Journal of Library and Information Studies 21:1 (June 2023) pp.1-16
https://doi.org/10.6182/jlis.202306_21(1).001

Revisit Girvan-Newman Algorithm for Research Topic Analysis: An Application on Library and Information Science Studies

Szu-Chia Lo[1,3], Chun-Chieh Wang[2,3]

Abstract

Research trend analysis gives the research community an essential chance to learn the past to support sustainable development. The topic of evolution analysis presents a chance to position the current research, linkages among research topics, and identify the research gap. In this study, the authors revisit a known mechanism, namely Girvan-Newman (GN) algorithm, and propose a new approach for research topic analysis. Based on the GN algorithm, author-keywords analysis approach, one-mode cluster, and duo GN algorithm analysis were suggested and applied to research topic analysis of Library and Information Science studies. The results show that the suggested approach could process major quantity materials and be able to avoid the possible distorted results gained by taking the small size of samples, or two-mode cluster, to ensure the validity of the results. The topics' hierarchy structure also suggests a different approach that could be used to deconstruct the linkages among research topics for future study.

Keywords: Research Topic Evolution; Girvan-Newman Algorithm; Library Information Science

1. Introduction

Research topic evolution analysis has always been one of the prominent issues for researchers to fully understand research trends and construct better research strategies. Different approaches are used in the study of topic evolution. Content analysis is one of the most common methods for studies on research trends. However, subject detection and labeling can take much manual work. New technologies and computerized mechanisms now offer an opportunity to largely reduce the human effort required in data analysis. Rather than manually reviewing and analyzing contents, a growing body of research applies some kind of natural language processing model, especially word embedding for topic detection. For instance, Word2vec (Mikolov et al., 2013) is a word embedding model used to study topic evolution. In addition to content analysis, bibliometrics is another computer-aided research strategy. It uses methods such as author keyword co-occurrence and citation relationship analysis to conduct similarity analysis. With the help of clustering technology, researchers can build a topic similarity network to identify research topics further.

Research trends in the Library and Information Science (LIS) field have been highly discussed in topic evolution studies. Järvelin and Vakkari

[1] Department of Library and Information Science, National Taiwan University, Taipei, Taiwan
[2] Department of Bio-Industry Communication and Development, National Taiwan University, Taipei, Taiwan
[3] Center for Research in Econometric Theory and Applications, National Taiwan University, Taipei, Taiwan
*Corresponding Author: Chun-Chieh Wang, E-mail: wangcc@ntu.edu.tw

(1990) reviewed articles that were issued in 37 LIS journals in 1985 to observe the topic evolution in LIS studies and identified ten core subjects covered: (1) the professions in the field of library and information service; (2) library history; (3) publishing (include book history); (4) education in LIS; (5) methodology; (6) analysis of LIS; (7) library and information service activities; (8) information storage and retrieval; (9) information seeking; (10) scientific and professional communication. In their work in 1993, Järvelin and Vakkari continued to examine articles issued in major LIS journals in 1965, 1975, and 1985 and discussed the changes over time. Related research later (Järvelin & Vakkari, 2022; Ma & Lund, 2021; Tuomaala et al., 2014) followed a similar research framework and reviewed articles within a single year in different time zones. Järvelin and Vakkari (2022) took the same approach; two authors analyzed 142 papers published in 1965, 449, 718, and 1,210 articles in 1985, 2005, and 2015, respectively, to view the LIS research trends across 50 years.

However, it requires much manual work, like reading contents and tagging papers. Researchers have to either devote massive amounts of time and effort to do data analysis, or compromise with a smaller dataset when revealing the shifting topic in research development. To enhance data processing efficiency, some scholars have recently adopted a topic modeling algorithm to extract semantic topics from documents. Latent Dirichlet Allocation (LDA) is one of the methods to classify texts to a certain topic. The work by Figuerola et al. (2017) conducted topic modeling analysis with LDA and analyzed 92,705 LIS articles issued from 1978 to 2014 by the articles' titles and abstracts. The algorithm allowed the authors to identify the 19 LIS topics and their changes over time. Based on a great number of documents rather than sampled works, these topics are further grouped into four major research areas. Han (2020) and Miyata et al. (2020) also applied similar LDA techniques to analyze high quantities of papers to show the research topics and their corresponding vocabulary.

Besides content analysis and topic modeling algorithms, bibliometrics techniques such as keyword co-occurrence were also used in research on topic evolution. Wang et al. (2021) took author-defined keywords as the tokens and constructed a co-keyword network, presenting the topic evolution of research in the Library Science and Computer Science field between 2014 and 2019. In the study, the authors proposed a homegrown app, NetViewer, which utilizes the Bonedel algorithm for research community detection, and presented the results with the Sankey diagram. In another research, Chang et al. (2015) adopted keyword analysis, bibliographic coupling, and co-citation analysis to uncover the research trends of LIS from 1995 to 2014.

