

Identifying Emerging Trends in Scientific Texts Using TF-IDF Algorithm: A Case Study of Medical Librarianship and Information Articles

Meisam Dastani ¹, Afshin Mousavi Chelak ², Soraya Ziaei ³, Faeze Delghandi⁴

¹PhD Candidate in knowledge and information science, Payame Noor University, Tehran, Iran

²Associate Professor, Department of Knowledge and Information Science, Payame Noor University, Tehran, Iran

³Associate Professor, Department of Knowledge and Information Science, Payame Noor University, Tehran, Iran

⁴Assistant Professor, Department of Knowledge and Information Science, Payame Noor University, Tehran, Iran

*Corresponding Author: Afshin Mousavi Chelak, Associate Professor, Department of Knowledge and Information Science, Payame Noor University, Tehran, Iran. Email: mousaviaf@gmail.com

Received 2020 August 27; Accepted 2020 October 06.

Abstract

Background: Nowadays, due to the increased publication of articles in various scientific fields, identifying the publishing trends and emerging keywords in the texts of these articles is essential.

Objectives:

Thus, the present study identified and analyzed the keywords used in the published articles on medical librarianship and information.

Methods: In the present investigation, an exploratory and descriptive approach was used to analyze librarianship and information articles published in specialized journals in this field from 1964 to 2019 by applying text mining techniques. The TF-IDF weighting algorithm was applied to identify the most important keywords used in the articles. The Python programming language was used to implement text mining algorithms.

Results: The results obtained from the TF-IDF algorithm indicated that the words "Library", "Patient", and "Inform" with the weights of 95.087, 65.796, and 63.386, respectively, were the most important keywords in the published articles on medical librarianship and information. Also, the words "Catalog", "Book", and "Journal" were the most important keywords used in the articles published between the years 1960 and 1970, and the words "Patient", "Bookstore", and "Intervent" were the most important keywords used in articles on medical librarianship and information published from 2015 to 2020. The words "Blockchain", "Telerehabilit", "Instagram", "WeChat", and "Comic" were new keywords observed in articles on medical librarianship and information between 2015 and 2020.

Conclusions: The results of the present study revealed that the keywords used in articles on medical librarianship and information were not consistent over time and have undergone a change at different periods so that nowadays, this field of science has also changed following the needs of society with the advent and growth of information technologies.

Keywords: Librarianship and Information; Medical; Analysis; Keyword; Text Mining; TF-IDF

1. Background

Today, many scientific documents are being produced by researchers, professors, and students from different universities worldwide, which often contain important and useful materials and are published in textual formats. Because of the increasing number of scientific articles and the massive volume of published papers, the evaluation and review of all articles and manual extraction of information and knowledge from this huge volume of texts is challenges and even impossible. However, identifying patterns and extracting potential knowledge in large volumes of textual data is an important issue in various scientific fields

(1). Therefore, providing tools and techniques that can analyze texts by automatic assessment has led to the development of text mining, which is also known as intelligent text analysis, text data mining, or knowledge discovery from the text, and generally refers to the process of extracting desired and important knowledge and information from a non-structured text collection (2-4). Also, identifying patterns and extracting potential knowledge in large volumes of data from scientific texts, news, and studies is an important issue in various scientific fields (5-7).

The keyword extraction is one of the text mining algo-



rithms and techniques to discover knowledge from many documents (8). The keyword extraction is an important technique for retrieving documents and web pages, categorizing documents, summarizing and extracting the texts, etc. Keywords are considered to be the main elements of the representation of knowledge concepts and are normally applied to indicate the knowledge structure of research fields (9).

In previous studies, researchers have mainly focused on identifying research topics (e.g., categorizing a research topic and discovering a network community) and interpreting their results (10-14). Moreover, less attention has been paid to the process of selecting the appropriate keywords for the analysis. Also, in many studies, important and extensively used keywords have been analyzed, including the identification of the evolution of library and information research based on keyword analysis, bibliography, and citations in which important and widely used words were obtained based on the highest frequency (15), keyword analysis based on the highest frequency in articles related to the application of Web 2.0 in medicine (16), and identifying the trend of public library research in India, in which the most used words were obtained based on the highest frequency (17). Salloum et al. (18) in 2018 also identified the most used words in mobile learning articles in part of their study.

