information Technology (IEIT 2024)*, Atlantis Highlights in Social Sciences, Education and Humanities 26,
https://doi.org/10.2991/978-94-6463-574-4_79

particular characteristics of scientific research groups, are the basic direction of the study.Robbins believes that there is a big difference between groups and teams[5], scientific research groups around the relevant theme of information exchange, integration[6], inherent organization, is not a sufficient factor to constitute the scientific research community, and has a pattern of solidification, division of labor alienation[6], and the phenomenon of gradual decrease in efficiency. Researchers group portrait, help to find potential cooperation object and research hotspot, promote knowledge exchange and innovation[7], the natural community construction of researchers, is a cut of the research community portrait.

2 Status of Research

With the increasing specialization and complexity of scientific research tasks, research teamwork and group communication are increasingly important[8]. The management of research teams, including members, management organizations, and management methods, has always been a hot research topic about scientific research[9]. The use of data to establish the portrait of research workers is the core of the research, and the basic information, keywords, research achievements and other information is more thoroughly analyzed[10]. The methods of group behavior research are also increasingly rich, and the behavioral cluster analysis for various types of feedback data has achieved certain results in the field of E-commerce.

In group analysis such as research team portrait construction based on heterogeneous data from multiple sources, and researchers' group portrait construction method of acquaintance and interest, there are complex data sources, models in the form of line graphs and keyword visualization, and insufficient research on comparison between associated groups, statistical analysis of data, and trend analysis. In this paper, based on the Coda Matrix, clear data model with matrix diagrams, statistical curves and other methods, better expresses the achievements of the community, the connection within the community, the growth of the community and the development trend and other issues.

3 Theoretical Model

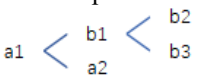
Define a thesis dataset G, that contains the authors of each thesis in the thesis data as the author dataset A, then the dataset of A1={a1, a2, ,a3……}, A2={b1, b2, ,b3……}. Authors can reach each other within the same paper, while different authors with different purposes cannot reach each other, record as P(A)=[Pij]n*n, among other:$Pij = \begin{cases} 1 & \text{If vj is reachable from vi} \\ 0 & \text{If vj is not reachable from vi} \end{cases}$.The author association matrix is constructed through the thesis data set. Because each author to its own reachable, but for the study of this paper is not meaningful, so set Pij = 0. Through the matrix A1 + A2, you can get the authors in A1 and the authors in A2 association relationship.

An example of the process of calculating the author association relation reachability matrix is given below:

Paper 1 contains authors {a1, a2, b1}, paper 2 contains authors {b1, b2, b3}, where b1 is a common element and generates a 5th order matrix of {a1, a2, b1, b2, b3}, i.e.:

$$\begin{bmatrix}0&1&1\\1&0&1\\1&1&0\end{bmatrix}+\begin{bmatrix}0&1&1\\1&0&1\\1&1&0\end{bmatrix}=\begin{bmatrix}0&1&1&0&0\\1&0&1&0&0\\1&1&0&1&1\\0&0&1&0&1\\0&0&1&1&0\end{bmatrix}$$

By gradually increasing the number of matrices, adding them together to form higher-order matrices, and recording their reachable correlation relationships. Further traversing the correlation tree of the 5 * 5th order matrix yields:



Create an author association set for dataset Ci: {a1, a2, b1, b2, b3}.

If there is no correlation between the two matrices, paper 1 contains authors {a1, a2, a3}, and paper 2 contains authors {b1, b2, b3}, then generate a 6th order matrix of {a1, a2, a3, b1, b2, b3}, that is:

$$\begin{bmatrix}0&1&1\\1&0&1\\1&1&0\end{bmatrix}+\begin{bmatrix}0&1&1\\1&0&1\\1&1&0\end{bmatrix}=\begin{bmatrix}0&1&1&0&0&0\\1&0&1&0&0&0\\1&1&0&0&0&0\\0&0&0&0&1&1\\0&0&0&1&0&1\\0&0&0&1&1&0\end{bmatrix}$$

It can be seen that the connectivity is 0, resulting in two independent tree like structure sets Ci: { a1, a2, a3 }、Ci+1:{ b1, b2, b3 }. Define the overall correlation matrix R. If R is the nth order square matrix of all author elements contained in dataset G, then: R = A1+A2+A3+…An . By traversing the tree data of matrix R, a basic correlation dataset based on scientific research paper data can be obtained for researchers.