An important issue in topic evolution analysis is the evolution of dynamic, meaning the shift of research topics over time. Due to the limitation of manual work, it is common for researchers who adopted content analysis to extract data from certain years and present the overall development. For example, Järvelin and Vakkari (2022) analyzed journal articles that were issued in 1965, 1985, 2005, and 2015 to observe the changes in LIS research topics in 50 years. This may cause concerns about result validity since details in the non-chosen years might be missing as well as the evidence to

present the evolution of emerging topics. With the topic modeling algorithm, which applies the text mining technique to extract features in the documents, one could process a high quantity of materials. However, knowing only the frequency of certain terms' appearance is not enough; an in-depth interpretation of the results needs to be acquired from the domain's experts. Highly frequent words might also interfere and alter the true representation of research trends since those frequent words might not be semantically meaningful in terms of contributing to identifying research topics. To reduce the impact of multiple keywords while presenting the topics, researchers need to select the most representative keywords for analysis; certain research topics presented by the unchosen keywords will be overlooked (Kim et al., 2022). Furthermore, by backtracking articles using the keywords obtained through the topics and keywords analysis, the models will lead to a subdivision. Kim et al. (2022) pointed out that selecting highly-used keywords in articles as the basis of topic classification can avoid analysis errors caused by subdivisions.

The challenges previous methods encountered were the stress of labor work for analysis and the limited data that could be processed. Although researchers could handle a large quantity of articles and proceed to topics modeling by applying LDA techniques, the limitation of the method is the distorted results caused by treating the common words, such as method, algorithm, advantage, process, etc., as representations of the core topics, since the high-frequency terms are always identified as labels for the core topics. It still requires more workforce to review the results for the topic naming. This challenge becomes more difficult as the number of publications grows. This study took the author-defined keywords for better tokens for the topic modeling to avoid the possible mis-tagging of the articles. The algorithm used in this study was the Girvan-Newman (GN) algorithm, which could process a high quantity of documents and downsize the workforce needed for topic naming. GN algorithm is one of the benchmark methods for cluster detection. Nowadays, cluster detection in large networks has become a very important issue. Many algorithms introduced previously have been evaluated on a limited number of networks with a small number of nodes (Xiao et al., 2020). These cluster detection algorithms work well on small networks, but the performance of these algorithms on real-world networks with millions of nodes is severely reduced (Alghamdi & Greene, 2019). The GN algorithm greatly promoted the development of cluster detection methods, detecting clusters by gradually deleting edges with high edge betweenness. While high computational demands are required, the advantage of the GN algorithm is its greatly improved computing performance (Liu & Ma, 2019).

This study proposes an automatic solution based on the bibliometrics mechanism, which allows researchers to process massive amounts of data. To conduct research topic evolution analysis using the GN algorithm, this study adopted a new method to review a large quantity of literature. As proposed in this study, the labor-saving process in analyzing topic evolution is described as follows. First, the author-defined keywords were taken as tokens to present the research topics; then, co-keyword analysis was applied for further clustering and research topic tagging. Besides

the overall observation, this study also examined the data by three time zones to reveal the shift of research topics, including new development, slipping, merging, and disappearance.

2. Methodology

The authors applied the GN algorithm and took the author-defined keywords in LIS journal articles for topic detection based on their similarity. The GN analysis was run two times in this study; the first identified the topics, and the second grouped the topics into categories. The same steps were taken for all the articles, including those issued at different intervals, to observe the topic's evolution.

Step 1. Data collection and keyword cleaning

This study adopted the data collection methods as proposed in the work by Huang et al. (2019). The authors checked the journals under the category of "Information Science and Library Science" (ISLS) by Journal Citation Report, which includes journals related to Management Information Systems (MIS), Library Science (LS), Information Science (IS), and Informetrics (IM). Among the 86 journals under the category ISLS, 25 journals from MIS and three non-research-oriented journals were excluded, leaving 58 journals under the category of IS and LIS. The authors further searched articles issued from 2007 to 2021 on the Web of Science for topic analysis. Besides analyzing the data as a whole, the authors observed the changes across three periods to reveal topic evolution. The three periods are 2007 to 2011, 2012 to 2016, and 2017 to 2021.

In this study, the co-occurrence of author-defined keywords was applied to define similarity among journal articles, and a document relation network was constructed accordingly for topic detection. After downloading the bibliographic information of the journal articles, author-defined keywords were extracted to detect the research topics. If there were no author-defined keywords, the keywords provided by Keyword Plus would be used. If there were neither author-defined keywords nor keywords provided by Keyword Plus, the works would then be excluded from the study. Authority control and word stemming were employed. Based on the import objects from Python modules, necessary word segmentation and word stemming processes took place. The process is as follows: Step 1. Importing simple_preprocess from gensim.utils for word segmentation, $1 \leq$ length of word ≤ 35; Step 2. Importing WordNetLemmatizer and SnowballStemmer from nltk.stem for word stemming; Step 3. Importing MWETokenizer from nltk.tokenize for the multi-words. It was also found that the downloaded data carried HTML control codes, for example, "hypothè ses passerelles" for "hypothèses passerelles," as well as non-English characters – thèorie des jeux. These all required data cleaning before performing data analysis.