The most used and general words may be useful in giving an overview of a scientific discipline; however, they are less successful in demonstrating the exact topics of a research area. For instance, “library”, “information resources”, and “services” are the frequently used keywords in the field of digital library (DL); however, in reality, they are probably universal concepts present in many other research areas of the library and information science (LIS). Therefore, they are not an excellent representation of research topics in the field of DL (19). The Term Frequency-inverse document frequency (TF-IDF) algorithm is one of the most useful approaches to identify important keywords in texts, which is also a method to measure the importance of a term in a docu-

ment. Specifically, TF-IDF is a numerical statistic that indicates the importance level of a word relative to a document in a set of documents. The value of TF-IDF is balanced by the number of documents in the collection that contain the word, and it increases in proportion to the number of word repetitions in the document. It means that if a word appears in many texts, it is probably a common word and is not so valuable in evaluating the text (13, 20). The TF-IDF algorithm combines two aspects of a word: the importance of a word for a document and its discriminative power within the whole collection. However, the TF-IDF scale automatically assumes that words are completely independent. Extracting keywords from a text is closely related to ranking words in the text by their relevance (21).

2. Objectives

Due to the increasing trend of science and published scientific articles, reviewing and analyzing the published scientific texts is of particular importance for organizations, researchers, and scientific policymakers. Therefore, the use of text mining techniques and keyword extraction can be considered a solution to extract knowledge from a large volume of scientific texts. Accordingly, in the present study, the TF-IDF algorithm has been applied to extract and analyze important keywords from published articles on librarianship and information.

3. Methods

Text mining is the method of the present study carried out with an exploratory approach. Text mining is a data-based exploratory method applied to find patterns and trends in large data sets (22).

The statistical population included all articles published in specialized journals of medical librarianship and information indexed in PubMed and Scopus. All the published articles were then extracted by searching the titles of these journals in the PubMed and Scopus databases using an appropriate search strategy (Table 1) on 2020-01-20.

Table 1. The Search Strategy Used in This Study

Database	Search Strategies
PubMed	(((((“Journal of eScience librarianship”[Journal]) OR (“Journal of the European Association for Health Information and Libraries”[Journal])) OR “Biomedical digital libraries”[Journal]) OR “The journal of the Canadian Health Libraries Association”[Journal]) OR “Journal of the Medical Library Association: JMLA”[Journal]) OR (“Health information and libraries journal”[Journal])) OR “Journal of hospital librarianship”[Journal]) OR “Hospital libraries”[Journal]) OR “Health libraries review”[Journal]) OR “Bulletin of the Medical Library Association”[Journal])OR (“Medical library and the historical journal”[Journal]))OR “Bulletin of the Association of Medical Librarians”[Journal]) OR “Medical libraries”[Journal]) OR “Journal of Medical Internet research”[Journal]
Scopus	(ISSN (1542-4073) OR ISSN (1540-9597))

Next, the titles and abstracts of the articles published in these journals from 1964 to the end of the year 2019 were extracted from the mentioned 12,819 databases. Items such as editor tips, book reviews, content analy-

sis, and letters to the editor were excluded from all articles because the purpose of the present study was to identify the emerging keywords in articles based on the titles and abstracts of the selected articles. It is also

noteworthy that these cases were not among research articles and had no abstracts (23). After excluding these items, the sample data was reduced to the number of 7,599 articles.

The text mining operations applied in the present study included three stages of text preprocessing, text mining operations, and post-processing (24). Text preprocessing included the following operations: Data selection, categorization, feature extraction, normalization, and removing stop words. The second stage of text mining in this study was the use of the root-finding algorithm, TF-IDF word weighting, and data visualization. The knowledge was interpreted, analyzed, and visualized throughout the third stage.

The Porter root-finding algorithm was used to find the root of words before implementing the TF-IDF algorithm. This algorithm prunes the words' suffixes from the same root and finally converts different forms of words and those from the same root into a single form (25-27). For example, it converts the words Models, Modeling, and Modeled into the word Model.

Besides, for visualizing data in images, word cloud visualization was used, which is one of the most common methods applied for the graphical display of textual data, which is also useful to analyze various forms of textual data, including articles, short answers, and comments given to surveys and questionnaires (28). Word clouds provide a unique way to summarize the content of text documents (29). They are produced by words in textual data. The size of a word in the cloud is proportional to the number or importance (word weight) that the word is employed in the whole dataset.

