Identifying Emerging Relationships in Healthcare Domain Journals via Citation Network Analysis

Kuo-Chung Chu,
Hsin-Ke Lu,
and Wen-I Liu

ABSTRACT

Online e-journal databases enable scholars to search the literature in a research domain or to cross-search an interdisciplinary field. The key literature can thereby be efficiently mapped. This study builds a web-based citation analysis system consisting of four modules: (1) literature search; (2) statistics; (3) articles analysis; and (4) co-citation analysis. The system focuses on the PubMed Central dataset and facilitates specific keyword searches in each research domain for authors, journals, and core issues. In addition, we use data mining techniques for co-citation analysis. The results could help researchers develop an in-depth understanding of the research domain. An automated system for co-citation analysis promises to facilitate understanding of the changing trends that affect the journal structure of research domains. The proposed system has the potential to become a value-added database of the healthcare domain, which will benefit researchers.

INTRODUCTION

Healthcare is a multidisciplinary research domain of medical services provided both inside and outside a hospital or clinical setting. Article retrieval for systematic reviews in the domain is much more elusive than retrieval for reviews in clinical medicine because of the interdisciplinary nature of the field and the lack of a significant body of evaluative literature. Other connecting research fields consist of the respective research fields of the application domain (i.e., the health sciences, including medicine and nursing).[1] In addition, valuable knowledge and methods can be taken from the fields of psychology, the social sciences, economics, ethics, and law. Further, the integration of those disciplines is attracting increasing interest.[2]

Researchers may use bibliometrics to evaluate the influence of a paper or describe the relationship between citing and cited papers. Citation analysis, one of several possible bibliometric approaches, is more popular than others because of the advent of information technologies.[3] Citation analysis counts the frequency of cited papers from a set of citing papers to determine the most influential scholars, publications, or universities in a discipline. It can be classified into two basic types: the first type counts only the citations in a paper that are authored by an individual, while the second

Kuo-Chung Chu (kcchu@ntunhs.edu.tw) is Professor, Department of Information Management, and Dean, College of Health Technology, National Taipei University of Nursing and Health Sciences; **Hsin-Ke Lu** (sklu@sce.pccu.edu.tw) is Associate Professor, Department of Information Management, and Dean, School of Continuing Education, Chinese Culture University; **Wen-I Liu** (wenyi@ntunhs.edu.tw, Corresponding author) is Professor, Department of Nursing, and Dean, College of Nursing, National Taipei University of Nursing and Health Sciences.

type analyzes co-citations to identify intellectual links among authors in different articles. This paper focuses on the second type of citation analysis.

Small defined co-citation analysis as "the frequency with which two items of earlier literature are cited together by the later literature."[4] It is not only the most important type of bibliometric analysis, but also the most sophisticated and popular method. Many other methods originate from citation analysis, including document co-citation analysis, bibliographic coupling,[5] author co-citation analysis,[6] and co-word analysis.[7]

There are levels of co-citation analysis: document, author, and journal. Co-citation could be used to establish a cluster or "core" of earlier literature.[8] The pattern of links between documents can establish a structure to highlight the relationship of research areas. Citation patterns change when previously less-cited papers are cited more frequently, or old papers are no longer cited. Changing citation patterns imply the possibility of new developments in research areas; furthermore, we can investigate changing patterns to understand the scientific trend within a research domain.[9]

Co-citation analysis can help obtain a global overview of research domains.[10] The aim of this paper is to detect emerging issues in the healthcare research domain via citation network analysis. Our results can provide a basis for knowledge that researchers can use to construct a search strategy. Structural knowledge is intrinsic to problem solving.

Because of the interdisciplinary nature of the healthcare domain and the broadness of the term, research is performed in several research fields, such as nursing, nursing informatics, long-term care, medical informatics, geriatrics, information technology, telecommunications, and so forth. Although electronic journals enable searching by author, article, and journal title using keywords or full text, the results are limited to article content and references and therefore do not provide an in-depth understanding of the knowledge structure in a specific domain. The knowledge structure includes the core journals, core issues, the analysis of research trends, and the changes in focus of researchers.

For a novice researcher, however, the literature survey remains a troublesome process in terms of precisely identifying the key articles that highlight the overview concept in a specific domain. The process is complicated and time-consuming, and it limits the number of articles collected for retrospective research. The objective of this paper is to provide information about the challenges and methodology of relevant literature retrieval by systematically reviewing the effectiveness of healthcare strategies. To this end, we build a platform for automatically gathering the full text of e-journals offered by the PubMed Central (PMC) database.[11] We then analyze the co-citation results to understand the research theme of the domain.