Step 2. Calculating document similarity based on author-defined keywords and keywords provided by Keyword Plus

The keywords, author-defined or provided by Keyword Plus, were taken as tokens for constructing co-occurrence relationships as the two-mode networks (document to keyword) transferring to one-mode networks (document to document weighted) to show the similarities among documents. As the number of keywords

increases, the linkage between documents is stronger. To make the co-keyword networks more meaningful, the authors included document pairs with a keyword-co-occurrence value equal to or greater than 3.

Step 3. Topics detected with the Girvan-Newman Algorithm

The authors took the one-mode networks of document-to-document weighted pairs from the first GN analysis as the input data for the second GN analysis for research topic detection (Girvan & Newman, 2002). The GN algorithm is a popular topology-based community detection approach, which partitions the network by gradually removing edges with high betweenness centralities to output the hierarchical cluster. The Girvan-Newman algorithm detects communities, the connected components of the remaining network, by progressively removing edges from the network (Newman, 2004). The steps of the Girvan-Newman algorithm could be described as follows (Despalatović et al., 2014):

(1) Calculate edge betweenness for every edge in the graph.
(2) Remove the edge with the highest edge betweenness.
(3) Calculate edge betweenness for remaining edges.
(4) Repeat step 2 to step 4 until all edges are removed.

However, when is the optimal decomposition moment reached for a network to turn into communities? Newman and Girvan (2004) proposed "modularity," a method of a qualitative measure of network decomposition. To divide a network into k clusters. A $k \times k$ symmetric matrix e is defined, in which the element e_{ij} is the fraction of all edges in the network that link vertices in cluster i to vertices in cluster j. The trace of this matrix $tr(e) = \Sigma_i e_{ii}$ gives the fraction of edges in the network that connect vertices in the same cluster. A good division into clusters should have a high value of this trace. The row (or column) sums is defined as $a_i = \Sigma_j e_{ij}$, representing the fraction of edges connecting to vertices in cluster i. In a network where the edges fall between vertices without regard for the clusters they belong to, $e_{ij} = a_i a_j$ would be established (Newman & Girvan, 2004). Thus, a modularity measure could be defined as:

$$Q = \sum_i (e_{ii} - a_i)^2 = tr(e) - \|e^2\|$$

where $\|\boldsymbol{x}\|$ is the sum of elements of the matrix $\mathbf{x}$. The e_{ij} is the fraction of edges number that connects vertices between community i to community j in the total edge number. Then on the diagonal of the matrix e is the fraction of edges located within the same community, so the trace of the matrix $tr(e)$ is the fraction of edges that will not be removed in the process of removing edges. When the fraction of edges within the communities is higher than in a random graph, the value is $Q = 0$. As Q is approaching value 1, the community structure in the network is better. In most cases, the value of Q is between 0.3 and 0.7. Value Q is calculated in every step of divisible algorithms; the maximum value Q gives us the best partition of the graph (Newman & Girvan, 2004). This study applied the GN algorithm to conduct clustering analysis on article pairs with a keyword co-occurrence value greater than 3.

Step 4. Topics merged up with cosine similarity

There were more than 100 topics identified after the first GN algorithm analysis. In this study, the authors applied cosine similarity analysis

to merge the resultant topics from the first GN algorithm analysis and to gain topical categories. The first GN algorithm clusters documents as topics based on their keyword co-occurrence; thus, each topic includes non-repeating documents. In the cosine similarity analysis, keywords of author-defined or provided by Keyword Plus of each document in a topic were extracted as the weighted word vectors in each topic. Then cosine similarity was used to merge topics based on their weighted word vectors. Those with a cosine similarity greater than 0.5 were considered from similar topics merged up. Only the core topics identified in the first GN algorithm analysis were included in the cosine similarity analysis. The cosine similarity, generally used to measure the similarity among documents, was used for word vectors, estimating each topic pair. It measures the cosine of an angle between two vectors projected in a multi-dimensional space (Elavarasi et al., 2014). In this study, cosine similarity is calculated as follows:

$$Cosine\ similarity = \frac{A * B}{\|A\|\|B\|} = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2}\sqrt{\sum_{i=1}^{n} B_i^2}}$$

where A_i and B_i are components of word vectors in topic A and topic B respectively. The second cycle of the topic network was made based on the cosine similarity values, and all the topics were merged according to the similarity.

Step 5. Evolution of research topics, views revealed with duo GN algorithm analysis

Changes in research topics over time, namely: newly developed, ceased, merged, split, grown, and declined, are essential issues in topic evolution analysis. The cosine similarity of the weighted word vectors among topics in consecutive intervals was calculated. Then the second GN algorithm analysis was applied to cluster the topics in two consecutive time-windows based on their cosine similarity of the weighted word vectors. Finally, those topics in two consecutive time-windows were clustered together, and their cosine similarity greater than 0.5 were considered from the same topic flows, representing the topic's evolution. This step aimed to calculate the similarity between the topic appearing in two time-windows clustered together. The topic flows were used as tokens to present in the Sankey diagram.

Step 6. Topics naming

The final step of data analysis was naming the detected topics, topical groups, and evolution trends. Researchers with LIS expertise were invited to review the clusters, keywords, and works grouped under various clusters and to label the topics and topical groups. Further analysis of context was also carried out to have an in-depth interpretation of the research topic evolution in LIS studies.