The basic association dataset loses a large amount of attribute data, and as the number of authors increases, the connection diagram becomes chaotic and difficult to analyze the existence of the community. Meanwhile, in terms of the amount of information available for statistical analysis, the amount of data that can be observed and analyzed is also very limited. To solve this problem, the model adds a multidimensional attribute data model U[row][col], which has the following structure:

Table 1. Data Structure

[Author's Name]	[Correlation value]	[Research Credits]	[department1]
[Author's ID card]	[contribution]	[Thesis discipline]	[department2]
[Author's number]	[main character]	[Impact factors]	[department3]
[Other information]	[NULL]	[Achievements]	[……]

Table 1 presents the data structure of the model. The main parameters of this multidimensional data model include correlation value, contribution value, scientific re-

search integral, subject and unit attributes. Among them, the author's job number, unit, and basic personnel database of the scientific research organization are matched to form an identity information column. Use the job number to avoid the problem of duplicate names. Use the ID number to better accommodate a larger sample space. This can also be extended to a 3D data model to record the impact values of each article.

4 Data Validation

In order to verify the validity of the model, this paper selects the research workers of the same unit and department in a university during the three years of 2020, 2021 and 2022, including the departed and new recruits. Among them, 367 research workers in 2020, published 573 research papers of various types; 375 research workers in 2021, published 611 research papers of various types; 378 research workers in 2022, published 629 papers of various types.

Selected data in the order of the collection of papers, then the points of the reachability matrix is scattered distribution, according to the reachability matrix points connecting line segments, also masked the orderliness of the community. And according to the order of individual cluster members, from large groups to small groups in order, can be obtained as shown in the figure of the matrix array, its orderliness and distribution status is clear at a glance, concise and clear. As shown in the figure below, according to the formation of community data and the reachable matrix, and then by the number of communities from the largest to the smallest arrangement, you can get the following distribution map:

