ISTL Issues in Science and Technology Librarianship

Tips from the Experts

Prior Steps into Knowledge Mapping: Text Mining Application and Comparison

Faizhal Arif Santosa
Librarian for Polytechnic Institute of Nuclear Technology
National Research and Innovation Agency
D. I. Yogyakarta, ID
faizhal.arif.santosa@brin.go.id

Abstract

Bibliometrics is increasingly being used by the knowledge community and librarians to easily analyze patterns in knowledge. In the field, the use of data from databases that provide bibliometric information is not always completely clean, so pre-processing is required. Several previous studies have shown that bibliometric analysis begins with a simple pre-processing step. The goal of this research is to use text mining to perform pre-processing to find the basic terms of the keywords that appear – to essentially construct a controlled vocabulary for a bibliographic dataset. The method used in this study is cleaning keywords with the stemming method using RapidMiner software. Bibliometrix was used to compare the results. A total of 85 keywords were combined into basic words. Using the built process, this study discovers differences in the network built between raw data and data that has been pre-processed, resulting in differences in the analysis that will be produced. The built process can also be reused in a variety of real-world situations.

Keywords: Bibliometric mapping, Science mapping, Stemming, Text pre-processing

Recommended citation:

Santosa, F. A. (2022). Prior steps into knowledge mapping: Text mining application and comparison. *Issues in Science and Technology Librarianship, 102.* https://doi.org/10.29173/istl2736

Introduction

Today's publications are rapidly expanding, and volumes are larger than ever before, making bibliometrics an excellent tool for mapping knowledge over time. Librarians are frequently involved in this matter to assist researchers and general users in identifying trends and gaps in a subject, and this response has resulted in bibliometrics becoming a service offered by many libraries today. Bibliometrics has become a suitable area for modern librarians due to the profession's extensive understanding of bibliographic sources (Gumpenberger et al., 2012). However, bibliometrics can presents libraries with a challenge. This mapping is very complicated and time-consuming because it necessitates a wide range of software and many steps (Aria & Cuccurullo, 2017).

Various software packages are now available to make it easier for librarians to provide bibliometric services, each with its own set of features and capabilities. Bibliometrix, CiteSpace, CitNetExplorer, SciMAT, Sci2 Tool, Pajek, and VOSviewer are some examples of bibliometric analysis software. Moral-Muñoz et al. (2020) add that Bibliometrix provides a diverse set of techniques, whereas VOSviewer provides good visualization to facilitate a more detailed examination. In general, tools for performing bibliometric analysis can see the relationship between authors and the relationship between keywords in one document with another, and several others offer analysis using abstracts from various databases.

For librarians who have conducted bibliometric analysis, the existence of Web of Science (WoS) and Scopus can aid in the retrieval of bibliographic data, and some software has even provided a feature to pull data directly and process it via an API (application programming interface). This data contains a variety of information that was used in the analysis, such as the names of all authors, keywords used, and abstracts if they were available. However, the resulting data is typically not clean enough for immediate analysis, requiring a different approach during the data pre-processing stage.

According to CRISP-DM, data quality improvement can be handled during the data preparation phase by cleaning the data prior to the modeling phase (Schröer et al., 2021). Data cleaning is a generic task in and of itself, but when faced with actions that require specific situations, specialized task levels must be carried out (Chapman et al., 2000), and data cleaning is a prerequisite for performing bibliometric analysis (Moral-Muñoz et al., 2020). Checking the data before analyzing it becomes necessary in order to produce precise and useful information, particularly in bibliometric analysis, which is based on bibliographic data entered into the application.