3. Results: LIS Research Topics

There were 43,352 journal articles issued in 58 journals from 2007 to 2021. To ensure the availability of the data sources for analysis, the Keyword Plus provided by Web of Science was used for articles without author-defined keywords. Five thousand three hundred seventy-three articles had neither author-defined keywords nor Keyword Plus, excluded from the further topic analysis. This study included 37,979 articles with 55,938 author-defined keywords and 3,967 Keyword Plus. After authority control and the removal of duplicates, finally, 51,975 keywords were extracted as the

basis for further analysis. Keywords provided by Keyword Plus used in this study accounts for 7.63% (3,967 divided by 51,975) of the overall authority-controlled keywords. Table 1 shows the results of the paper count based on co-keywords. For topic evolution analysis, the authors divided the data into three sets by publication years, which contain 9,103 articles for the data set from 2007 to 2011, 12,986 for the set from 2012 to 2016, and 15,890 for the set from 2017 to 2021.

For further topic analysis, this study included document pairs with a keyword-co-occurrence value equal to or greater than 3. Table 2 shows the counting results of the author-defined keywords co-occurrence for the LIS articles from 2007 to 2021, and Figure 1 is the visualized presentation of the topic clusters. Analyzing the data set, there were 10,489 articles included and 269 clusters identified; two major groups, which cover 80 clusters and have 7,807 articles, present the core research topics, while there are 189 isolated research topics. Each node in the figure represents a topic. To confirm the validity of the results, the authors further calculated the Max *Q* value, which was 0.77, and the value was in the confidence interval (Newman & Girvan, 2004).

For the topic evolution analysis, the data set was divided into three sub-sets, 2007 to 2011, 2012 to 2016, and 2017 to 2021. Table 3 presents the statistical results, and Figure 2 shows the clustering results' visualization. Same as in Figure 1, each node represents a topic. The Q values for the three periods were 0.79, 0.75, and 0.73, all in the confidence interval (Newman & Girvan, 2004). Table 3 lists the number of articles covered by the core topics for the three time-windows, which are 61.21%, 53.86%, and 64.71%. The

Table 1. Author-defined Keywords Co-occurrence in LIS

Co-keywords	2007-2011			2012-2016			2017-2021			2007-2021		
	Article pair	No. of articles	Cumulate articles (%)	Article pair	No. of articles	Cumulate articles (%)	Article pair	No. of articles	Cumulate articles (%)	Article pair	No. of articles	Cumulate articles (%)
≥8	4	8	0.09	5	10	0.08	26	38	0.24	38	44	0.12
7	4	8	0.18	13	21	0.24	36	41	0.50	49	53	0.26
6	15	25	0.45	45	51	0.63	115	91	1.07	223	196	0.77
5	66	90	1.44	199	170	1.94	395	265	2.74	940	643	2.46
4	421	368	5.48	947	656	6.99	1,579	727	7.31	5,170	2,229	8.33
3	3,055	1,383	20.67	5,394	1,996	22.36	9,137	2,550	23.36	39,131	7,324	27.62
2	38,545	3,779	62.19	58,989	5,390	63.87	97,834	6,511	64.34	488,388	16,246	70.39
1	877,351	3,115	96.41	1,331,500	4,392	97.69	1,915,822	5,335	97.91	11,421,824	10,725	98.63
0	-	327	100.00	-	300	100.00	-	332	100.00	-	519	100.00
Total	877,351	9,103		1,331,500	12,986		1,915,822	15,890		11,421,824	37,979	

Table 2. Author-defined Keywords Co-occurrence in LIS, 2007-2021

	2007-2021 (Q = 0.77)	
Total topics (%)	269	(100.00)
No. of articles (%)	10,489	(100.00)
Core topics (%)	80	(29.74)
No. of articles (%)	7,807	(74.43)
Isolated topics (%)	189	(70.26)
No. of articles (%)	2,682	(25.57)