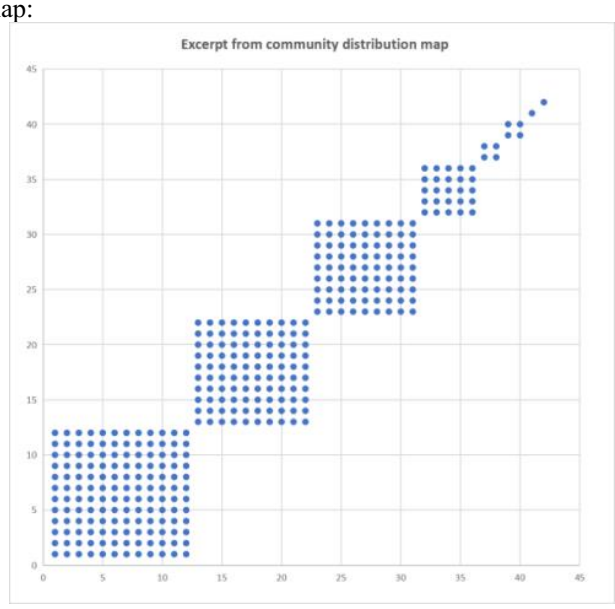


Fig. 1. Excerpt from community distribution map

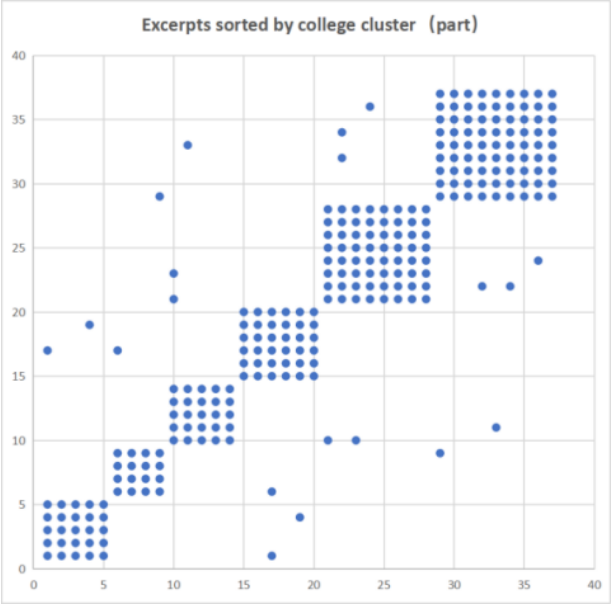


Fig. 2. Excerpts sorted by college cluster

As shown in Fig.1, excerpt from community distribution map, reachable element points are distributed symmetrically up and down with a large head and a small tail, using the coordinate angle bisector as the axis of symmetry. The base data gives a clear indication of the size of the cluster, the distribution of cluster sizes, and the number of lone researchers, converging lone researchers, whose clusters are small or only one point.

Arranging the elements of each author in the base matrix in the order of colleges (based on the first unit of each author) gives more information, as shown in Fig. 2: enumerating a number of clusters in the three colleges, and naturally plotting the reachability matrices by clusters, there are 10 correlations among the colleges. Such as (1,17), (4,19), (6,17) and so on. The correlation relations have mutual nature and while drawing the correlation diagram, bidirectional reachable correlation relations are taken. The following issues can be clearly seen from the diagram:

1. The distribution of scientific research clusters in each college, such as the three groups in College A are small in size and the matrix range is not large, so it can be seen that there is relatively little cooperation within the college. The figure can clearly see the scale of cooperation and internal association of researchers in colleges and other information.
2. With the members of each group in the college constituting the square matrix of the college researchers, the extended areas above and below, left and right of their squares are the associations between the college and other colleges. The number of points in it shows the degree of association between colleges and is the association relationship outside the college.

3. This square array can also use disciplines as horizontal and vertical coordinates in order to count the inter-discipline association and cooperation relationship, as well as the association within the disciplines. The matrix with disciplines as horizontal and vertical coordinates can highlight the scientific research in cross disciplines.

In terms of statistical dimensions, we mainly examine the information of cluster size, number of clusters, number of articles, strength of association, scientific research score, and article category. Among them, because the university paper score assessment is more complex, and this paper to study its trend and relevance of the main, so simplify the calculation process, the scientific research score of various types of papers to take the average value of the grade, the data set without the T class papers (high level papers), A class 120 points (core class), B class 40 points (general journals), and C class 10 points (open publications). The simplified data, which do not affect the performance of its characteristics, meet the research needs of this paper. The statistical data table is as follows:

Table 2. Analysis of research clusters and connection strengths, research outputs in 2022