Various researchers who employ bibliometric analysis have also investigated the significance of this pre-processing stage. Wang et al. (2020) used VOSviewer and CiteSpace to perform pre-processing directly on the WoS Core Collection database and generate 831 publications, but the steps are unclear. To eliminate duplicate data, several supporting software can be used, such as CiteSpace (Li et al., 2020; Wang et al., 2021) and Excel (Obidat, 2022). Han et al. (2020) used Python to perform pre-processing by removing common terms, compiling a list of synonyms, combining singular and plural words, and academic compound nouns. Meanwhile, CheshmehSohrabi & Mashhadi

(2022) use text mining by using Excel and RapidMiner software to perform pre-processing in the form of deeper checking of the dataset obtained from Scopus and finding 134 book reviews listed as article reviews, as well as filling in the blank cells with the number zero.

RapidMiner has been used in the scope of the libraries due to the ease with which various algorithms can be applied without the need for programming skills. Lamba and Madhusudhan (2018) performed a sentiment analysis where the data from Twitter required for his study was obtained using RapidMiner and analyzed using AYLIEN. Meanwhile, Moore (2017) employed RapidMiner to analyze sentiment in open-ended qualitative LibQUAL+ comments.

Manual pre-processing of large amounts of data will be costly and time-consuming for librarians. The author attempts to perform pre-processing using text mining using the RapidMiner application without the need for programming skills, with the main focus that the author proposes, namely keywords. The issue is that keywords with plural and singular values must be combined. This process's main contribution is that it can be reused by librarians to support bibliometric services in libraries.

Methods

As a test sample, data from Scopus was used as an application example, and 849 records about academic libraries from 2016 to 2021 were imported into RapidMiner by selecting the UTF-8 encoding file. The information gathered was in the form of "Citation information" and "Abstracts & keywords". Meanwhile, the authors used Bibliometrix to compare results.

Stemming Snowball (Porter, 2001) is used to find the root of a keyword. For example, "librarian" and "librarians" would be changed to "librarian", while "library" and "libraries" would become "librari". Both author and index keywords are processed, with the final result being a CSV (Comma-separated values) file. The trial on a co-occurrence network was performed in Bibliometrix without changing any of the available parameters by selecting author keywords and the clustering algorithm, Walktrap.

Findings

The initial file was counted as having 1987 keywords. "Academic libraries" and "academic librarians" are the top keywords (Figure 1). According to Bibliometrix, the use of original data from Scopus results in author keywords that are divided into many clusters, with a total of 12 clusters (Figure 2). "Academic libraries" and "academic librarians" are the most numerous nodes in Cluster 1. Meanwhile, "academic librarian" is in cluster 3, and "academic library" is in cluster 8. These are actually matching words and are separated from each other because they are treated as distinct entities, resulting in a multitude of nets and clusters.