METHODS

This paper tries to build a value-added literature database system for co-citation analysis of healthcare research. The results of the analysis will be visually presented to provide the structure of the domain knowledge to increase the productivity of researchers.

https://doi.org/10.6017/ital.v37i1.9595

Dataset

For co-citation analysis, a data source of related articles on healthcare is required. For this paper, the articles were retrieved from the PMC database using search terms related to the healthcare domain. To build the article analysis system, we used bibliometrics to locate the relevant references while analysis techniques were implemented by the association rule algorithm of data mining. The PMC database, which is produced by the US National Institutes of Health and is implemented and maintained by the US National Center for Biotechnology Information of the US National Library of Medicine, provides electronic articles from more than one thousand full-text journals for free. We could understand the publication status from the Open Access Subset (OAS) and access to the OAI (Open Archives Initiative) Protocol for Metadata Harvesting, which includes the full text in XML and PDF. Regarding access permission, PMC offers a dataset of many open access journal articles. This paper used a dedicated XML-formatted dataset (https://www.ncbi.nlm.nih.gov/pmc/tools/oai/). The XML-formatted dataset followed the specification of DTD (document type definition) files, which are sorted by journal title. Each article has a PMCID (PMC identification), which is useful for data analysis. In addition to the dataset, the PMC also provides several web services to help widely disseminate articles to researchers.

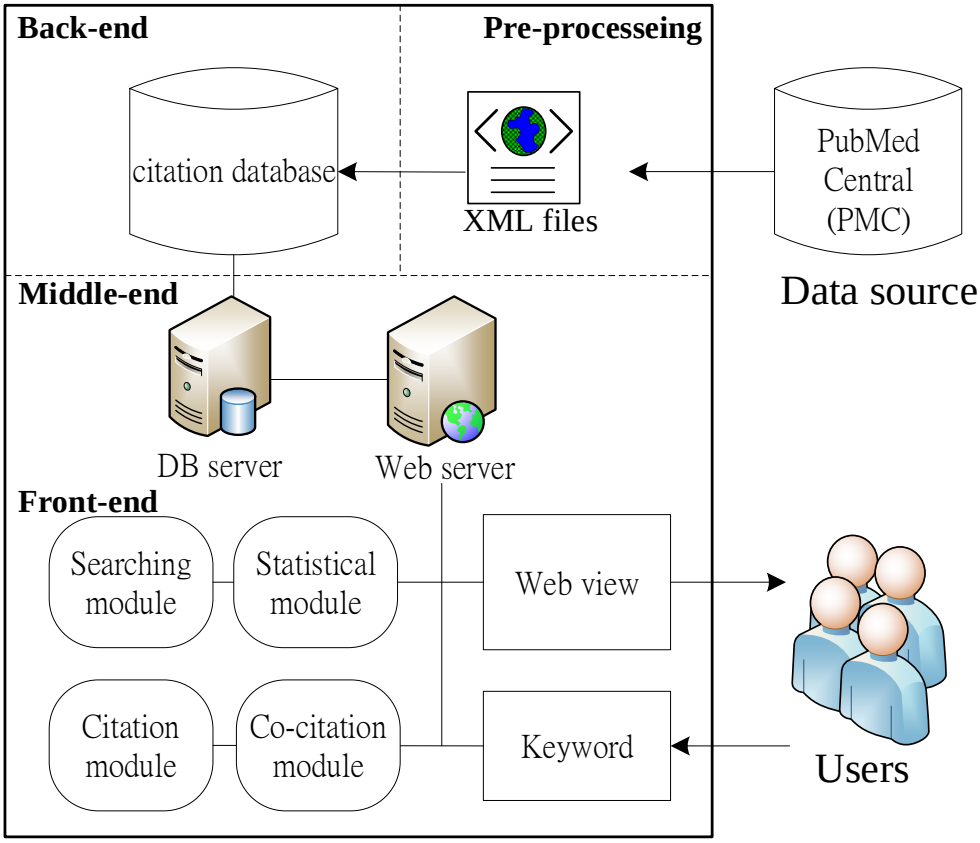

Figure 1. The system architecture of citation analysis with four subsystems.

System Architecture

Our development environment consisted of the following four subsystems: front-end, middle-end, back-end, and pre-processing. The front-end creates a "web view," a visualization of the results for our web-based co-citation analysis system. The system architecture is shown in figure 1.

Front-End Development Subsystem
We used Adobe Dreamweaver CS5 as a visual development tool for the design of web templates. The PHP programming language was chosen to build the co-citation system that would be used to access and analyze the full-text articles. In terms of the data mining technique, we implemented the Apriori algorithm with the PHP language.[12] The results were exported as XML to a charting process, where we used amCharts (https://www.amcharts.com/), to create stock charts, column charts, pie charts, scatter charts, line charts, and so forth.

Middle-End Server Subsystem
The system architecture was a Microsoft Windows-based environment with a XAMPP 2.5 web server platform (https://www.apachefriends.org/download.html). XAMPP is a cross-platform web development kit that consists of Apache, MySQL, PHP, and Perl. It works across several operating systems, such as Linux, Windows, Apache, macOS, and Oracle Solaris, and provides SSL encryption, a phpMyAdmin database management system, Webalizer traffic management and control suite, a mail server (Mercury Mail Transport System), and FileZilla FTP server.