Group sequence	Community size	Number of communities	Total number of people	Number of articles/		Correlation strength coefficient		Scientific research score		Number of articles in ABC category		
				Per capita article count		(Progressive/ Average)		(Progressive/ Average)				
1	31	1	31	57	1.8387	114	3.6774	2250	72.5806	6	34	17
2	29	1	29	56	1.9310	93	3.2069	2040	70.3448	5	31	20
3	25	1	25	33	1.32	55	2.2	1310	52.4	4	18	11
4	22	1	22	40	1.8182	72	3.2727	1600	72.7273	3	29	8
5	19	2	38	77	2.0263	135	3.5526	2810	73.9473	6	46	25
6	18	1	18	33	1.8333	46	2.5556	1050	58.3333	3	13	17
7	14	2	28	49	1.75	74	2.6429	1800	64.2857	4	29	16
8	13	1	13	16	1.2308	28	2.1538	560	43.0769	2	6	8
9	12	2	24	42	1.75	66	2.75	1500	62.5	3	25	14
10	11	2	22	36	1.6364	49	2.2273	970	44.0909	2	13	21
11	10	3	30	47	1.5667	64	2.1333	1470	49	2	26	19
12	9	4	36	74	2.0556	111	3.0833	2330	64.7222	3	42	29
13	8	1	8	6	0.75	11	1.375	240	30	0	6	0
14	7	3	21	22	1.0476	36	1.7143	750	35.7143	1	14	7
15	6	1	6	7	1.1667	8	1.3333	250	41.6667	0	6	1
16	5	2	10	12	1.2	18	1.8	440	44	1	7	4
17	3	3	9	14	1.5556	13	1.4444	290	32.2222	0	5	9
18	2	1	2	2	1	2	1	50	25	0	1	1
19	1	6	6	6	1	0	0	180	30	0	4	2
Total		38	378	629	1.6640		0	21890	57.9101	45	355	229

From the data in Table 2, it can be seen that the more the number of articles, the stronger the association, and in the case of the same number of articles, or the average number of articles is not much different, the strength of association is strong, which indicates that the number of authors of the paper has increased. The level of association strength is also related to the level of the published articles, with high level, the association strength is relatively high, and with low level, the association strength is also low. For example, for core papers, the average authors are around five, while for average papers, the authors are two to three. At the same time, there is also a close relationship between the association strength and the average scientific research score, with small clusters having low association strength and low average scientific research scores. This is particularly evident in the third and seventh data sets.

In the data used in this paper, the large clusters are from scientific research institutions and units related to computer and artificial intelligence technology, and compared with other clusters, they are characterized by high scientific research efficiency, high quality, and closer cooperation. The small clusters, on the other hand, are significantly lower in quantity and quality than the large clusters, on the one hand, the personnel in the small clusters are mostly teaching and researching and heavy, and the cooperation is not sufficient. Papers from the same faculty and specialty, about hot research, are not related to each other, which can be shown from the keywords and the content of the papers.

According to the data analysis of multidimensional attributes and assignment weights research association reachable matrix, it can be seen that the same number of people in the cluster, on the number of articles, quality, average workload, the degree of research connection is different, and there is a certain positive correlation during the period.

The graphs of cluster development trend, cluster size vs. group association statistics, and association vs. average scientific research achievement statistics (2020 to 2022) are listed below:

