IAES International Journal of Artificial Intelligence (IJ-AI)
Vol. 11, No. 3, September 2022, pp. 1057~1065
ISSN: 2252-8938, DOI: 10.11591/ijai.v11.i3.pp1057-1065

A text mining and topic modeling based bibliometric exploration of information science research

Tipawan Silwattananusarn[1], Pachisa Kulkanjanapiban[2]
[1]Faculty of Humanities and Social Sciences, Prince of Songkla University, Mueang Pattani, Thailand
[2]Khunying Long Athakravisunthorn Learning Resources Center, Prince of Songkla University, Hat Yai, Thailand

Article Info

Article history:

Received Oct 1, 2021
Revised May 19, 2022
Accepted Jun 17, 2022

Keywords:

Bibliometrics
Latent dirichlet allocation
Research trends
Text mining
Topic modeling

ABSTRACT

This study investigates the evolution of information science research based on bibliometric analysis and semantic mining. The study discusses the value and application of metadata tagging and topic modeling. Forty-two thousand seven hundred thirty-eight articles were extracted from Clarivate Analytic's Web of Science Core Collection 2010-2020. This study was divided into two phases. Firstly, bibliometric analyzes were performed with VOSviewer. Secondly, the topic identification and evolution trends of information science research were conducted through the topic modeling approach latent dirichlet allocation (LDA) is often used to extract themes from a corpus, and the topic model was a representation of a collection of documents that is simplified using topic-modeling-toolkit (TMT). The top 10 core topics (tags) were information research design, information health-based, model data public, study information studies, analysis effect implications, knowledge support web, data research, social research study, study media information, and research impact time for the studied period. Not only does topic modeling assist in identifying popular topics or related areas within a researcher's area, but it may be used to discover emerging topics or areas of study throughout time.

Corresponding Author:

Pachisa Kulkanjanapiban
Khunying Long Athakravisunthorn Learning Resources Center, Prince of Songkla University
15 Karnjanavanich Rd., Hat Yai, Songkhla 90110, Thailand
Email: pachisa.ku@psu.ac.th

1. INTRODUCTION

Many businesses generate large amounts of text and image data, which they store. This makes it challenging to manage large amounts of data and extract relevant information for decision-making. New tools and techniques are required to manage this explosion of electronic documents better. Topic modeling is one of the new techniques for finding patterns of words in many documents developed in the last decade by machine learning and statistics for effective information retrieval. Numerous topic modeling applications include tag recommendation, text categorization, keyword extraction, and similarity search. Text mining, information retrieval, and statistical language modeling are just a few applications.

Several techniques for building knowledge models based on topics extracted using text mining procedures have been developed in recent years. Moro *et al.* [1] describes that latent semantic analysis and topic modeling are two of the most used techniques. The former is a natural language processing technique that analyzes relationships between textual terms and documents founded on the notion that words with similar meanings will appear incomparable material. At the same time, the latter takes as input the structure obtained by text mining, with the relevant terms and their frequency gathered into an orderly structure in which the documents are split into subjects [2]. Both techniques generate themes that summarize the body of information included in the documents, resulting in a literature synthesis.

Critical studies on applying latent dirichlet allocation (LDA), topic modeling, and text mining have been reviewed. Text mining enables the identification and retrieval of high-quality new semantic information through the automated assessment of textual patterns and trends in the literature under review, which provides a more in-depth understanding of the contents than a fundamental word count analysis [3]. A topic model is a valuable tool for text mining to identify research topics and hotspots in scientific and technological papers. The LDA model is popular in various fields. Some articles apply LDA, topic modeling, and text mining, such Lee and Cho [4] proposed a web document ranking method using topic modeling for effective information collection and classification. Allahyari *et al.* [5] described several of the most fundamental text mining tasks and techniques in biomedical and health care domains. The paper aims to identify major academic branches and detect research trends in design research using text mining techniques. Lubis *et al.* [6] proposed the topic modeling approach on helpful subjective reviews. Subeno *et al.* [7] aimed to determine the optimal number of corpus topics in the LDA method. The proposed approach in [8] can cluster the text documents of research papers into meaningful categories which contain a similar scientific field using a title, abstract, and keywords of the paper to the categories topics. Chauhan and Shah [9] introduced the preliminaries of the topic modeling techniques and reviewed its extensions and variations. The research in [10] is to survey the body of research revolving around big data and analytics in hospitality and tourism using bibliometric techniques, network analysis, and topic modeling. Chen *et al.* [11] used the LDA model to extract the subject of each paper published by authors in 239 educational journals from China and the United States during 20 years (2000-2019). Suominen *et al.* [12] uses LDA to create topic-based linkages between publications and patents based on the semantic content in the documents. In empirical investigations [13], topic modeling has analyzed textual data. Between 2009 and 2020, the study analyzed subject modeling in 111 publications from the top ten ranked software engineering journals. The most common topic modeling techniques are LDA and LDA-based strategies.

