Research on Subject Hot Spots Based on Keyword Co-occurrence Analysis and Social Network Analysis——Taking Economics for Example

Linrui Li*

School of Economics and Management NUST, Nanjing University of Science and Technology, Nanjing, China

*Corresponding author: 15822737061@163.com

Abstract: [Method / Process] In this paper, the literatures in the field of economics from 2005 to 2019 in SSCI database are used as the data source, and the literature keywords in the dataset are analyzed based on the co-occurrence analysis and social network analysis methods. [Purpose / Significance] The purpose of this study is to understand the research hot spots in the field of economics in the past 15 years. On the one hand, it can provide reference for researchers in the field, and on the other hand, make the economic research results better serve the society. [Result / Conclusion] Based on the drawing of keyword co-occurrence network and the analysis of co-occurrence network by social network method, eight hot issues in the field of economics in recent 15 years are obtained by clustering, and these hot spots are summarized.

1. Introduction

Subject hotspots are issues that have a high degree of attention in academia within a certain period and have received more research. Understanding the subject hotspots helps field researchers grasp the research trends and focus of the field, and can be used for scientific research planning, topic selection, demonstration, and provide important basis for evaluation. This research takes economics as an example which has relatively obvious characteristics of discipline schools and takes the keywords of academic papers published in important journals in this field from 2005 to 2019 in Web of Science as the data source. Then on the basis of keyword co-occurrence analysis and social network analysis, the subject hot spots are found.

2. Data Sources and Preprocessing

2.1 Data Sources

The research collects data of academic papers published in important journals in the field from 2005 to 2019 from Web of Science. Specifically, select the database as SSCI in the data collector, set WC (WOS category) = economics, and set the document type to: Article or Review, the time range is from 2005 to 2019. Finally, a total of 248128 literatures were obtained. Based on the obtained data sets, the subject hot spots are explored.

2.2 Data Preprocessing

Standardized processing of keywords: The initial case of the same keyword may be different in different articles, so the keywords are uniformly treated in lowercase, and the literature with the keyword field of "None" is excluded. Finally, a total of 216473 keywords are obtained.

3. Analysis Methods

This study mainly uses co-occurrence analysis of keywords and social network analysis methods. In keyword co-occurrence analysis, the complex network analysis package Networkx is imported into

Python to calculate the degree of each node in the co-occurrence network. By doing so, the nodes with high centrality in the network are determined preliminary. The modular keyword co-occurrence network is drawn by using social network analysis tool Gephi, so as to form a preliminary judgment on the research hotspots. To go further, based on social network analysis, the network structure is analyzed by using the social network analysis software UCINET. Nodes are analyzed quantitatively by the connection between nodes, and clustering is carried out, and the research hotspots in economics in recent 15 years are obtained.

3.1 Word Frequency Analysis

Word frequency analysis is the quantitative statistics of the occurrence times of each keyword that can represent the content and theme of the literature. By understanding the frequency of the keywords, we can have a preliminary intuitive judgment of the hot spots in the research field.

3.2 Keyword Co-occurrence Analysis

Keyword co-occurrence refers to the phenomenon that key words appear together in more than two literatures. Keyword co-occurrence analysis is a quantitative analysis of keyword co-occurrence, with the purpose of revealing the related information content and the hidden knowledge of feature items [1]. The co-word network can be used to determine the relationship between subjects, reveal the hot research direction and research frontier of specific topics.

3.3 Social Network Analysis

Social network refers to the collection of social actors and their relationships. Social network analysis (SNA) is a quantitative analysis method developed by sociologists based on mathematical methods and graph theory [2]. The introduction of social network analysis method in information science greatly broadens the perspective of information analysis work and can effectively discover the characteristics of keywords in each node of keyword co-occurrence network and the relationship between each node [3]. Social network analysis can be carried out from the following aspects:

3.3.1 Network Centrality Analysis

Centrality represents the central position of members in the social network. The higher the centrality, the more important its role in the network. The study of centrality is generally divided into three indicators: degree centrality, betweenness centrality and proximity centrality. In this paper, the degree centrality and the betweenness centrality are calculated for the keyword co-occurrence network. The degree centrality represents the total number of times that the node and other nodes co-occur. The higher the degree centrality, the stronger the centrality of the keyword. If a node is in the path between many other two points, it is considered that the point has a higher betweenness centrality, and it has the ability to control the communication between the other two actors. The more an actor occupies such a position in the network, the more it represents its high centrality in the middle, and the more actors need to connect through it [4]. It is an indicator to measure the ability of a keyword in the network to influence other keywords to co-occur in a journal paper.