Figure 1. Topic Clusters Detected by GN Algorithm, 2007-2021

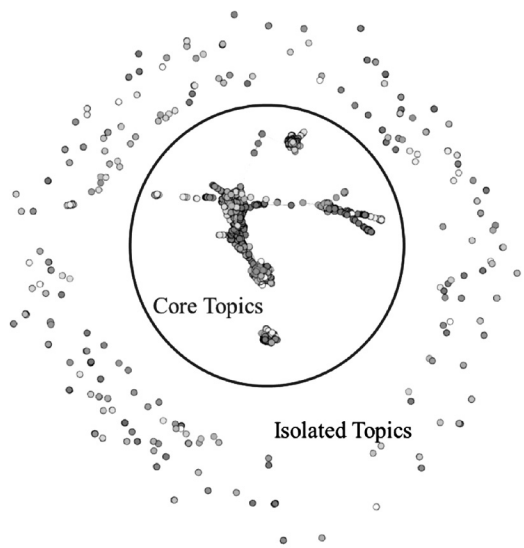


Table 3. Author-defined Keywords Co-occurrence in LIS, Three Time-windows

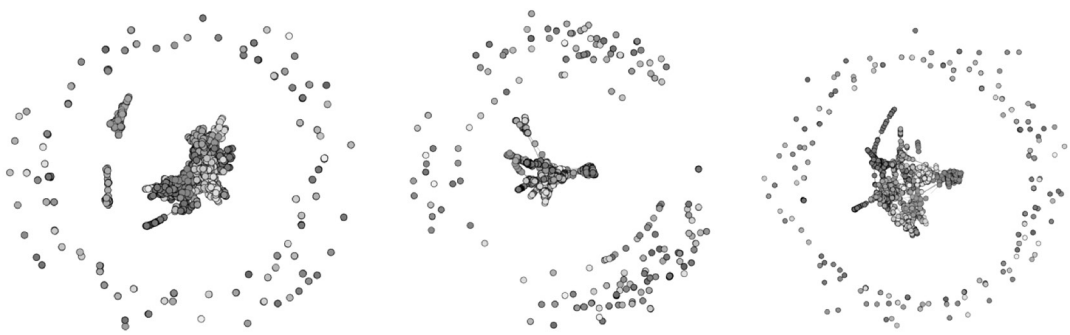
	2007-2011 (Q = 0.79)		2012-2016 (Q = 0.75)		2017-2021 (Q = 0.73)	
Topics (%)	115	(100.00)	177	(100.00)	161	(100.00)
No. of articles (%)	1,882	(100.00)	2,904	(100.00)	3,712	(100.00)
Core topics (%)	27	(23.48)	21	(11.86)	22	(13.66)
No. of articles (%)	1,152	(61.21)	1,564	(53.86)	2,402	(64.71)
Isolated topics (%)	88	(76.52)	156	(88.14)	139	(86.34)
No. of articles (%)	730	(38.79)	1,340	(46.14)	1,310	(35.29)

other side of the result is that nearly 60% of the LIS journal articles contributed to the core topics targeted in the study for further analysis and discussion. As for the isolated topics scattered in the LIS field presented in 40% of the LIS publications were excluded from the research topic evolution analysis.

3.1 Topical categories-topics merged up: 2007-2021

Table 4 lists the topical categories and the sub-categories. There were 4 topical categories: information seeking, information behavior, themes over social media, and bibliometrics, which were identified from 80 core topics. The first topic category, which covers 3,846 articles, included the discussion on the quality of information obtained, the factors influencing the seeking results, and the information transmission; the focus was more emphasized on the outcome of information seeking. The second topic group was constructed by 2,690 articles, and the center of the discussion was information behavior, from the various user types and information in specific domains to the research approach adopted. Among the subject domains, health information attracted more attention. As social media came into the information society, research shifted the focus onto the themes on social media, themes from different platforms, and special issues. This category was formed by 1,123 articles. The fourth noticeable topic category was

Figure 2. Topic Clusters Detected by the GN Algorithm of Three Time-windows



2007-2011 2012-2016 2017-2021

Table 4. LIS Core Topics in the Whole Period

Category	Sub-category	No. of topics	No. of articles
Information seeking	Information quality, influential factor, etc.	37	3,846
Information behavior	User types, qualitative, health information behavior, etc.	35	2,690
Social media	Social media, themes	7	1,123
Bibliometrics	Data sources, indicators, etc.	1	148

bibliometrics, and there were 148 articles listed in this cluster. The issues covered included the sources used for analysis, and the methods and indicators applied.

3.2 Topical categories: Observations from the three time-windows

The data was processed separately by the three intervals to observe the possible differences in research trends. Table 5 illustrates the results in more detail. In the first one (2007 to 2011), there were 5 topical categories based on 27 core topics; research evaluation and information behavior were the two major categories. Research evaluation includes 3 sub-issues of data sources, methods applied, and the use for rankings; as for information behavior, health information behavior drew high attention from researchers, and a shift of research methods was also brought up in various studies. In the second interval (2012 to 2016), there were 3 topical categories identified from 21 core topics, and they were all related to information behavior. One of them continued the discussion on health information behavior as a lasting trend from the first period; the other topics that were issued in this period, such as information retrieval, behavior on social media, and information literacy competencies, were also related to information behavior. An interesting shift was found in the third interval (2017 to 2021), which revealed 3 topical categories from 22 core topics. Besides the special issues related to information behaviors, more discussions on library service design were covered, and studies about research evaluation with scholarly communication have become popular again.

3.3 Topic evolution

To unclose the paradigm shift of research topics, the authors applied the Sankey technique, which provides an opportunity to observe the birth, development, and decline of certain research topics (Figure 3). From the overall research results, information seeking and information behavior studies took up major research efforts. However, some interesting observations were found when the data was put into different time-windows. The discussion in this section will mainly focus on two research topics, information behavior studies and research evaluation. The prior was the explicit topic, and the latter was the implicit research focus. Platforms that provided observing bases for and methods used in both topics were the two main research streams in the first time-window. Moving towards the second time-window, the shift of research focus was observed.