Back-End Database Subsystem
To speed up co-citation analysis, the back-end database system used MySQL 5.0.51b with interface phpMyAdmin 2.11.7 for easy management of the database. MySQL includes the following features:

- Using C and C++ to code programs, users can develop an application programming interface (API) through Visual Basic, C, C + +, Eiffel, Java, Perl, PHP, Python, Ruby, and Tcl languages with the multithreading capability that can be used in multi-CPU systems and easily linked to other databases.
- Performance of querying articles is quick because SQL commands are optimally implemented, providing many additional commands and functions for a user-friendly and flexible operating database. An encryption mechanism is also offered to improve data confidentiality.
- MySQL can handle a large-scale dataset. The storage capacity is up to 2TB for Win32 NTS systems and up to 4TB for Linux ext3 systems.
- It provides the software MyODBC as an ODBC driver for connecting many programming languages, and it several languages and character sets to achieve localization and internationalization.

Pre-processing Subsystem
The PMC provides access to the article via OAS, OAI services, e-utilities, and FTP. We used FTP to download a compressed (ZIP) file packaged with a filename following the pattern "articles?-?.xml.tar.gz" on October 28, 2012 (ftp://ftp.ncbi.nlm.nih.gov/pub/pmc), where "?-?" is "0-9" or "A-Z". The size of the ZIP file was approximately 6.17GB. After extraction, the size of the articles was approximately 10GB. The 571,890 articles from 3,046 journals were grouped and

https://doi.org/10.6017/ital.v37i1.9595

sorted by journal title in a folder labeled with an abbreviated title. An XML file would, for example, be named "AAPSJ-10-1-2751445.nxml," where "AAPSJ" was the abbreviated title of the journal *American Association of Pharmaceutical Scientists Journal*, "10" was the volume of the journal, "1" was number of the issue, and "2751445" was the PMCID.

We used related technologies for developing systems that include PHP language, array usage, and the Apriori algorithm to analyze the articles and build the co-citation system.[13] Finally, several analysis modules were created to build an integrated co-citation system.

RESEARCH PROCEDURE

The following is our seven-step research procedure to fulfill the integrated co-citation system:

1. Parse XML file: select tags for construction of database; choose fields for co-citation analysis (for example, <*journal-title*>, <*author-list*>, and <*reference-list*>).
2. Present web-based article: design webpage and CSS style; present web-based XML file by indexing variable <*PMCID*>.
3. Build an abstract database: the database consists of several fields: <*author-list*>, <*article-title*>, <*journal-title*>, <*volume*>, <*issue*>, <*pub-date*>, and <*abstract*>.
4. Develop searching module: pass the keyword to the method "POST" in SQL query language and present the search result in the webpage.
5. Develop statistical module: the statistical results include number of article and cited articles, the journals and authors cited in all articles, and the number of cited articles.
6. Develop citation module: visually present the statistical results in several formats; rank searched journals; rank searched and cited journals in all the articles.
7. Develop co-citation module: analyze the association between articles with the Apriori algorithm.

Association Rule Algorithms

The association rule (AR), usually represented by $A \rightarrow B$, means that the transaction containing item *A* also contains item *B.* There are many such rules in most of the dataset, but some were useless. To validate the rules, two indicators, support and confidence, can be applied. Support, which means usefulness, is the number of times the rules feature in the transactions, whereas confidence means certainty, which is the probability that *B* occurs whenever the *A* occurs. We chose the rules for which the values of both support and confidence were greater than a predefined threshold. For example, a rule stipulating "toast→jam" has support of 1.2 percent and confidence of 65 percent, implying that 1.2 percent of the transactions contain "toast" and "jam" and that 65 percent of the transactions containing "toast" also contained "jam."

The principle for generating the AR is based on two features of the documents: (1) find the high-frequency items that set their supports greater than the threshold; (2) for each dataset *X* and its subnet *Y*, check the rule $X \rightarrow Y$ if the support is greater than the threshold, in which the rule $X \rightarrow Y$ means that the occurrence in the rule containing *X* also contains *Y.* Most studies focus on searching high-frequency item sets.[14] The most popular approach for identifying the item sets is Apriori algorithm, as shown in figure 2.[15] The algorithm rationale is that if the support of item set *I* is less

than or equal to the threshold, *I* is not a high-frequency item set. New item set *I* that inserts any item *A* into *I* would not be a high-frequency item set. According to the rationale, the Apriori algorithm is an iteration-based approach. First, it generates candidate item set *C1* by calculating the number of occurrences of each attribute and finding that the high-frequency item set *L1* has support greater than the threshold. Second, it generates item set *C2* by joining *L1* to *C1*, iteratively finding *L2* and generating *C3,* and so on.

```
1: L1 = {large 1-item sets};
2: for (k=2; Lk-1; k++) do begin
3:    Ck = Candidate_gen (Lk-1);
4:    for all transactions t∈D do begin /* generate candidate k-dataset*/
5:          Ct = subset (Ck, t);
6:          for all candidates c ∈ Ct do
7:                      c_count=c_count+1;
8:          end
9:          Lk ={c∈Ck | c_count ≥ minsuppport}
10:   end
11: return L = ∪Lk;
```

Figure 2. The Apriori algorithm.