3.3.2 Cohesive Subgroups Analysis

In a set of actors, there are relatively strong, direct, close, frequent or positive relationships between actors, so that a secondary small group is formed. Such groups are called cohesive subgroups in social networks [5]. For keywords, keywords in the same subgroup have strong relationship with each other. In social network analysis, we can analyze cohesive subgroups from different aspects. In this paper, K-Core analysis is used to analyze cohesive subgroups based on degree. The cohesive subgroups based on degree are obtained by limiting the number of adjacent points of each member in the subgroup, that is, the number of co-occurrence between keywords. K-Core means that all nodes in a subgraph are connected to less than k other points of the subgraph. Through the K-Core analysis, we can determine the meaningful cohesive subgroups in the keyword network and find the isolated points in the network.

3.3.3 Positions and Roles in the Network:

By investigating and calculating the peer relationship model of member structure in the network, we can analyze the similarity between various actors [6]. CONCOR is a method to calculate the positions and roles of network members in UCINET software. It is a correlation iterative convergence method, which measures the similarity between pairs of actors by Pearson product distance coefficient, so as to determine how many "blocks" or "clusters" exist in the network.

4. Data Analysis and Results Discussions

4.1 Word Frequency Analysis

The word frequency of the standardized keywords is counted and arranged in descending order according to the word frequency. In this study, the keywords with the top 200 frequencies are used as the high-frequency keywords to be analyzed. The frequency table of some keywords are shown in TABLE I.

We can see that the word "china" (Due to the unified lowercase processing, that is China) with the highest frequency. With the rapid development of China's economy in recent years, China and its development situation have become more and more concerned by economists of various countries. In addition, through the high-frequency words such as "economic growth", "innovation" and "inequality", we can preliminarily determine the focus of attention in the field of economics in recent years.

4.2 Keyword Co-occurrence Analysis

The construction of keyword co-occurrence matrix is the basis of constructing keyword co-occurrence network and using UCINET to analyze social network. Python is used to construct the keyword co-occurrence matrix of 201 * 201 according to the determined high-frequency keywords.

4.2.1 Node Centrality Judgment

In this paper, we import the social network analysis package Networkx in Python, set the threshold value of co-occurrence times as 10, and calculate the out degree of each node in the co-occurrence network. The out degree is the number of nodes directly connected with the node. The higher the keyword out degree is, the higher the correlation degree between the keyword and other keywords in the network, and it can be preliminarily considered that the node has a high centrality. The output degree of some nodes in the network is shown in TABLE II. It can be seen that "economic growth", "inequality" and "poverty" have the highest degree of output. It can preliminarily show that these issues have a high degree of concern in the field of economics and are easily mentioned by different research issues.

Table 1. Keyword Frequency Statistics Table (partial)

Keyword	frequency	Keyword	frequency
china	3591	education	1273
economic growth	2714	regulation	1223
monetary policy	1982	unemployment	1218
innovation	1803	institutions	1189
panel data	1613	india	1133
inequality	1477	financial crisis	1099
climate change	1470	poverty	1092
human capital	1455	inflation	1083
productivity	1441	gender	1082
growth	1381	cointegration	1003
efficiency	1359	entrepreneurship	1001
uncertainty	1291	forecasting	976

Table 2. Out Degree of Partial Keyword Node

Keyword	Degree	Keyword	Degree
economic growth	34	education	11
inequality	19	development	11
poverty	18	gender	11
growth	18	unemployment	10
productivity	16	monetary policy	9
innovation	14	inflation	9
human capital	12	panel data	8
institutions	12	employment	8

4.2.2 Preliminary Judgment of Sub Network

Gephi is a visualization software for exploratory data analysis. The data is imported into Gephi, and the network nodes are modularized. The main function of modular statistics is to cluster each node. According to the correlation between nodes, the whole network is divided into several modules, and different modules are represented by different colors. Set the threshold of co-occurrence times to 10 times to get the keyword modular co-occurrence network, as shown in Fig. 1. The larger the node in the graph, the higher the correlation between the node and other nodes in the network, and the more important position in the network. As can be seen from the figure, "economic growth" is an important connection point of each sub network, and the whole network has formed a divergent trend with "economic growth" as the core, forming different sub networks. It can be preliminarily explained that economic growth is always concerned in various fields of economics or economic growth is the background of economic research problems, which is consistent with the keyword node out degree analysis above.

In order to clearly reflect each sub network in the network, each sub network is displayed separately. The network can be divided into six sub networks, as shown in Fig. 2.