Observation 1: Internal shifts of research focus in information behavior studies

As various information systems appeared in the first time-window, it guided more attention to studies on information behavior presented in different platforms (A2.1). Information behavior remained a major issue for research. However, the focus of relevant studies split into information behavior as a process (B2.4), special interest in information retrieval (B2.3), and design of information systems (B2.5). As we got into the third time-window, information behavior as a process remained, and information retrieval and system design were turned into part of the special issues related to information behavior. One new focus was also spotted in this period: health librarianship growing out of health information behavior study, becoming a special research interest.

Table 5. LIS Core Topics in Three Time-windows

Category	Sub-category	No. of topics	No. of articles
2007-2011			
A1 Research evaluation	A1.1 Platform, data, and Indicators	9	109
	A1.2 Scientific collaboration	3	305
A2 Information behavior	A2.1 Information seeking and communication	9	526
	A2.2 Communicate technology	1	5
A3 Research performance ranking		2	17
A4 Health information behavior		2	126
A5 Information retrieval service (cyberspace)		1	64
2012-2016			
B1 Health Behavior (research methodology)		1	460
B2 Information behavior	B2.1 On-line information access behavior	2	9
	B2.2 Science impact evaluation	1	59
	B2.3 Information seeking model	1	84
	B2.4 On-line knowledge communication	4	355
	B2.5 Impact from information system	1	138
	B2.6 Bibliometrics in social media	7	357
B3 Information literacy	B3.1 Information literacy and librarianship	1	73
	B3.2 Information literacy in health information needs	1	29
2017-2021			
C1 Library services with special focus on information behavior	C1.1 Library service related	8	275
	C1.2 Information behavior in specific groups	1	585
	C1.3 Information behavior in special issues	6	1,078
	C1.4 Health librarianship	1	131
C2 Bibliometrics, application in scholarly communication		4	284
C3 Health information analysis (methodology)		2	49

Observation 2: Qualitative approach was appraised in information behavior studies

From the viewpoint of research strategies, groups of information behavior studies with qualitative approaches were observed in all three time periods (A4, B1, C1.2). However, more interest in health information behavior studies (B1) was observed during the second time-window, and the output doubled in productivity. A similar amount of research output remained in the third time-window (C1.2).

Observation 3: The rise and fall of research attention on research evaluation in LIS

The discussion on research evaluation, the first time-window, and the proportion of discussion on data sources (A1.1) and methods (A1.2) was about 1:3, and the discussion declined, and only the discussion on the concepts and results of research evaluation remained. Research evaluation was out of the scope of the core topics as we looked at the research outputs of the third time-window.

Observation 4: New focus, from information literacy to scholarly communication

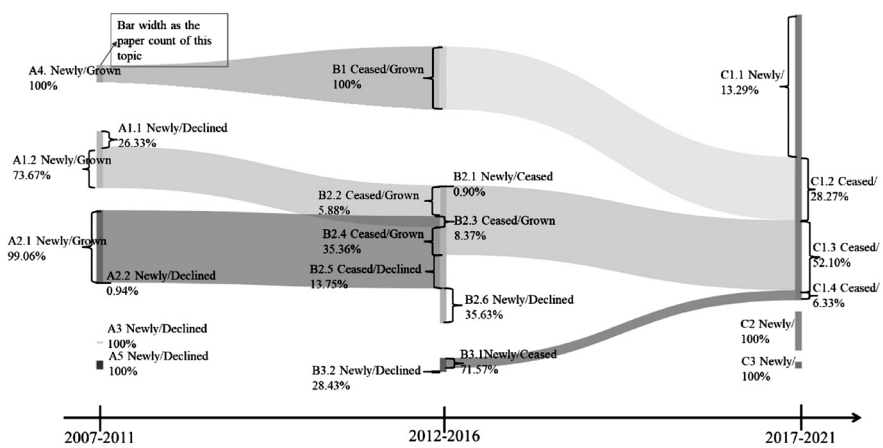
Reviewing the research topics by time-windows, new topics formed during later time-windows. For example, research on information literacy was identified in the second time-window and continued to be discussed as one of the issues related to library services. Another new focus is scholarly communication, from the research output and impact to ethical issues gained more attention in the third time-window.

Besides the topics mentioned above, library-based studies had a higher presentation. There were more than 200 studies linked to library services, especially the service seen in the setting of academic libraries.

4. Conclusion

In their work, Järvelin and Vakkari (2022) reviewed LIS research for 50 years, including 142

Figure 3. Sankey Flow of LIS Core-topic Categories Evolution



Note. Newly: newly developed topic, not traced from the previous time-window; Grown: topic extended to the next time-window; Ceased: continued topic from the previous time-window, but no following trace; Declined: short-term developed topic.

articles issued in 1965, 449 articles in 1985, 718 articles in 2005, and 1,210 articles in 2015, leading to a total of 2,519 articles. The five preset facets include (a) topics, (b) viewpoints, (c) social levels, (d) research strategies, and (e) the application of research strategies. To overcome the restriction of manual work, the authors took advantage of automated data process algorithms to include a high quantity of data and obtain the results within a more feasible time frame. This study included 8,498 journal articles issued over 15 years. (Please reference the numbers of articles in Table 3) With the massive data, the results show more details and are closer to the real development picture.