The importance and complexity of information science issues have drawn scholars from various disciplines, including bibliometrics, social media analytics, text mining, machine learning, knowledge management, knowledge sharing, qualitative research, social science. Although information science has received much attention in recent years, few studies have attempted to conduct a large-scale evaluation of academic literature on the subject. One of the essential functions of information science research is that it aids in identifying a variety of current public policy issues. This function responds to the growing need for information science in rational decision-making. Building conceptual frameworks of relevant information science research are required to make more reasonable policies. In order to assist in the deployment of a rational information science development plan, a topic modeling-based bibliometrics examination of peer-reviewed literature representing information science research with 42,738 target articles published between 2010 and 2020 was conducted.

This study combines the bibliometric method and LDA model to analyze the development trend of information science research from statistical analysis and text mining. This research fills in the blanks of existing literature by employing a technique that examines disciplines and applies them as tags to information science journals. This research benefits information retrieval, the semantic web, and linked data. This publication will be helpful to researchers, documentation and information professionals, students, and others interested in the field.

2. RESEARCH METHOD

This study considers core journals in information science and library science from 2010 to 2020 and provides a method to identify the disciplinary identity in information science research. Each document's title, abstract, and keywords were used for the topic analysis. Forty-two thousand seven hundred thirty-eight articles published from 2010-2020 were collected as shown in Table 1. LDA topic modeling was used to further process and analyze the data sets. The topics were modeled using the LDA modeling technique.

Recently, the availability of accessible software allows researchers to make use of topic modeling and other text mining methodologies, making these methods more approachable. This study's modeling process is based on the topic-modeling-toolkit (TMT) package [14]. Text mining front-end additions, such as the R package and *VOSviewer* [15], are required by the topic models package. The use of Microsoft Excel and PowerBI to aid in the processing and plotting of statistical data.

2.1. Data collection

The data used in this investigation was obtained from the Science Citation Index Expanded (SCI-Expanded), Social Sciences Citation Index (SSCI), and Arts and Humanities Citation Index (AHCI) databases in June 2021, provided by the Institute for Scientific Information (ISI). The search time frame is set from 2010 to 2020. The following search query selected: *SU=Information Science* AND DT=(Article) AND PY=2010-2020 Refined By: Web of Science Index: Social Sciences Citation Index (SSCI) or Science Citation Index Expanded (SCI-EXPANDED) or Arts and Humanities Citation Index (A&HCI)* to search articles published between 2010 and 2020 in the online SCI-Expanded, SSCI, and AHCI databases. The study was restricted to research papers only articles. Proceedings papers, early access, book chapters, retracted publications were excluded. A total of 42,738 publications were collected.

Table 1. Year-wise distribution of information science articles from 2010-2020

Year of publication	Number of articles
2010	3292
2011	3498
2012	3548
2013	3684
2014	3855
2015	3933
2016	4126
2017	4129
2018	4086
2019	4149
2020	4438
Total	42738

2.2. Latent dirichlet allocation (LDA)

LDA is a legal term that refers to (latent dirichlet allocation) [2]. In a text collection, topic modeling is a method for analyzing the distribution of semantic word clusters or "topics." It can explore a corpus' content and generate content-related features for computational text classification. Topic modeling is thus largely independent of language and orthographic convention because it relies solely on the analyzed texts; it does not use additional sources of information such as dictionaries or external training data. It is solely based on a statistical analysis of symbol co-occurrence (at the word level), then translated into possible semantic relationships [2], [16]–[20]

This paper focuses on applying LDA [2] to model the subjects from the corpus of Information Science articles based on dirichlet distribution. Each article is represented in this study as a pattern of LDA topics. LDA automatically infers the topic mentioned in a collection of articles, and these topics can be used to summarize and organize the articles. Bags of words per article are the variables observed, while the hidden random variables are the topic distribution of each article. The observable variables in LDA are: i) the bags of words per article based on probabilistic modeling. LDA's fundamental purpose is to compute the posterior of hidden variables given the observable variables' values. Articles with similar themes will employ similar groupings of words, ii) articles are a probability distribution over latent topics, and iii) topics are probability distributions over words [21] as shown in Figure 1.