Further detailed to each sub network, we can draw a preliminary analysis that in the past 15 years, the research in the field of economics can be divided into six categories. With only one sub network's central words are not obvious, while the central words of the other categories are monetary policy, productivity, inequality and poverty, economic growth and institutions.

Gephi's visualization results are only exploratory analysis of each small group in the keyword co-occurrence network, and then use the social network analysis tool UCINET to carry out quantitative statistical analysis.
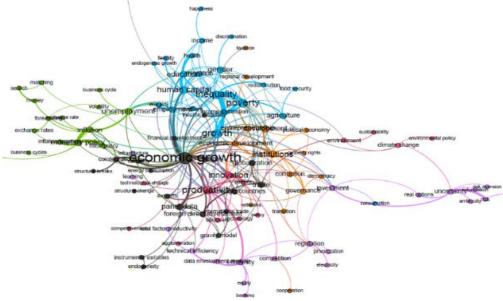

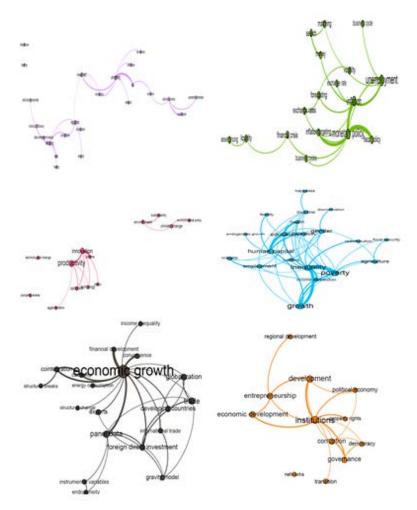Figure 1. Keyword co-occurrence network based on Modularization.

Figure 2. Sub networks.

4.3 Social Network Analysis

4.3.1 Centrality Analysis

UCINET is used to calculate centrality of each node. Some results are shown in Fig. 3(the upper part is degree centrality and the below part is betweenness centrality). The average degree centrality is 154.567, and there are 61 nodes higher than the average. Comparing the top20 keywords of degree centrality with the top20 of frequency, the two groups of keywords are basically consistent, and the two high-frequency keywords with obvious difference are "China" and "India". The degree centrality of the two is 5 and 0 respectively, which indicates that although "China" and "India" are frequently mentioned countries in economic research, they are not the center of the whole research network. Maybe they are more mentioned as the research background of a certain kind of national economic problems, which is consistent with the fact that economics itself studies economic problems rather than countries.

As for the betweenness centrality, it can be seen that "growth" has the highest betweenness centrality, that is, many nodes achieve co-occurrence through "growth"[7]. More attention is paid to the keyword "climate change" in the middle centrality, indicating that more and more attention has been paid to the possible impacts of global climate change.

4.3.2 Cohesive Subgroups Analysis

K-Core Analysis: We can analyze some "small groups" in the keyword co-occurrence network according to the calculation of K-Core, to roughly judge the keywords other than "small groups" in the whole network as relatively isolated points. Based on the co-occurrence matrix of keywords, the K-Core analysis using UCINET can roughly select the keywords that can be clustered and remove the

loose relationship of the noisy keywords. The results of K-Core analysis are shown in Fig. 4.It can be seen from the figure that 25 K values are obtained by K-Core analysis of 200 candidate keywords, including 1,2,4, 6, which means that the keyword co-occurrence network can be composed of 25. However, it is easy to find that there are many keywords which have loose relationship with other keywords and cannot form a small group with other keywords, such as (201) Vietnam, (184) Indonesia, etc., but the isolated points cannot be directly excluded, and the isolated points with high centrality can also indicate a certain degree of heat. The keywords with centrality less than 50 were eliminated, and 159 keywords with more clustering value were obtained.

CONCOR Analysis: Using UCINET's CONCOR to cluster the keywords obtained after K-Core analysis and processing. Eight clusters and the density matrix of each cluster are obtained. The top10 keywords in each category are shown in TABLE III.

The density matrix of clustering results is shown in Fig. 5. The value on the diagonal represents the density within each class. The larger the value, the closer the relationship among the members of the class is, that is, the better the clustering effect is; the value outside the diagonal represents the density between the clusters, and the larger the value, the stronger the coupling between the categories.