Instead of applying a text mining model, such as LDA, to extract keywords from the full text and form the topical categories, the authors took mainly the author-defined keywords, which were believed to be more representable to present the major concepts carried in the work, as the tokens to perform the topical clustering. This study took a post-clustering approach to reveal the topic evolution of LIS research, which provides a different perspective to examine research trends. The adopted algorithm allows the authors to manage a larger number of materials. To avoid the clustering bias, the GN algorithm analysis was modified, from two-mode clusters, document-keyword, to one-mode cluster, document-document, before topic modeling was applied in this study. With two GN algorithm analyses, the topics identified in the first GN algorithm analysis were merged into topical categories, which differed from categorizing the materials according to a pre-designed classification schema that required manually tagging the works and could only accommodate a limited number of works. The authors took one step further to practice the same analysis on the three data sets, in which the data was divided based on three time-windows. Doing so could present the mainstreams of LIS research topics in different periods and also provide a chance for the authors to reveal the changes in research topics over time: newly developed, ceased, merged, split, grown, and declined. Take the studies on health information behavior as examples; it was found that the research focuses in this area started with observing health information behavior, then moved on to the applications of various research methods and health information librarianship. The previous studies did not cover such an account of transformation and development in a research topic.

The results from the proposed mechanism produced a large number of topical clusters. With the duo GN analysis, the hierarchy structure was constructed. The benefit of the structure can not only be consolidated into a few categories but also subdivided into specific topics to discuss research trends in detail and depth. The advantage of the one-mode cluster is that it could clearly grasp the articles linked by each topic. The isolated topics, which have not been discussed in depth yet, could be further explored to reveal their links and significance among core and isolated topics and their impact on research trends (Burt, 2004). Moreover, more analytical facets, such as authors, institutions, and countries, could also be included for research strength analysis besides the paper level.

Acknowledgment

This work was financially supported by the Center for Research in Econometric Theory and Applications (Grant no. 111L900204) which

is under the Featured Areas Research Center Program by Higher Education Sprout Project of Ministry of Education (MOE) in Taiwan, the Universities and Colleges Humanities and Social Sciences Benchmarking Project (Grant no. 111L9A002), and the National Science and Technology Council (NSTC), Taiwan, with Grant No. 111-2634-F-002-018-.

References

Alghamdi, E., & Greene, D. (2019). Active semi-supervised overlapping community finding with pairwise constraints. *Applied Network Science, 4*, Article 63. https://doi.org/10.1007/s41109-019-0175-7

Burt, R. S. (2004). Structural holes and good ideas. *American Journal of Sociology, 110*(2), 349-399. https://doi.org/10.1086/421787

Chang, Y.-W., Huang, M.-H., & Lin, C.-W. (2015). Evolution of research subjects in library and information science based on keyword, bibliographical coupling, and co-citation analyses. *Scientometrics, 105*(3), 2071-2087. https://doi.org/10.1007/s11192-015-1762-8

Despalatović, L., Vojković, T., & Vukičević, D. (2014). Community structure in networks: Girvan-Newman algorithm improvement. In P. Biljanovic, Z. Butkovic, K. Skala, S. Golubic, M. Cicin-Sain, V. Sruk, S. Ribaric, S. Gros, B. Vrdoljak, M. Mauher, & G. Cetusic (Eds.), *2014 37th International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO)* (pp. 997-1002). IEEE. https://doi.org/10.1109/MIPRO.2014.6859714

Elavarasi, S. A., Akilandeswari, J., & Menaga, K. (2014). A survey on semantic similarity measure. *International Journal of Research in Advent Technology, 2*(3), 389-398.

Figuerola, C. G., García Marco, F. J., & Pinto, M. (2017). Mapping the evolution of library and information science (1978–2014) using topic modeling on LISA. *Scientometrics, 112*(3), 1507-1535. https://doi.org/10.1007/s11192-017-2432-9

Girvan, M., & Newman, M. E. J. (2002). Community structure in social and biological networks. *Proceedings of the National Academy of Sciences of the United States of America, 99*(12), 7821-7826. https://doi.org/10.1073/pnas.122653799

Han, X. (2020). Evolution of research topics in LIS between 1996 and 2019: An analysis based on latent Dirichlet allocation topic model. *Scientometrics, 125*(3), 2561-2595. https://doi.org/10.1007/s11192-020-03721-0

Huang, M.-H., Shaw, W.-C., & Lin, C.-S. (2019). One category, two communities: Subfield differences in "Information Science and Library Science" in Journal Citation Reports. *Scientometrics, 119*(2), 1059-1079. https://doi.org/10.1007/s11192-019-03074-3

Järvelin, K., & Vakkari, P. (1990). Content analysis of research articles in library and information science. *Library & Information Science Research, 12*(4), 395-421.