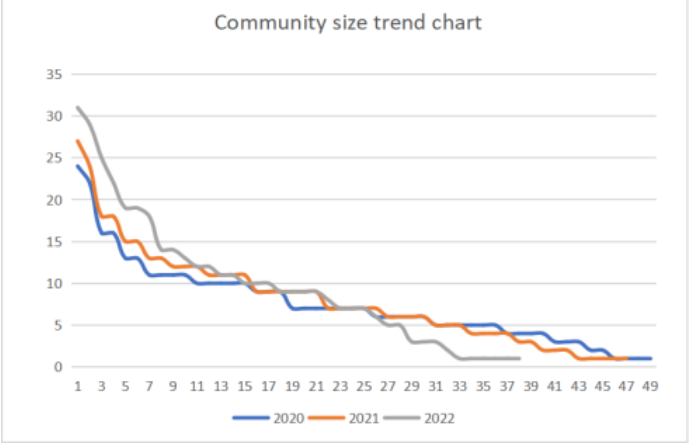


Fig. 3. Community size trend chart

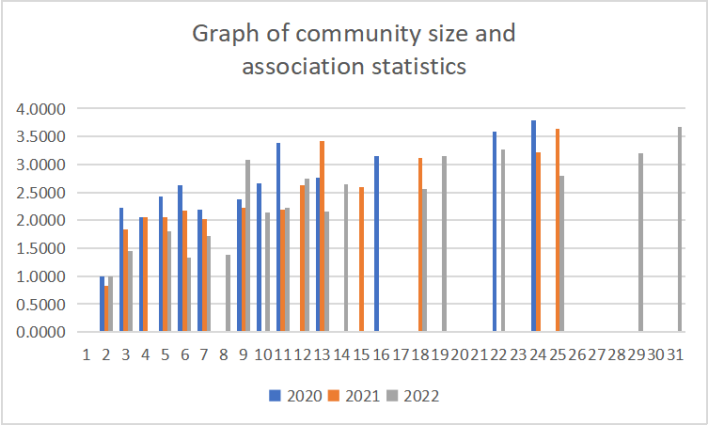


Fig. 4. Graph of community size and association statistics

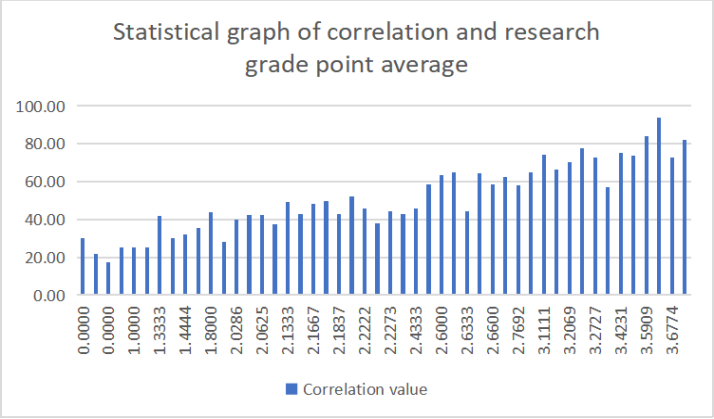


Fig. 5. Statistical graph of correlation and research the number of grade point average

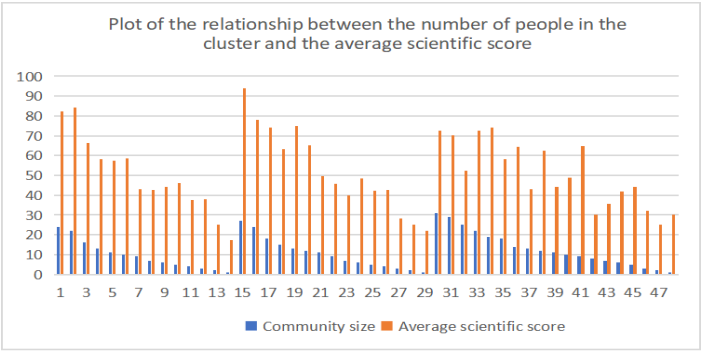


Fig. 6. Plot of the relationship etween people in the cluster and the scientific score

Figure 3 shows that between 2020 and 2022, research groups are gradually increasing and the number of research clusters is decreasing; Figure 4 shows that the larger the number of people in a research group, the higher the degree of relatedness of its researchers, there is a certain degree of chance, but the overall trend is obvious; and Figure 5 shows, more clearly, the relationship between the degree of relatedness and the scientific achievement; the higher the degree of relatedness, the richer the results of its scientific research. Figure 6 shows that research groups with large clusters have better research achievements. From the multi-dimensional statistical graphs, it can be seen that the larger the research group cluster identified by the model, the better its relevance and scientific research achievements, and the development of the size of an organization's research cluster is positively correlated with the development trend of the organization's scientific research.