The Python programming language and related libraries were used to implement text mining algorithms. This language is compact and multifunctional, possesses a simple syntax, develops easily, and provides the user with various libraries to work with texts (30).

4. Results

4.1. Most Important Keywords Used in Articles on Medical Librarianship and Information

Table 2 shows the 10 most important keywords of published articles on medical librarianship and information. It is noteworthy that these words had been rooted before extraction using Porter's root-finding algorithm. Moreover, Figure 1 indicates word clouds, the important words used in the published articles on medical librarianship and information. In a word cloud, words with larger sizes are of more importance than others, and the distinct colors of each word are used to better separate and display the words. Table 2 and Figure 1 demonstrate that the words librari, patient, and inform with the weights of 95.087, 65.796, and 63.386, respectively, were the most important keywords in the articles on medical librarianship and information.

Table 2. The Most Important Keywords in Articles on Medical Librarianship and Information Extracted by Term Frequency-Inverse Document Frequency (TF-IDF) Algorithm

	Word	Weight
1	Librari	95.08718
2	Patient	65.79694
3	Inform	63.38658
4	Journal	61.94757
5	Librarian	59.67304
6	Health	59.46848
7	Servic	58.98901
8	Medic	56.89672
9	Scienc	52.78193
10	Intervent	50.42498



Figure 1. The word cloud of the most important keywords in articles on medical librarianship and information extracted by the term frequency-inverse document frequency (TF-IDF) algorithm

4.2. Most Important Keywords Used in Articles Published in Journals Over Time

Table 3 indicates the most important keywords applied in articles on medical librarianship and information at time intervals along with the TF-IDF weight of each word and the shape of their word clouds. One hundred important words used in the published articles on medical librarianship and information are shown in the word cloud of each time interval, in which the larger size of each word indicates the importance of that word. For instance, the word Catalog with a weight of 3.065228 showed the most application in the period of 1960 to 1970, and the word Patient with a weight of 26.27959 was mostly used from 2015 to 2020.

Table 5 indicates 15 emerging keywords in the published articles in journals of medical librarianship and information from 2010 to 2015. The data in Table 5 shows that the words tweet, iPad, and mhealth with the weights of 3.676,

2.430, and 2.273, respectively, were the most important emerging keywords in the published articles in journals of medical librarianship and information in the period of 2010 to 2015.

Table 5. The Emerging Keywords in Published Articles on Medical Librarianship and Information From 2010 to 2015

	Words	Weight TF-IDF			
		2000 - 2005	2005 - 2010	2010 - 2015	2015 - 2020
1	Tweet	0	0	3.676234	7.330024
2	Ipad	0	0	2.430934	1.133248
3	Mhealth	0	0	2.273765	5.122682
4	Crowdsourc	0	0	1.644879	1.983541
5	App	0	0	1.542563	4.20206
6	Mooc	0	0	1.405142	1.116956
7	Clinicalkey	0	0	1.216506	0.450108
8	Youtub	0	0	1.154643	1.895998
9	Avatar	0	0	1.132791	1.090212
10	Altmetr	0	0	1.01116	0.868911
11	Quantum	0	0	0.970428	0
12	Isabel	0	0	0.950463	0
13	Rfid	0	0	0.900783	0
14	Webpag	0	0	0.776538	0.899917
15	Wikipedia	0	0	0.693369	3.127697

Table 6 presents 15 emerging keywords in the published articles in journals of medical librarianship and information from 2015 to 2020. The data in Table 6 indicates that the words blockchain, telerehabilit, and Instagram with

the weights of 4.136, 1.608, and 1.376, respectively, were the most important emerging keywords in the published articles in journals of medical librarianship and information in the period of 2015 to 2020.