The Apriori algorithm is one of the most commonly used methods for AR induction. The Candidate_gen algorithm, as shown in figure 3, includes join and prune operations for generating candidate sets.[16] Steps 1 to 4 generate all possible candidate item sets *c* from *Lk-1*. Steps 5 to 8: delete the item set that is not a frequent item set by the Apriori algorithm. Step 9 returns candidate set *Ck* to the main algorithm.

```
1: for each item set X1∈ Lk-1
2: for each item set X2∈ Lk-1
3:    c = join (X1[1], X1[2], X1[k-2], X1[k-1], X2[k-1])
4: Where X1[1] = X2[1], X1[k-2] = X2[k-2], X1[k-1] < X2[k-1];
5: for item sets c ∈ Ck do
6:    for all (k-1)-subsets s of c do
7:          if (s ∈ Lk-1) then add c to Ck;
8:          else delete c from Ck;
9: return Ck;
```

Figure 3. The Candidate_gen algorithm.

https://doi.org/10.6017/ital.v37i1.9595

RESULTS

We searched the PMC database with keywords "healthcare," "telecare," "ecare," "ehealthcare," and "telemedicine" and located 681 articles with a combined 14,368 references. Values were missing from the year field for 4 of the references; this was also the case for 635 of a total of 52,902 authors.

According to the keyword search for the healthcare domain, a pie chart of the journal citation analysis, as shown in figure 4, the top-ranked journal in terms of citations was the *British Medical Journal* (*BMJ*). It was cited approximately 439 times, 18.89 percent of the total, followed by the *Journal of the American Medical Association* (*JAMA*), which was cited approximately 344 times, 14.80 percent of the total. The trend of healthcare citation 1852 to 2009 peaked in 2006 at approximately 1,419 citations, with more than half of the total occurring in this year.

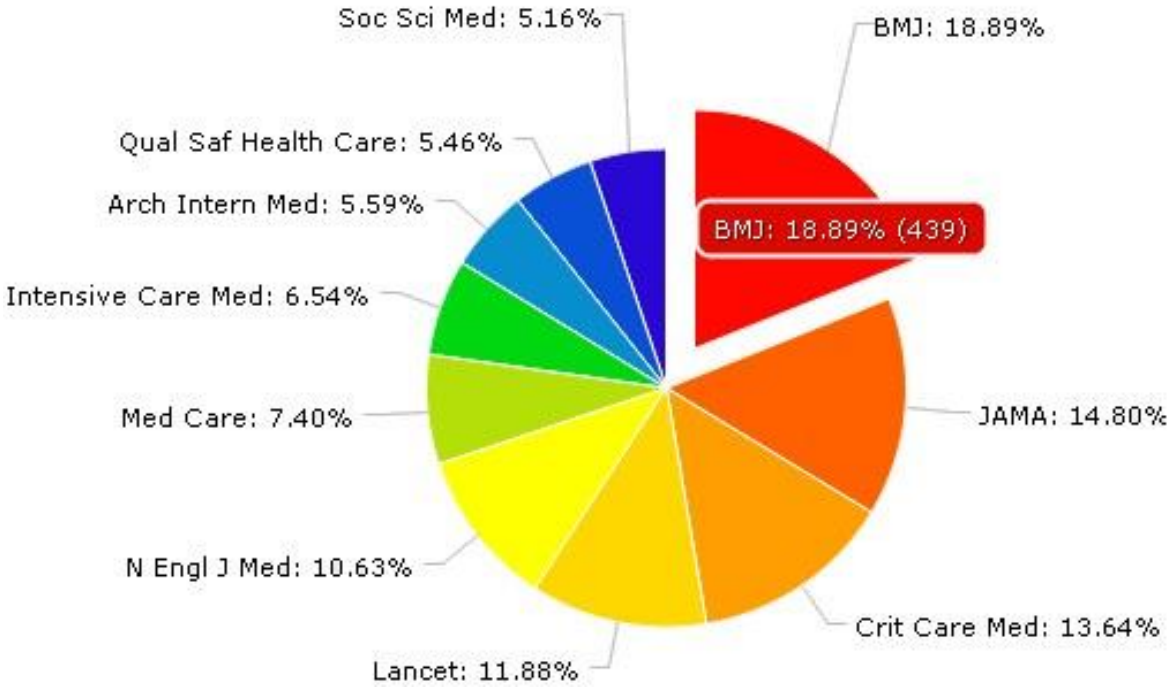

Figure 4. Top-cited journals in the healthcare domain by percentage of total citations (*N* = 2324)

With the keyword search for the healthcare domain, Figure 5 shows a pie chart of the author citations. The most-cited author was J. W. Varni, professor of pediatric cardiology at the University of Michigan Mott Children's Hospital in Ann Arbor. This author was cited approximately 149 times, equivalent to 23.24 percent of the total, followed by D. N. Herndon, professor at the Department of Plastic and Hand Surgery, Friedrich-Alexander University of Erlangen in Germany. This author was cited approximately 73 times, 11.39 percent of the total. By identifying the affiliations of the top-ranked authors, researchers can access related information in their field of interest.