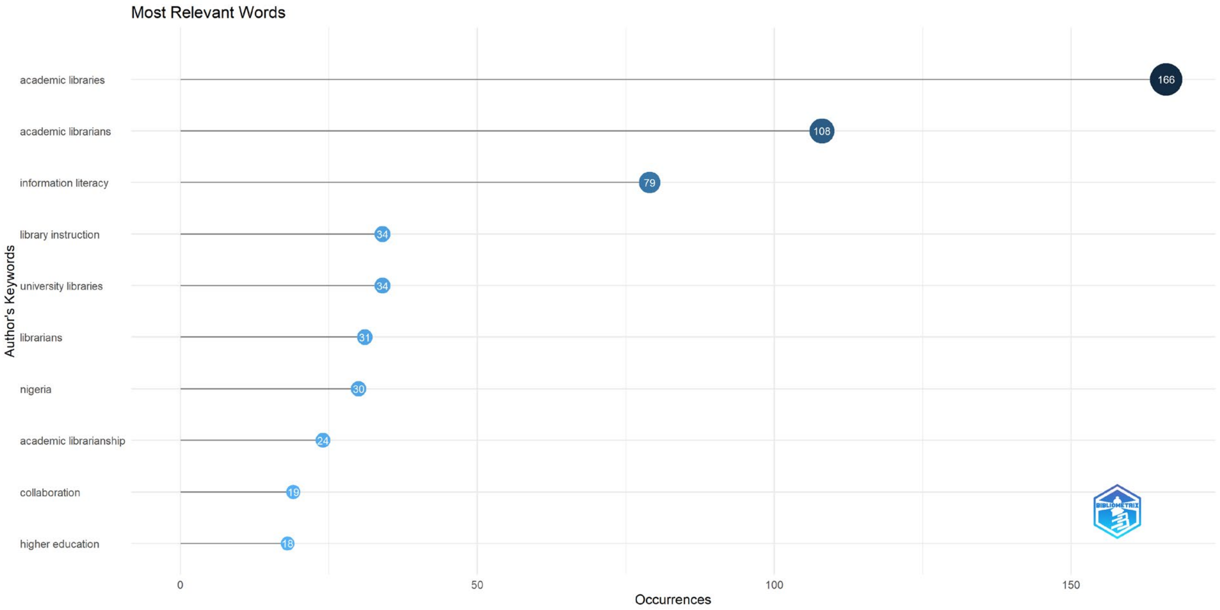

Figure 1. Top ten author keywords before processing

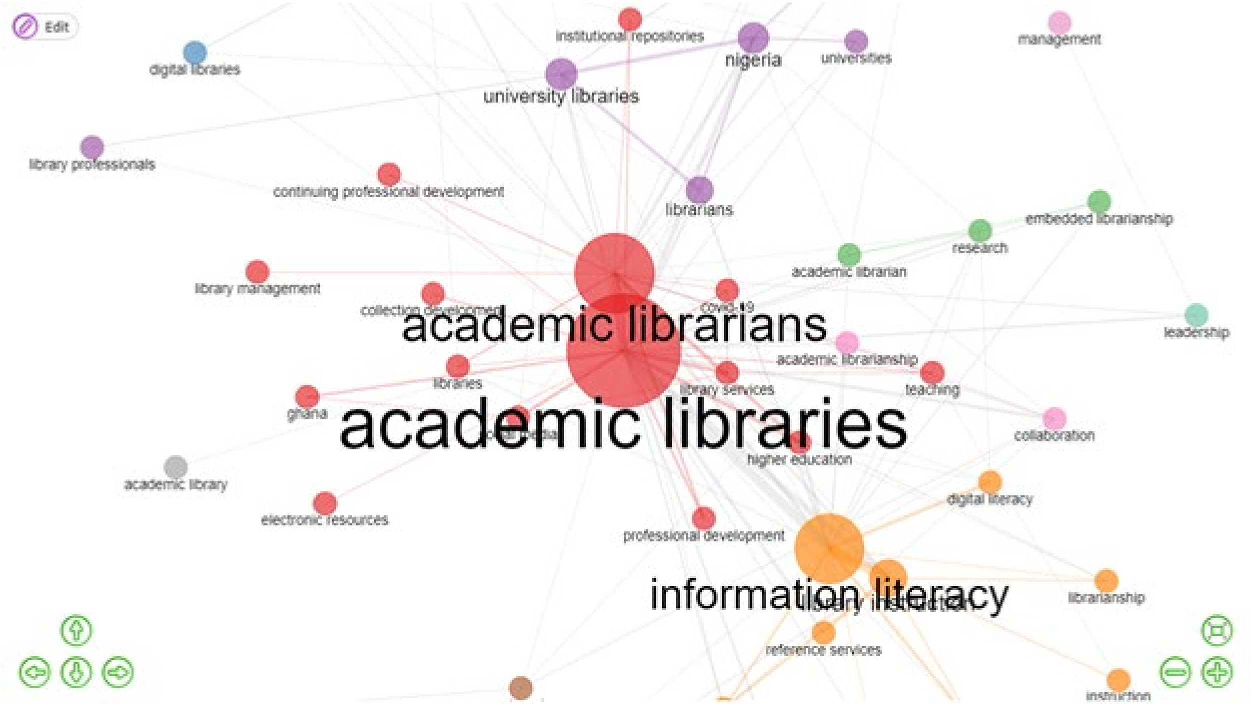

Figure 2. Bibliometrix's keyword network using Scopus data

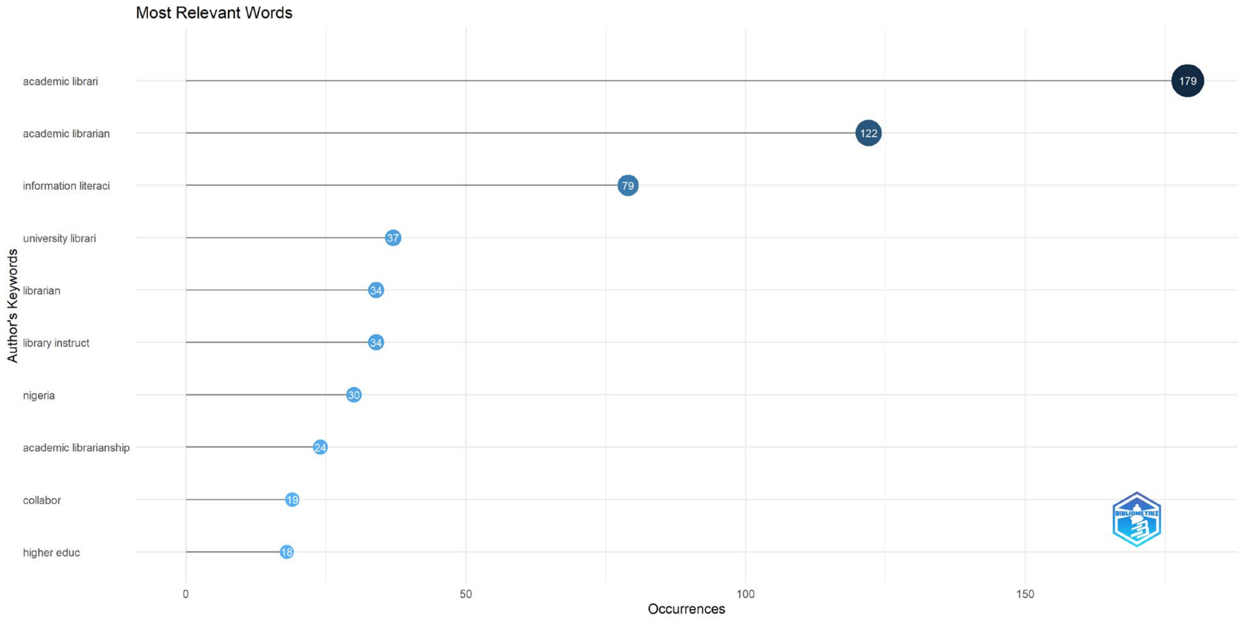

Figure 3. Top ten author keywords after pre-processing

Using pre-processed data, the number of keywords is reduced to 1902, or as many as 85 keywords have been combined into certain words. The increasing number of keyword occurrences in several keywords demonstrates this transformation (Figure 3). Different clustering results are displayed in pre-processed files (Figure 4). The word appears to change from "academic libraries" and "academic library" to a new word, "academic librari". Meanwhile, "academic librarians" was merged into the previously existing word "academic librarian". The formed cluster also changed significantly by having only 5 clusters, whereas clusters 4 and 5 only had 1 keyword. The large cluster is split into two parts, with "academic librari" and "academic librarian" as the largest nodes in each.