Table 6. The Emerging Keywords in Published Articles on Medical Librarianship and Information From 2015 to 2020

	Words	Weight TF-IDF			
		2000 - 2005	2005 - 2010	2010 - 2015	2015 - 2020
1	Blockchain	0	0	0	4.136238
2	Telerehabilit	0	0	0	1.608465
3	Instagram	0	0	0	1.376558
4	Wechat	0	0	0	1.289766
5	Comic	0	0	0	1.219835
6	Chatbot	0	0	0	1.120755
7	Orcid	0	0	0	1.049601
8	Exergam	0	0	0	0.95089
9	Anatomag	0	0	0	0.946656
10	Pokemon	0	0	0	0.909836
11	Jove	0	0	0	0.75018
12	Mobilepdr	0	0	0	0.739713
13	Infograph	0	0	0	0.734067
14	Hospitalist	0	0	0	0.732413
15	Cybersecur	0	0	0	0.638965

5. Discussion

The results of evaluating the most important keywords indicated that the words library, patient, inform, journal, librarian, health, service, medic, scienc, and intervent are 10 essential keywords with the highest TF-IDF weights. The analysis of these keywords provides an overview of the main topics of published articles on medical librarianship and information. Funk has also indicated that the words library, information, medical, health, libraries, librarians, services, research, hospital, and time were the most frequently used words in the articles published in the *Bulletin of the Medical Library Association* and *Journal of the Medical Library Association* from 1960 to 2010. Funk also claimed that the words used in these articles were in the categories of environment, management, technology, and research (31). Sahoo and Bhui (17) showed that the words library, public, study, development, district, and services were the most frequently used words in the Indian public library.

An evaluation of the most important keywords employed in the published articles on medical librarianship and information indicated that they have had changes in different periods, and the use of words in the articles did not follow the same trend. For instance, Catalog, Book, and Journal were the most important keywords used in the published articles between 1960 and 1970. Also, the keywords Patient, Librarian, Intervent, Librarian, Inform, Review, Social, Care, Student, and Onlin were the most important keywords used in the published articles in the period of 2015 to 2020. Therefore, it can be seen that the importance of the keywords used in the published articles changes in different periods, and other words are considered as the most extensively applied and the most important keywords.

Chang et al. showed that the words Search, and Online Data Retrieval were the most frequently used keywords from 1995 to 1999. They also indicated that the World Wide Web and online data retrieval were the most repetitive keywords from 2000 to 2004. It was also shown that the Word Citation Analysis was the most frequently used keyword from 2005 to 2009, and the word Scientific Publication was the most repetitive keyword from 2010 to 2014 in the studies on medical librarianship and information (15).

The results of the present study also indicated that the words Patient and Care were important keywords since 2000 among other important and specialized words used in the articles on medical librarianship and information. These two important keywords, along with the word Information, present the significance of health information for patients in the field of health care. There is no doubt that patients need to have access to high-quality information, and the ability to understand and interpret this type of information, which is the task of medical librarianship and informants, has become more prominent in recent years (32, 33). Medical libraries are always health care pro-

viders and support the information needs of consumers. Library science is naturally a socially friendly field, which is always adapted to the changing needs and behaviors of its user community. Medical librarians also know how to handling, organizing, searching, finding, locating, and delivering accurate, reliable, and relevant information. They know how and when they can provide information to physicians and patients and how to discover the correct information from the vast amount of data (34). In general, the growth of health information-related issues has been influenced by the advent of the Internet (12).

The results revealed some changes in the used keywords over time. The trend of keywords used in the articles on medical librarianship and information demonstrates that although these articles have retained the main format and subject of the scientific field, evaluating the emerging keywords in this area shows the movement toward information technology.

The results of the present research conducted on emerging keywords in the published articles in the journals of medical librarianship and information revealed that the keywords Blockchain, telerehabilit, Instagram, WeChat, comic, chatbot, orcid, exergam, anatomag, and pokemon were introduced to the articles and studies published in journals on medical librarianship and information from 2015 to 2020, which represents a movement toward the information and communication technologies and new topics in information and digital technology. The words related to Web 2.0 technology and social networks are among the emerging keywords in recent years. In this regard, the words weblog and webcast from 2005 onwards, the words tweet and Wikipedia from 2010 onwards, and the words related to instant messengers such as Instagram and WeChat from 2015 onwards have been observed in the medical librarianship and information literature. Accordingly, Boudry (16) has also indicated that the articles related to Web 2.0 were published in medical sciences from 2002 to 2012, and the word weblog was the most frequently used keyword in the published articles on Web 2.0 in medical sciences.