The co-citation analysis was conducted using the Apriori algorithm. The relationship of co-citation journals with a supporting degree greater than 38 from 1852 to 2009 is shown in figure 6. Each

journal was denoted by a node, where the node with double circle meant the journal is co-cited with the other in a citing article. *BMJ*, which covers the fields of evidence-based nursing care, obstetrics, healthcare, nursing knowledge and practices, and others, is the core journal of the healthcare domain.

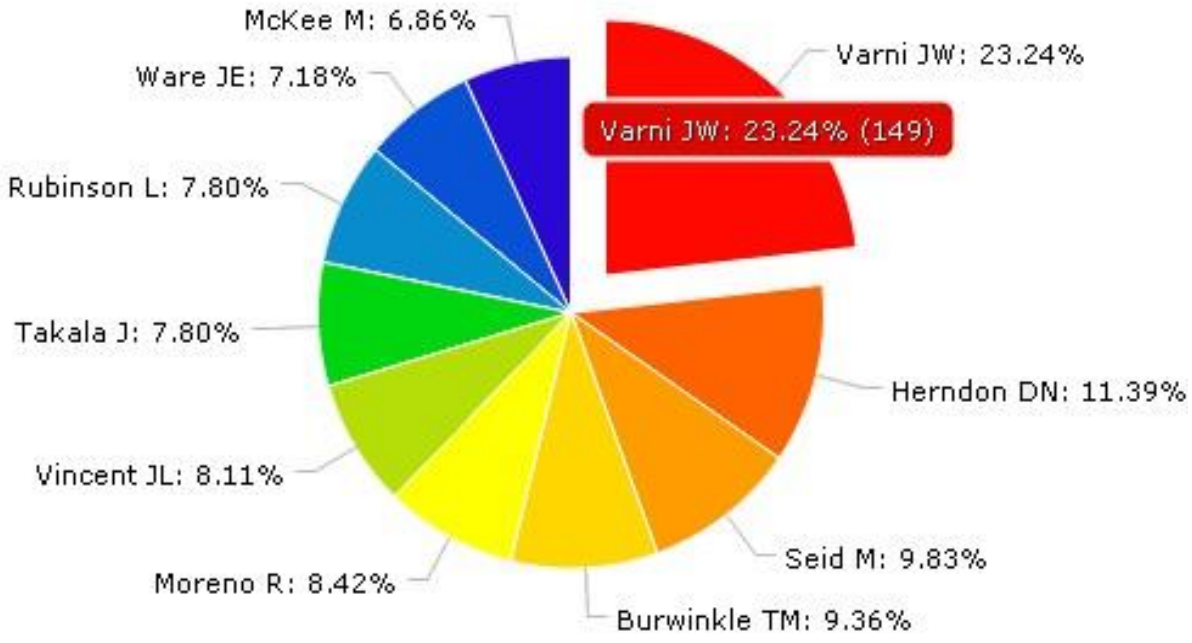

Figure 5. Top-cited authors in journals of the healthcare domain by percentage of total citations (*N* = 641)

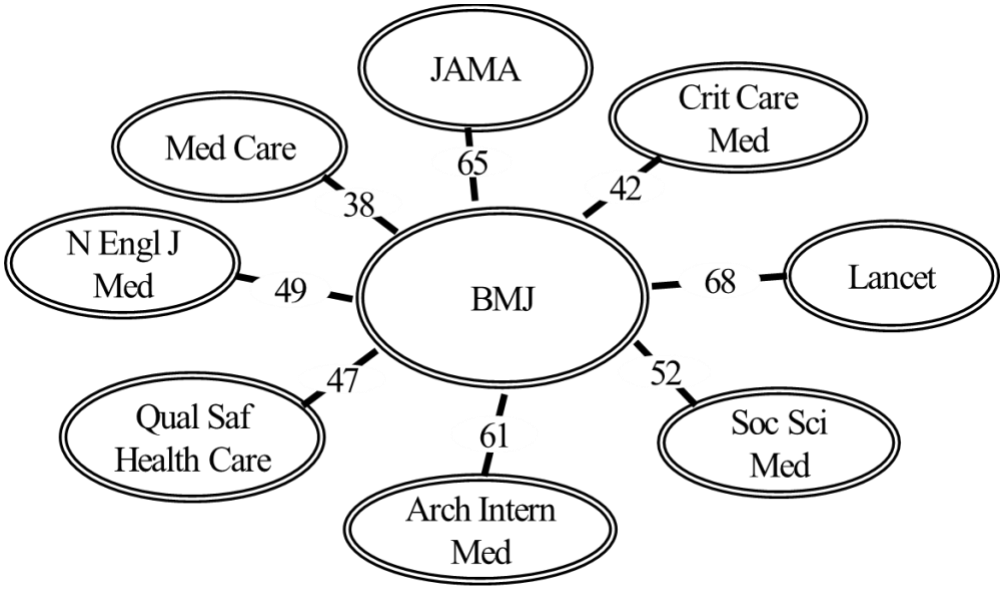

Figure 6. The relationship of co-citation journals with *BMJ*.

https://doi.org/10.6017/ital.v37i1.9595

To identify the focus of the journal, we analyze the co-citation in three periods. In 1852–1907, journals are not in co-citation relationships; in 1908–61, five candidates had a supporting degree greater than 1 (see table 1); and in 1962–2009, twenty-eight candidates had a supporting degree greater than 14 (see table 2 (for example, *BMJ* and *Lancet* had sixty-eight co-citations).