Järvelin, K., & Vakkari, P. (1993). The evolution of library and information science 1965–1985: A content analysis of journal articles. *Information Processing &*

Management, 29(1), 129-144. https://doi.org/10.1016/0306-4573(93)90028-C

Järvelin, K., & Vakkari, P. (2022). LIS research across 50 years: Content analysis of journal articles. *Journal of Documentation, 78*(7), 65-88. https://doi.org/10.1108/JD-03-2021-0062

Kim, E. H. J., Jeong, Y. K., Kim, Y., & Song, M. (2022). Exploring scientific trajectories of a large-scale dataset using topic-integrated path extraction. *Journal of Informetrics, 16*(1), Article 101242. https://doi.org/10.1016/j.joi.2021.101242

Liu, Z., & Ma, Y. (2019). A divide and agglomerate algorithm for community detection in social networks. *Information Sciences, 482*, 321-333. https://doi.org/10.1016/j.ins.2019.01.028

Ma, J., & Lund, B. (2021). The evolution and shift of research topics and methods in library and information science. *Journal of the Association for Information Science & Technology, 72*(8), 1059-1074. https://doi.org/10.1002/asi.24474

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). *Efficient estimation of word representations in vector space*. arXiv. https://doi.org/10.48550/arXiv.1301.3781

Miyata, Y., Ishita, E., Yang, F., Yamamoto, M., Iwase, A., & Kurata, K. (2020). Knowledge structure transition in library and information science: Topic modeling and visualization. *Scientometrics, 125*(1), 665-687. https://doi.org/10.1007/s11192-020-03657-5

Newman, M. E. J. (2004). Detecting community structure in networks. *The European Physical Journal B, 38*(2), 321-330. https://doi.org/10.1140/epjb/e2004-00124-y

Newman, M. E. J., & Girvan, M. (2004). Finding and evaluating community structure in networks. *Physical Review E, 69*(2), Article 026113. https://doi.org/10.1103/PhysRevE.69.026113

Tuomaala, O., Järvelin, K., & Vakkari, P. (2014). Evolution of library and information science, 1965–2005: Content analysis of journal articles. *Journal of the Association for Information Science & Technology, 65*(7), 1446-1462. https://doi.org/10.1002/asi.23034

Wang, X., Wang, H., & Huang, H. (2021). Evolutionary exploration and comparative analysis of the research topic networks in information disciplines. *Scientometrics, 126*(6), 4991-5017. https://doi.org/10.1007/s11192-021-03963-6

Xiao, J., Ren, H.-F., & Xu, X.-K. (2020). Constructing real-life benchmarks for community detection by rewiring edges. *Complexity, 2020*, Article 7096230. https://doi.org/10.1155/2020/7096230

(Received: 2022/11/10; Accepted: 2022/12/14)

圖書資訊學刊　第21卷第1期　（2023.6）　頁1-16
https://doi.org/10.6182/jlis.202306_21(1).001

重新審視Girvan-Newman演算法的研究主題分析：以圖書資訊學研究為例

Revisit Girvan-Newman Algorithm for Research Topic Analysis: An Application on Library and Information Science Studies

羅思嘉[1,3]　王俊傑[2,3]

Szu-Chia Lo[1,3], Chun-Chieh Wang[2,3]

摘　要

研究趨勢的分析為學術界提供一個可以了解過去並藉以支持未來持續發展的重要機會。主題演化分析能用來定位當前研究、連結研究主題間的關係，以及辨識研究主題間的落差。在本研究中，作者重新審視現有的Girvan-Newman（GN）演算法在主題演化分析的應用，提出了一個新的主題演化分析的方法。在作者－關鍵詞關係、單模叢集分析和雙重GN演算法的基礎上，作者進行圖書資訊學的研究主題分析。研究結果顯示，作者提出的方法可以處理大量資料文獻，並且能夠避免因為小樣本或雙模叢集分析導致的偏誤結果，進而確保研究結果的有效性。最後作者更提出建構研究主題的階層來衡量研究主題之間的連結關係，可作為後續深入研究的方法。

關鍵字：研究主題演化、Girvan-Newman演算法、圖書資訊學

[1] 國立臺灣大學圖書資訊學系
Department of Library and Information Science, National Taiwan University, Taipei, Taiwan
[2] 國立臺灣大學生物產業傳播暨發展學系
Department of Bio-Industry Communication and Development, National Taiwan University, Taipei, Taiwan
[3] 國立臺灣大學計量理論與應用研究中心
Center for Research in Econometric Theory and Applications, National Taiwan University, Taipei, Taiwan
* 通訊作者Corresponding Author: 王俊傑Chun-Chieh Wang, E-mail: wangcc@ntu.edu.tw

註：本中文摘要由作者提供。
以APA格式引用本文：Lo, S.-C., & Wang, C.-C. (2023). Revisit Girvan-Newman algorithm for research topic analysis: An application on library and information science studies. *Journal of Library and Information Studies, 21*(1), 1-16. https://doi.org/10.6182/jlis.202306_21(1).001
以Chicago格式引用本文：Szu-Chia Lo and Chun-Chieh Wang, "Revisit Girvan-Newman algorithm for research topic analysis: An application on library and information science studies," *Journal of Library and Information Studies* 21, no. 1 (2023): 1-16. https://doi.org/10.6182/jlis.202306_21(1).001