		1 Degree	2 NrmDegree	3 Share			1 Betweenness	2 nBetweenness
2	economic growth	1086.000	4.641	0.035	10	growth	357.225	1.795
10	growth	780.000	3.333	0.025	7	climate change	334.840	1.683
6	inequality	773.000	3.303	0.025	2	economic growth	318.354	1.600
4	innovation	618.000	2.641	0.020	5	panel data	312.277	1.569
9	productivity	616.000	2.632	0.020	12	uncertainty	289.062	1.453
19	poverty	576.000	2.462	0.019	4	innovation	284.035	1.427
16	institutions	557.000	2.380	0.018	15	unemployment	249.276	1.253
8	human capital	555.000	2.372	0.018	146	environmental policy	248.088	1.247
13	education	507.000	2.167	0.016	14	regulation	211.127	1.061
5	panel data	492.000	2.103	0.016	11	efficiency	209.577	1.053
3	monetary policy	431.000	1.842	0.014	113	electricity	204.386	1.027
20	inflation	416.000	1.778	0.013	35	risk	202.946	1.020
11	efficiency	411.000	1.756	0.013	46	agriculture	193.176	0.971
53	development	409.000	1.748	0.013	16	institutions	192.758	0.969
15	unemployment	402.000	1.718	0.013	51	globalization	187.039	0.940
12	uncertainty	368.000	1.573	0.012	34	welfare	178.086	0.895
31	economic development	365.000	1.560	0.012	30	investment	174.794	0.878
21	gender	364.000	1.556	0.012	26	competition	174.749	0.878
51	globalization	360.000	1.538	0.012	9	productivity	173.237	0.871
32	employment	360.000	1.538	0.012	13	education	164.672	0.827
14	regulation	348.000	1.487	0.011	18	financial crisis	150.939	0.758
40	trade	346.000	1.479	0.011	21	gender	145.797	0.733
25	foreign direct investment	332.000	1.419	0.011	43	learning	143.654	0.722
35	risk	323.000	1.380	0.010	6	inequality	130.932	0.658
30	investment	321.000	1.372	0.010	87	information	130.681	0.657
46	agriculture	316.000	1.350	0.010	31	economic development	127.450	0.640
39	migration	304.000	1.299	0.010	32	employment	125.416	0.630
34	welfare	291.000	1.244	0.009	33	sustainability	124.121	0.624
23	entrepreneurship	290.000	1.239	0.009	40	trade	118.967	0.598
47	health	281.000	1.201	0.009				
57	developing countries	277.000	1.184	0.009				
22	cointegration	277.000	1.184	0.009				
7	climate change	268.000	1.145	0.009				

Figure 3. Degree centrality and betweenness centrality of part keywords.

Figure 4. the result of K-Core analysis(partial).

```
Density Matrix

           1       2       3       4       5       6       7       8
         _____   _____   _____   _____   _____   _____   _____   _____
    1    3.218   1.564   1.359   1.439   1.093   0.452   0.558   0.394
    2    1.564   2.539   0.511   1.332   0.717   0.248   0.434   0.517
    3    1.359   0.511   5.697   1.328   0.667   2.364   0.522   0.344
    4    1.439   1.332   1.328   4.251   0.434   0.207   0.352   0.402
    5    1.093   0.717   0.667   0.434   2.733   0.903   1.058   0.369
    6    0.452   0.248   2.364   0.207   0.903   2.582   0.497   0.158
    7    0.558   0.434   0.522   0.352   1.058   0.497   2.286   0.902
    8    0.394   0.517   0.344   0.402   0.369   0.158   0.902   2.543

R-squared = 0.086
```

Figure 5. CONCOR clustering density matrix.

Based on the principle of "high cohesion and low coupling", the density between clusters should be small. The overall network density can be calculated as 1.184 by UCINET. The clustering density of each diagonal of the cluster density matrix is compared with the overall network density. When the density is greater than the overall network density, the clustering effect is better. The clustering density of the eight clustering modules are 3.218, 2.539, 5.697, 4.251, 2.733, 2.582, 2.286, 2.543, which are all bigger than the overall network density. The clustering result is good and can be adopted completely.

Table 3. The Result of CONCOR Analysis

Class number	TOP 10 Keywords
1	economic growth, total factor productivity, innovation, panel data, competitiveness, exports, externalities, productivity, gravity model, efficiency
2	transaction costs, social networks, democracy, taxation, institutions, sustainability, entrepreneurship, regional development, social capital, policy
3	heterogeneity, bootstrap, exchange rate, inflation, financial crisis, structural breaks, business cycles, monetary policy, unemployment, consumption
4	education, discrimination, employment, poverty, welfare, redistribution, propensity score matching, higher education, happiness, migration
5	regulation, capital structure, simulation, market power, data envelopment analysis, insurance, corporate governance, banks, risk, uncertainty
6	exchange rates, financial crises, emerging markets, money, volatility, asset pricing, inflation targeting, forecasting, stock returns, liquidity
7	information, moral hazard, search, quality, auctions, general equilibrium, competition, health insurance, adverse selection, mechanism design
8	voting, willingness to pay, altruism, contingent valuation, fairness, game theory, experimental economics, bargaining, experiment, public goods, experiments