Table 1. Candidates in co-citation analysis with a supporting degree greater than 1 (1908–61).

No	Journals	No. of Journals Co-cited	Support
1	*Publ Math Inst Hung Acad Sci*, *Publ Math*	2	3
2	*JAOA*, *J Osteopath*	2	1
3	*Antioch Rev*, *J Abnorm Soc Psychol*	2	1
4	*N Engl J Med*, *Am Surg*	2	1
5	*Arch Neurol Psychiatry*, *J Neurol Psychopathol*, *Z Ges Neurol Psychiat*	3	1

Table 2. Candidates in co-citation analysis with a supporting degree greater than 14 (1962–2009).

No	Journals	No. of Journals Co-cited	Support
1	*BMJ*, *Lancet*	2	68
2	*BMJ*, *JAMA*	2	65
3	*JAMA*, *Med Care*	2	64
4	*BMJ*, *Arch Intern Med*	2	61
5	*Lancet*, *JAMA*	2	52
6	*Soc Sci Med*, *BMJ*	2	52
7	*JAMA*, *Arch Intern Med*	2	51
8	*Lancet*, *Med Care*	2	50
9	*Crit Care Med*, *Prehospital Disaster Med*	2	49
10	*N Engl J Med*, *BMJ*	2	49
11	*N Engl J Med*, *Lancet*	2	49
12	*N Engl J Med*, *JAMA*	2	47
13	*N Engl J Med*, *Med Care*	2	47
14	*Qual Saf Health Care*, *BMJ*	2	47
15	*BMJ*, *Crit Care Med*	2	42
16	*Med Care*, *BMJ*	2	38
17	*N Engl J Med*, *J Bone Miner Res*	2	33

18	*N Engl J Med*, *J Pediatr Surg*	2	26
19	*Lancet*, *J Pediatr Surg*	2	25
20	*JAMA*, *Nature*	2	25
21	*Lancet*, *JAMA*, *BMJ*	3	24
22	*N Engl J Med*, *Lancet*, *BMJ*	3	21
23	*Intensive Care Med*, *BMJ*	2	21
24	*BMJ*, *N Engl J Med*, *JAMA*	3	20
25	*N Engl J Med*, *JAMA*, *Lancet*	3	20
26	*JAMA*, *Med Care*, *Lancet*	3	14
27	*JAMA*, *Med Care*, *N Engl J Med*	3	14
28	*BMJ*, *JAMA*, *Lancet*, *N Engl J Med*	4	14

The link of co-citation journals in three periods from 1852 to 2009 can be summarized as follows: (1) three journals were highly cited but were not in a co-citation relationship in 1852–1907 (see figure 7); (2) five clusters of the healthcare journals in co-citation relationships were found for the years 1908–61 (see figure 8); and (3) 1962–2009 had a distinct cluster of four journals within the healthcare domain (see figure 9).

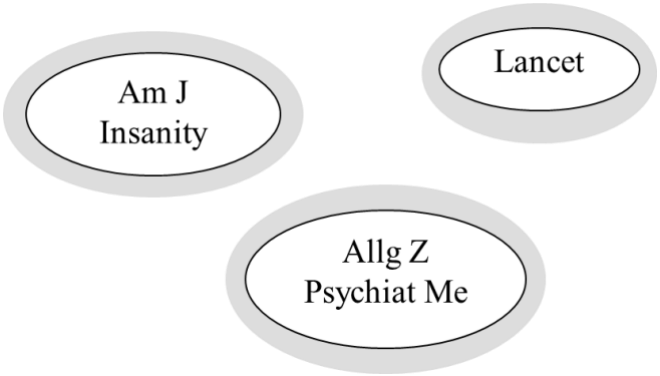

Figure 7. The relationship of co-citation journals for the healthcare domain in 1852–1907.