Table 3. Comparison of three years of data for the same subject community

Year	Community size	Differential numbers	No. of papers published	Community association strength	Average strength	Scientific research score	Average scientific research score	High level thesis
2020	24	0	46	91	3.7917	1970	82.08	5
2021	27	7	62	98	3.6296	2540	94.07	6
2022	31	9	57	114	3.6774	2250	72.58	6

Similarity operations were performed on the clusters to easily obtain data for the clusters over the years. As shown in Table 3. analyzing the data for one of the clusters, it can be obtained that the cluster grows from 24 people in 2020 to 31 people in 2022. There are some differentiated people in its cluster, with new people entering and original members exiting. And the number is about one-third of the population, which is a big difference. For the association strength and scientific research score can be seen, the overall performance of the large cluster is better than that of the small cluster, but the data of 2022 shows that although the number of people increased, the association strength and scientific research score are obviously inferior to that of 2021, which indicates that the scientific research of this cluster has a large fluctuation.

5 Conclusions

The study of groups of research workers is a direct observation of the productivity, stability, and state of production of research work. This paper summarizes the significance by organizing and computing the data as follows:

1. Being able to understand the community size, research content, activity frequency, research collaboration (coupling degree), and achievement quality of the internal research group in universities, and observe the research status of the school from another perspective.

2. For the introduction of talents in universities, the contribution of talents to the growth of universities can be examined. An individual, especially an individual who is in a research cluster outside the university, has limited contribution to the growth of the university itself.
3. The model focuses on the presentation of group performance and trend prediction, so that the model to a certain extent to avoid the error caused by the performance of individuals, or the performance of a few people to flatten the average value. A more in-depth study of the group effect of scientific research work has practical value for the overall state, trend and policy orientation of university scientific research work.

This paper is an effective practical research by modeling the cluster of college teachers' thesis data and teachers' basic information data in order to understand the scientific research of the school and individuals from a tangential perspective. This modeling study, there are still great limitations in terms of the amount of data, data dimensions, only to put forward a different perspective, how to make this perspective richer, more perfect, and achieve better results, is the main content of further research.

Acknowledgments

Scientific Research Foundation, Hubei Provincial Natural Science Foundation Project, No.2022CFC041.

Scientific Research Foundation, China University Industry University Research Innovation Fund--New generation information technology innovation project, No.2021ITA03010.

Scientific Research Foundation, Jianghan University Campus level Science and Technology Special Research Projects, No.2023KJZX17.

References

1. Wei ZG, Zheng KD. Reflections on the assessment of scientific research level in colleges and universities [J]. Journal of Anhui University of Technology: Social Science Edition, 2021, 38(5):112-114.
2. Lu Genshu. The application of big data in higher education and the challenges it faces[J]. Chongqing Higher Education Research, 2022, 10(4):31-38.
3. Wang Mengqian, Fan Yizhou, Guo Wenge, Wang Qiong. A review of research on clustering analysis of MOOC learners' characteristics[J]. China Distance Education, 2018(7), 9-19.
4. Lanping Feng, Fengping Wu. Research on ontology credibility of collaborative construction based on group behavior[J]. Journal of Intelligence, 2015, 34(6):163-168.
5. ROBBINS P S, COULTER M. Management [M]. New Jersey:Prentice Hall, 2012: 370 -401.
6. ZHANG Hongbo, ZHAO Xiaoning. Discussion on the model of scientific research group and scientific research team[J]. Science-Talent-Market, 2023(5):25-27.

7. XU Wei, DOU Yongxiang, LI Weibo. A group portrait construction method for researchers considering similar interests[J]. Intelligence Theory and Practice, 2021, 044 (011): 166-172.
8. Mo Junlan, Dou Yongxiang, Kai Qing. Construction of research team portrait based on multi-source heterogeneous data[J]. Intelligence Theory and Practice, 2020, 43 (9): 100-106.
9. Pei Li-Shen, Zhao Xue-Zhu. A review of research on deep learning methods for group behavior recognition[J]. Computer Science and Exploration, 2022, 16(4):775-790.
10. Li Feng. Cluster analysis based on core keywords--Another discussion on the insufficiency of co-word cluster analysis[J]. Intelligence Science, 2017(8), 35(8):68-78.