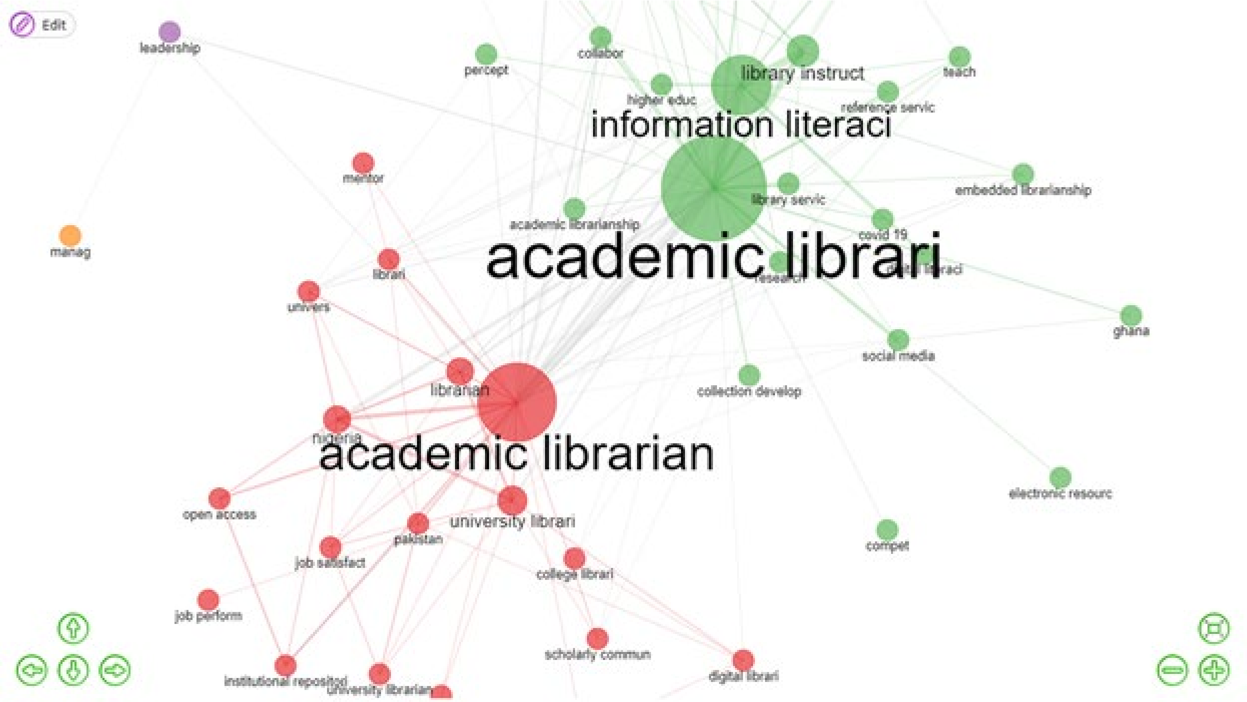

Figure 4. Keyword network on Bibliometrix after pre-processing

Conclusions

Pre-processing prior to bibliometric analysis can help provide better analytical results. “Garbage in, garbage out,” a popular computer phrase, indicates the need for anyone performing bibliometric analysis, including librarians, to perform pre-processing such as data cleaning and uniformity. Pre-processing keywords results in significant changes, which have implications for changing the network analysis. Furthermore, because of the unification of words with the same root word, this process can make analysis easier.

Anyone can re-create this process using existing data and RapidMiner. On the other hand, adequate hardware is required to install and run this software. The use of a web interface for pre-processing will be very useful and convenient for librarians and the general public.

Additional Material

RapidMiner for Academics

Data Availability: https://hdl.handle.net/20.500.12690/RIN/XMCUUX

References

Aria, M., & Cuccurullo, C. (2017). bibliometrix: An R-tool for comprehensive science mapping analysis. *Journal of Informetrics, 11*(4), 959–975. https://doi.org/10.1016/j.joi.2017.08.007

Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C., & Wirth, R. (2000). *CRISP-DM 1.0: Step-by-step data mining guide.* SPSS. https://www.kde.cs.uni-kassel.de/wp-content/uploads/lehre/ws2012-13/kdd/files/CRISPWP-0800.pdf