https://doi.org/10.6017/ital.v37i1.9595

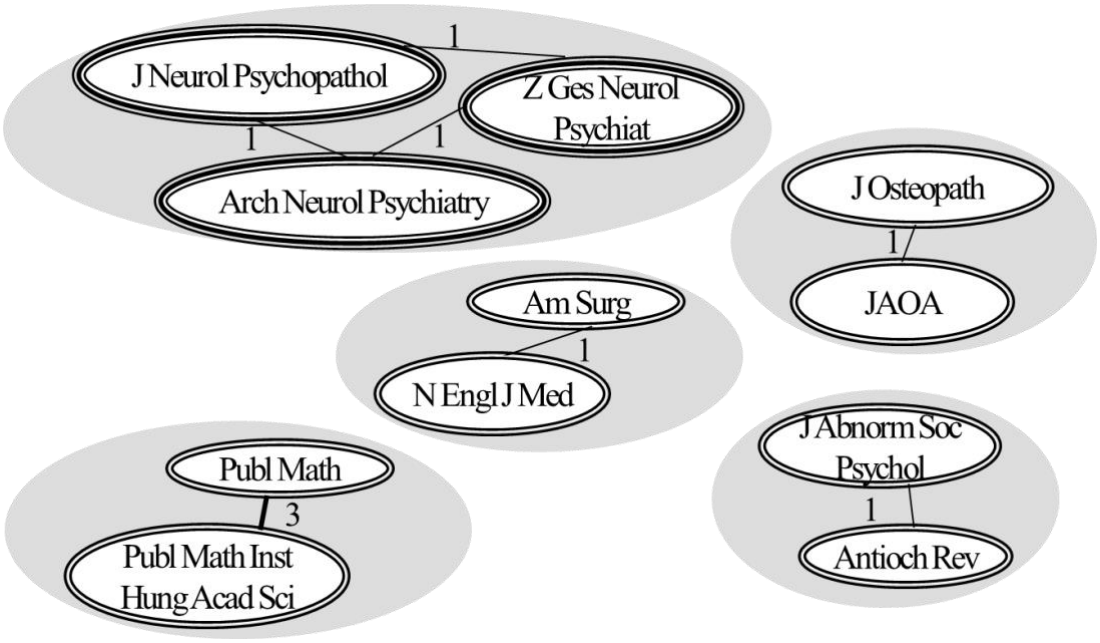

Figure 8. The relationship of co-citation journals for the healthcare domain in 1908–61. Journals with double circles are co-cited with the other in a citing article. Journals with triple circles are co-cited with the other two in a citing article.

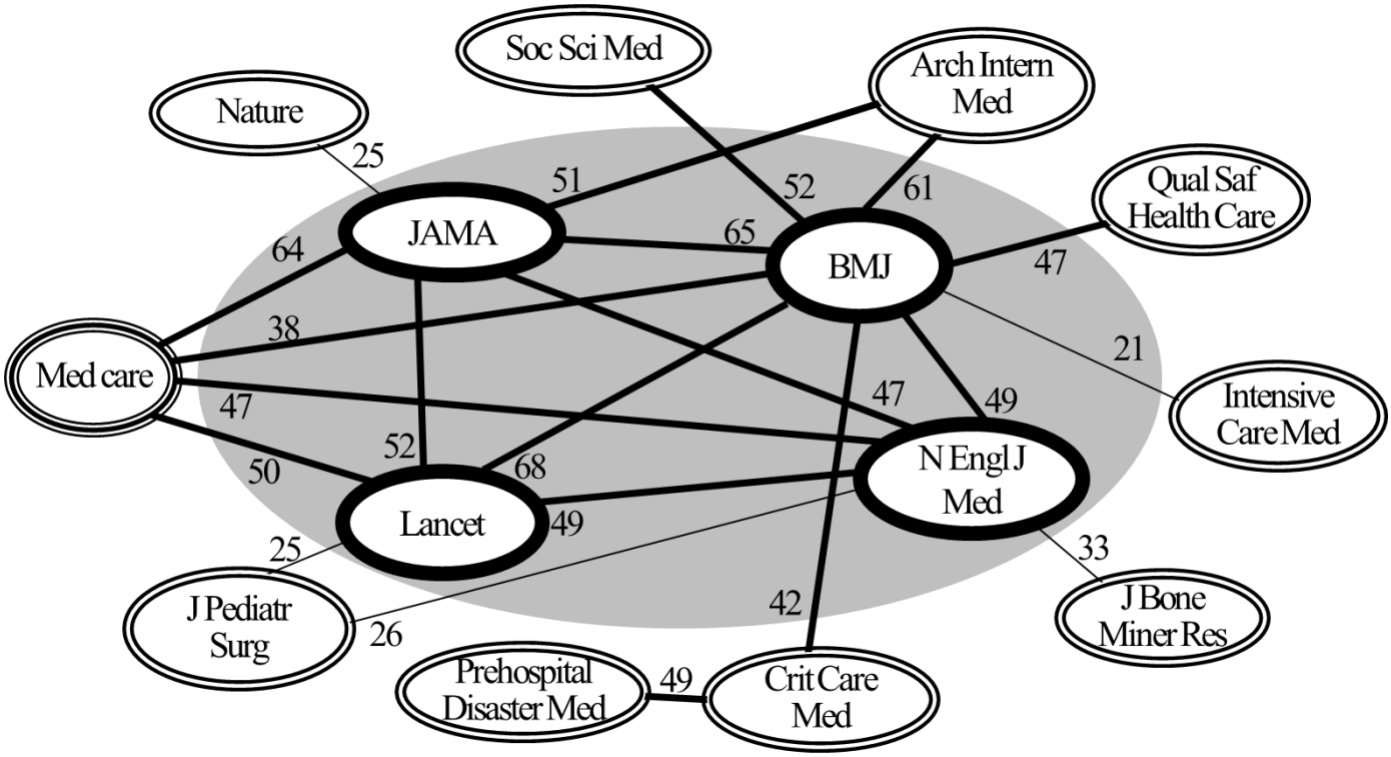

Figure 9. The relationship of co-citation journals for the healthcare domain in 1962–2009. The thick line and circle indicates the journals are co-cited in a citing article.