Table 4. Cluster Label

Class number	Cluster label
1	Economic Growth and Sustainable Development
2	About Social
3	Economic Policy under the Financial Crisis
4	Social Livelihood Issues and Economy
5	Corporate Governance
6	Financial Crisis and Financial Market
7	Risk Assessment and Decision Management
8	Experimental Economics

The labels of the eight clusters are summarized, as shown in TABLE IV. According to the clustering results, the following conclusions can be drawn：

Economic growth and sustainable development: Economic growth is the most frequent keyword among all the keywords, indicating that economic growth has always been the most concerned issue in the field of economics. In the research of economic growth, the more concerned direction in recent 15 years can be summarized as the factors directly related to economic growth, such as production efficiency and technical factors; Besides, the sustainable development has been paid more and more attention. After the stage developing economic by exploiting resources without scruple in the last century, people begin to pay attention to the sustainable utilization of resources.

Social problems: From all aspects of social life, such as technology, politics, culture, etc., comprehensively discuss the problems faced in social development.

Economic policy under the financial crisis: The financial crisis that broke out in 2008 has a serious and sustained impact on the whole world. In this case, various regulatory policies to deal with the financial crisis have become the focus of economic discussion.

Social livelihood issues: Social and livelihood issues have always been the focus in the field of economics, such as education, employment, social welfare, etc., which are closely related to and concerned about by each of us.

Corporate governance and risk management: This area mainly includes the risk response of the company to the uncertain environment, the supervision and governance of the company itself.

Financial crisis and financial market: Different from class 3 which focuses on real economic policies, this cluster pays more attention to the impact of financial crisis on stock market and currency exchange rate.

Risk assessment and decision-making management: Information is an important factor affecting economic decision-making. Based on the psychology of risk aversion, how to use information to effectively avoid risks is the research focus in the field of economics.

Experimental economics: Experimental economics is an important branch of economics. In the past two decades, some economists, psychologists and sociologists have begun to use laboratory controlled experiments to conduct empirical tests on the theory, providing more and more experimental evidence for the previous relevant theories [8].

5. Conclusions

Through the co-occurrence analysis and social network analysis of the keywords of the journal literature in the economics from 2005 to 2019, this article clearly summarizes the hotspots of the economics research in the past 15 years, which can be summarized as macroeconomic development and microscopic social issues. In addition, through the application of co-occurrence analysis and social network analysis, it is found that the social network analysis method plays an important role in the research of hot spots in the subjects, which can be used to explore hot spots in other disciplines.

References

[1] Jia,X.N.(2019).Research on Hot Topics of Enterprise Competitive Intelligence in China Based on Keyword Co-occurrenc and Social Network Analysis . Information Research, (08),114-121. doi.

[2] Qi B.B.,Xiang X. & Zhu X.F.(2020).The Application of Social Network Method in Quantitative Analysis of Think Tank:Based on the Practice of RAND Corporation. Information and Documentation Services (04),104-112. doi.

[3] Chen Y.W. (2019). Development of Social Network Analysis (SNA) Using Intelligence Analysis. Journal of The China Society for Scientific and Technical Information (01),21-28. doi.

[4] Wei R.B. (2009). An Empirical Study of Keywords Network Analysis Using Social Network Analysis. Journal of Intelligence (09),46-49. doi.

[5] Zhu Q.H. & Li L. (2008). Social Network Analysis Method & Its Application in Information Science. Information studies: Theory & Application (02),179-183+174. doi: 10.16353/j.cnki.1000-7490.2008.02.021.

[6] Li M.,Zhang J.L.,Yin S.Q. & Wang D.W.(2015).Empirical Study on Subject Research Hotspots Based on Social Network Analysis——Taking Economic Subject as an Example. Sci-Tech Information Development & Economy (22),119-122. doi.

[7] Ji Y.L.,Tian W.J. & Dong Y.(2015).Thematic Study of Think Tank Hot Spots in China Based on Keywords Co-occurrence Analysis and Social Network Analysis. Information Science (03),108-111. doi: 10.13833/j.cnki.is.2015.03.021.

[8] Zhou Z.B. (2019). Experimental Examination on "Volunteers' Dilemma" Game of Public Goods Supply. Journal of Central University of Finance & Economics (05),74-88. doi: 10.19681/j.cnki.jcufe.2019.05.007.