CheshmehSohrabi, M., & Mashhadi, A. (2022). Using data mining, text mining, and bibliometric techniques to the research trends and gaps in the field of language and linguistics. *Journal of Psycholinguistic Research.* https://doi.org/10.1007/s10936-022-09911-6

Gumpenberger, C., Wieland, M., & Gorraiz, J. (2012). Bibliometric practices and activities at the University of Vienna. *Library Management, 33*(3), 174–183. https://doi.org/10.1108/01435121211217199

Han, J., Kang, H.-J., Kim, M., & Kwon, G. H. (2020). Mapping the intellectual structure of research on surgery with mixed reality: Bibliometric network analysis (2000–2019). *Journal of Biomedical Informatics, 109*, 103516. https://doi.org/10.1016/j.jbi.2020.103516

Lamba, M., & Madhusudhan, M. (2018). Application of sentiment analysis in libraries to provide temporal information service: A case study on various facets of productivity. *Social Network Analysis and Mining, 8*(1), 63. https://doi.org/10.1007/s13278-018-0541-y

Li, D., Dai, F.-M., Xu, J.-J., & Jiang, M.-D. (2020). Characterizing hotspots and frontier landscapes of diabetes-specific distress from 2000 to 2018: A bibliometric study. *BioMed Research International, 2020*, 1–13. https://doi.org/10.1155/2020/8691451

Moore, M. T. (2017). Constructing a sentiment analysis model for LibQUAL+ comments. *Performance Measurement and Metrics, 18*(1), 78–87. https://doi.org/10.1108/PMM-07-2016-0031

Moral-Muñoz, J. A., Herrera-Viedma, E., Santisteban-Espejo, A., & Cobo, M. J. (2020). Software tools for conducting bibliometric analysis in science: An up-to-date review. *El Profesional de La Información, 29*(1). https://doi.org/10.3145/epi.2020.ene.03

Obidat, A. H. (2022). Bibliometric analysis of global scientific literature on the accessibility of an integrated e-learning model for students with disabilities. *Contemporary Educational Technology, 14*(3), ep374. https://doi.org/10.30935/cedtech/12064

Porter, M. F. (2001). *Snowball: A language for stemming algorithms.* http://snowball.tartarus.org/texts/introduction.html

Schröer, C., Kruse, F., & Gómez, J. M. (2021). A systematic literature review on applying CRISP-DM process model. *Procedia Computer Science, 181*, 526–534. https://doi.org/10.1016/j.procs.2021.01.199

Wang, X., Xu, Z., & Škare, M. (2020). A bibliometric analysis of Economic Research-Ekonomska Istraživanja (2007–2019). *Economic Research-Ekonomska Istraživanja, 33*(1), 865–886. https://doi.org/10.1080/1331677X.2020.1737558

Wang, X., Xu, Z., Su, S.-F., & Zhou, W. (2021). A comprehensive bibliometric analysis of uncertain group decision making from 1980 to 2019. *Information Sciences, 547*, 328–353. https://doi.org/10.1016/j.ins.2020.08.036

Recommended Reading

Aria, M., & Cuccurullo, C. (2017). Bibliometrix : An R-tool for comprehensive science mapping analysis. *Journal of Informetrics, 11*(4), 959-975. https://doi.org/10.1016/j.joi.2017.08.007

Aria, M., & Cuccurullo, C. (n.d.). *A brief introduction to bibliometrix.* Bibliometrix R-Package. https://www.bibliometrix.org/vignettes/Introduction_to_bibliometrix.html

Kalra, V., & Aggarwal, R. (2017). *Importance of text data preprocessing & implementation in RapidMiner* [Paper presentation]. Proceedings of the First International Conference on Information Technology and Knowledge Management, New Delhi, India. https://doi.org/10.15439/2017km46

RapidMiner. (n.d.). *Text and web mining with RapidMiner.* RapidMiner Academy. https://academy.rapidminer.com/courses/text-and-web-mining-with-rapidminer