CONCLUSIONS

This paper presented an automated literature system for co-citation analysis to facilitate understanding of the sequence structure of journal articles cited in the healthcare domain. The system visually presents the results of its analysis to help researchers quickly identify the key articles that provide an overview of the healthcare domain. This paper used the keywords related to healthcare for its analysis and found that *BMJ* is a core journal in the domain. The co-citation analysis found a single cluster within the healthcare domain comprising four journals: *BMJ*, *JAMA*, *Lancet*, and the *New England Journal of Medicine*.

This paper focused on a co-citation analysis of journals. Authors, articles, and issues featured in the co-citation analysis can be further studied in an automated way. A period analysis of publication years is also important. Further analyses can facilitate understanding of the changes in a research domain and the trend of research issues. In addition, the automatic generation of a map would be a worthwhile topic for the future study.

ACKNOWLEDGEMENTS

This article was funded by the Ministry of Science and Technology of Taiwan (MOST), formerly known as National Science Council (NSC), with Grant No: NSC 100-2410-H-227-003. For the remaining authors none were declared. All the authors have made significant contributions to the article and agree with its content. There is no known conflict of interest in this study.

REFERENCES

[1] A. Kitson et al., "What are the Core Elements of Patient-Centered Care? A Narrative Review and Synthesis of the Literature from Health Policy, Medicine and Nursing," *Journal of Advanced Nursing* 69 (2013): 4–8, https://doi.org/10.1111/j.1365-2648.2012.06064.x.

[2] S. J. Brownsell et al., "Future Systems for Remote Health Care," *Journal of Telemedicine and Telecare* 5 (1999): 145–48, https://doi.org/10.1258/1357633991933503; B. G. Celler, N. H. Lovell, and D. K. Chan, "The Potential Impact of Home Telecare on Clinical Practice," *Medical Journal of Australia* 171 (1999): 518–20; R. Walker et al., "What It Will Take to Create New Internet Initiatives in Health Care," *Journal of Medical Systems* 27 (2003): 95–98, https://doi.org/10.1023/A:1021065330652.

[3] I. Marshakova-Shaikevich, The Standard Impact Factor as an Evaluation Tool of Science Fields and Scientific Journals," *Scientometrics* 35 (1996): 283–85, https://doi.org/10.1007/BF02018487; I. Marshakova-Shaikevich, "Bibliometric Maps of Field of Science," *Information Processing & Management* 41(2005):1536–45, https://doi.org/10.1016/j.ipm.2005.03.027; A. R. Ramos-Rodríguez and J. Ruíz-Navarro, "Changes in the Intellectual Structure of Strategic Management Research: A Bibliometric Study of the Strategic Management Journal, 1980–2000," *Strategic Management Journal* 25, no. 10 (2004): 982–1000, https://doi.org/10.1002/smj.397.

[4] H. Small, "Co-citation in the Scientific Literature: A New Measure of the Relationship between Two Documents," *Journal of American Society for Information Science* 24 (1973): 266–68.

[5] M. M. Kessler, "Bibliographic Coupling between Scientific Papers," *American Documentation* 14 (1963): 10–25, https://doi.org/10.1002/asi.5090140103; B. H. Weinberg, "Bibliographic Coupling: A Review," *Information Storage and Retrieval* 10 (1974): 190–95.

[6] H. D. White and B. C. Griffith, "Author Cocitation: A Literature Measure of Intellectual Structure," *Journal of the American Society for Information Science* 32 (1981): 164–70, https://doi.org/10.1002/asi.4630320302.

[7] Y. Ding, G. G. Chowdhury, and S. Foo, "Bibliometric Cartography of Information Retrieval Research by Using Co-word Analysis," *Information Processing & Management* 37 no. 6 (November 2001): 818–20, https://doi.org/10.1016/S0306-4573(00)00051-0.

[8] Small, "Co-citation," 266.

[9] D. Sullivan et al., "Understanding Rapid Theoretical Change in Particle Physics: A Month-by-Month Co-citation Analysis," *Scientometrics* 2 (1980): 312–16, https://doi.org/10.1007/BF02016351.

[10] N. Shibata et al., "Detecting Emerging Research Fronts based on Topological Measures in Citation Networks of Scientific Publications," *Technovation* 28 (2008): 762–70, https://doi.org/10.1016/j.technovation.2008.03.009.

[11] Weinberg, "Bibliographic Coupling."

[12] White and Griffith, "Author Cocitation."

[13] R. Agrawal and R. Srikant. "Fast Algorithm for Mining Association Rules in Large Databases" (paper, International Conference on Very Large Databases [VLDB], September 12–15, 1994, Santiago de Chile).

[14] R. Agrawal, T. Imielinski, and A. Swami, "Mining Association Rules between Sets of Items in Large Databases" (paper, ACM SIGMOD International Conference on Management of Data, Washington, DC, May 25–28, 1993.

[15] Agrawal and Srikant, "Fast Algorithm," 3.

[16] Ibid., 4.