In general, it can be noted that the advent of new information technologies has changed the process of storing, retrieving, managing, and distributing health information. Consequently, the conventional tasks and services of medical librarianship and information have been evolved. In other words, nowadays, the activities of medical librarianship and information are beyond the libraries, and they possess the capability to work and provide services in various health areas such as hospitals, libraries, research centers, publishing centers, scientometric centers, specialized journals, growth centers, and knowledge-based enterprises (35).

5.1. Conclusions

The published topics in science are evolving, new topics are emerging, and old topics are becoming obsolete over

time. The scientific field of medical librarianship and information is no exception. According to the results of the present study on lexical and thematic changes, lexical changes and the emergence of new keywords and topics are noticeable in this scientific field. The trend of changes in the use of different keywords in the articles on medical librarianship and information, as well as the identification of emerging keywords in these articles, indicated that this field has changed according to the needs of society, as well as the emergence and growth of information and communication technologies.

References

- Hashimi H, Hafez A, Mathkour H. Selection criteria for text mining approaches. *Comput Hum Behav*. 2015;51:729-33. doi:10.1016/j.chb.2014.10.062.
- Rebholz-Schuhmann D, Oellrich A, Hoehndorf R. Text-mining solutions for biomedical research: enabling integrative biology. *Nat Rev Genet*. 2012;13(12):829-39. doi:10.1038/nrg3337. [PubMed:23150036].
- Jensen PB, Jensen LJ, Brunak S. Mining electronic health records: towards better research applications and clinical care. *Nat Rev Genet*. 2012;13(6):395-405. doi:10.1038/nrg3208. [PubMed:22549152].
- Rodriguez-Esteban R, Bundschuh M. Text mining patents for biomedical knowledge. *Drug Discov Today*. 2016;21(6):997-1002. doi:10.1016/j.drudis.2016.05.002. [PubMed:27179985].
- Kao A, Poteet SR. *Natural Language Processing and Text Mining*. Springer Science & Business Media; 2007.
- Lee S, Lee S, Seol H, Park Y. Using patent information for designing new product and technology: keyword based technology roadmapping. *R&D Manag*. 2008;38(2):169-88. doi:10.1111/j.1467-9310.2008.00509.x.
- Hung J-L, Zhang K. Examining mobile learning trends 2003-2008: a categorical meta-trend analysis using text mining techniques. *Journal of Computing in Higher Education*. 2011;24(1):1-17. doi:10.1007/s12528-011-9044-9.
- Rajman M, Besançon R. Text Mining - Knowledge extraction from unstructured textual data. *Advances in Data Science and Classification*; 1998. p. 473-80.
- Su H-N, Lee P-C. Mapping knowledge structure by keyword co-occurrence: a first look at journal papers in Technology Foresight. *Scientometrics*. 2010;85(1):65-79. doi:10.1007/s11192-010-0259-8.
- Soleimani NA, Salajegheh M, Tayyebi NE. Clustering Scientific Articles based on the K-means Algorithm Case Study: Iranian Research Institute for Information Science and Technology (Iran-Doc). *Iran J Inform Process Manag*. 2019;34(2):874-96.
- Lamba M, Madhusudhan M. Mapping of topics in DESIDOC Journal of Library and Information Technology, India: a study. *Scientometrics*. 2019;120(2):477-505. doi:10.1007/s11192-019-03137-5.
- Figuerola CG, García Marco FJ, Pinto M. Mapping the evolution of library and information science (1978-2014) using topic modeling on LISA. *Scientometrics*. 2017;112(3):1507-35. doi:10.1007/s11192-017-2432-9.
- Kim YM, Delen D. Medical informatics research trend analysis: A text mining approach. *Health Informatics J*. 2018;24(4):432-52. doi:10.1177/1460458216678443. [PubMed:30376768].
- Moon C-S, Ohk* K, Kim M-J. Trend Analysis by Using Text Mining of Journal Articles Regarding Consumer Policy. *New Physics: Sae Mulli*. 2017;67(5):555-61. doi:10.3938/npsm.67.555.
- Chang Y-W, Huang M-H, Lin C-W. Evolution of research subjects in library and information science based on keyword, bibliographical coupling, and co-citation analyses. *Scientometrics*. 2015;105(3):2071-87. doi:10.1007/s11192-015-1762-8.
- Boudry C. Web 2.0 applications in medicine: trends and topics in the literature. *Med 2 0*. 2015;4(1):e2. doi:10.2196/med20.3628. [PubMed:25842175]. [PMC5823613:PMC5823613].
- Sahoo S, Bhui T. Trend of Public library research in India: a bibliometric study. *Lib Philos Pract*. 2018.
- Salloum SA, Al-Emran M, Monem AA, Shaalan K. Using Text Mining Techniques for Extracting Information from Research Articles. *Intelligent Natural Language Processing: Trends and Applications*; 2018. p. 373-97.
- Chen C. CiteSpace II: Detecting and visualizing emerging trends and transient patterns in scientific literature. *J Am Soc Inf Sci Technol*. 2006;57(3):359-77. doi:10.1002/asi.20317.
- Frakes WB, Baeza-Yates R. *Information retrieval: data structures and algorithms*. Prentice-Hall, Inc.; 1992.
- Wartena C, Brussee R, Slakhorst W. Keyword Extraction Using Word Co-occurrence. 2010 Workshops on Database and Expert Systems Applications. 2010. p. 54-8.
- Hastie T, Tibshirani R, Friedman J. *The elements of statistical learning: data mining, inference, and prediction*. Springer Science & Business Media; 2009.
- Dancy-Scott N, Dutcher GA, Keselman A, Hochstein C, Coptly C, Ben-Senia D, et al. Trends in HIV Terminology: Text Mining and Data Visualization Assessment of International AIDS Conference Abstracts Over 25 Years. *JMIR Public Health Surveill*. 2018;4(2):e50. doi:10.2196/publichealth.8552. [PubMed:29728344]. [PMC5960041:PMC5960041].
- Zhang Y, Chen M, Liu L. A review on text mining. 2015 6th IEEE International Conference on Software Engineering and Service Science (ICSESS). 2015. p. 681-5.
- Miner G, Elder IV J, Fast A, Hill T, Nisbet R, Delen D. *Practical text mining and statistical analysis for non-structured text data applications*. Academic Press; 2012.
- Abuhay TM, Kovalchuk SV, Bochenina KO, Kamps G, Krzhizhanovskaya VV, Lees MH. Analysis of Computational Science Papers from ICCS 2001-2016 using Topic Modeling and Graph Theory. *Proced Comput Sci*. 2017;108:7-17. doi:10.1016/j.procs.2017.05.183.
- Porter ME. An algorithm for suffix stripping. *Program*. 1980;14(3):130-7. doi:10.1008/eb046814.
- DePaolo CA, Wilkinson K. Get Your Head into the Clouds: Using Word Clouds for Analyzing Qualitative Assessment Data. *Tech Trends*. 2014;58(3):38-44. doi:10.1007/s11528-014-0750-9.
- Sanders AM, Stehle JR, Jr., Blanks MJ, Riedlinger G, Kim-Shapiro JW, Monjzab AM, et al. Cancer resistance of SR/CR mice in the genetic knockout backgrounds of leukocyte effector mechanisms: determinations for functional requirements. *BMC Cancer*. 2010;10:121. doi:10.1186/1471-2407-10-121. [PubMed:20356394]. [PMC2861034:PMC2861034].
- Rehurek R, Sojka P, editors. Software framework for topic modeling with large corpora. In Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks; 2010. Citeseer.
- Funk ME. Our words, our story: a textual analysis of articles published in the Bulletin of the Medical Library Association/ Journal of the Medical Library Association from 1961 to 2010. *J Med Libr Assoc*. 2013;101(1):12-20. doi:10.3163/1536-5050.101.1.003. [PubMed:23405042]. [PMC3543134:PMC3543134].
- Kelly K. *Health science libraries: future trends*. 2009.
- Kurata K, Miyata Y, Ishita E, Yamamoto M, Yang F, Iwase A. Analyzing library and information science full-text articles using a topic modeling approach. *Proceed Assoc Inf Sci Technol*. 2019;55(1):847-8. doi:10.1002/pr2.2018.14505501143.
- Gavani VZ, Mohan VV. Application of Web 2.0 Tools in Medical Librarianship to Support Medicine. *Webology*. 2008;5(1).
- Ashrafi RH, Hodhodinezhad N, Shahrzadi L, Soleymani M. A study on the novel services of medical librarians in health information services: a narrative review. *Health Inf Manag*. 2017;13(6):438-44.