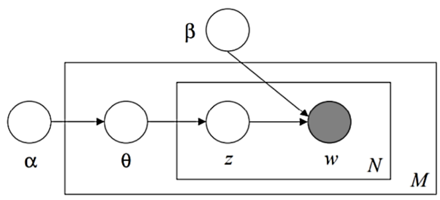


Figure 1. LDA is represented graphically in this model (source [2])

Figure 1 demonstrates the LDA model as a probabilistic graphical model. The LDA representation has three layers. The variables shown in the figure are defined [2], [21]:
α - parameter of Dirichlet prior on the per-document topic distribution
β - parameter of Dirichlet prior on per-topic word distribution
θ - topic distribution for the document, *d*
z - the topic for the *n*th word in the document, *d*
w - is the specific word
N - total number of words
M - total number of documents in the corpus

LDA is a corpus-based generative probabilistic model [2]. The core idea is that documents are represented as random mixes over latent topics, with each subject defined by a distribution of words. LDA assumes the following generative process for each document *w* in a corpus *D*:

- Choose $N \sim$ Poisson(ξ)
- Choose $\theta \sim$ Dir(α)
- For each of the *N* words w_n: i) choose a topic $z_n \sim$ Multinomial(θ) and ii) choose a word w_n from $p(w_n \mid z_n, \beta)$, a multinomial probability conditioned on the topic z_n

The outside box in Figure 1 represents documents in the LDA model, while the inner box represents the documents' repeated selection of subjects and terms. The alpha (α) and the beta (β) are corpus-level parameters assumed to be sampled once during the corpus generation process. The variables θ_d are document-level variables sampled only once for each document. The word-level variables z_{dn} and w_{dn} are sampled once for each word in each document.

2.3. Topic-modeling-toolkit (TMT)

Finding topics in a document is known as topic modeling [22]. The co-occurrence of terms in a document is one method of detecting the presence of a topic in a document. Topic models make analyzing large amounts of unlabeled text simple. A topic comprises a group of words that appear together regularly. Using contextual signals, topic models can connect words with similar meanings and discriminate between words with various meanings.

A topic model is a representation of a collection of documents that is simplified. TMT [14], [23], [24] is a topic modeling software that associates words with subject labels so that words that frequently appear in the same documents are more likely to have the same label applied to them. It can find similar themes in a collection of documents and trends in discourse through time and across borders. TMT is a graphical interface tool for LDA topic modeling [25]. All 42,738 articles were converted into text format and then processed using TMT. In the toolkit, the following parameters were being fixed for the study: i) number of topics: 20, ii) number of iterations: 400, iii) number of topic words to print: 20, iv) interval between hyperprior optimizations: 10, and v) number of training threads: 4.

2.4. Text mining functionality in *VOSviewer*

VOSviewer is a software application that creates maps based on network data and then visualizes and explores these maps [26], [27]. The *VOSviewer* functionality can create maps based on network data and visualize and explore maps. *VOSviewer* can be used to create, visualize, and explore maps based on any network data, in addition to analyzing bibliometric networks. Van Eck and Waltman [15] presents the text mining functionality of *VOSviewer* supports the creation of term maps from a corpus of texts. A term map is a two-dimensional map in which words are arranged so that the distance between two terms may be read to measure their relatedness. The closer two terms are related to each other, the smaller the distance between them. The co-occurrences of terms in documents are used to determine their relatedness. Titles, abstracts, and full texts in publications, patents, and newspaper articles are examples of these documents.

To create a term map based on a corpus of documents, *VOSviewer* identifies the following processes:

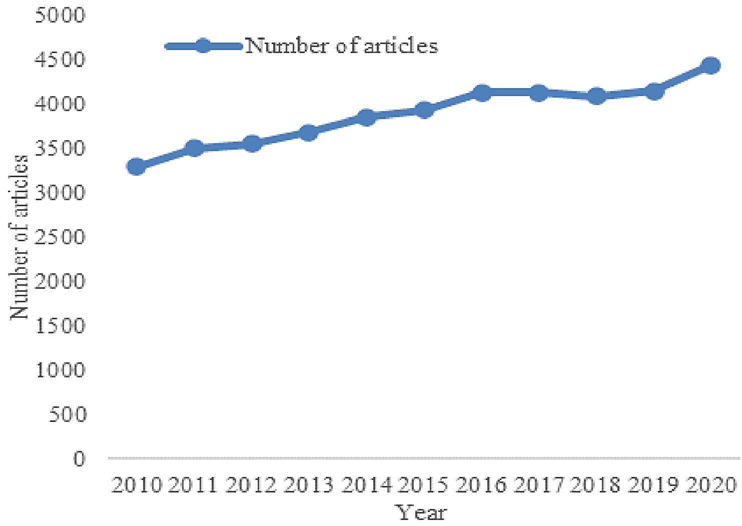
- Identification of noun phrases by: i) performing part-of-speech tagging, ii) using a linguistic filter to identify noun phrases, and iii) converting plural noun phrases into singular ones.
- Selection of the most relevant noun phrases by: i) determining the distribution of co-occurrences overall noun phrases, ii) comparing this distribution with the overall distribution of co-occurrences over noun phrases, iii) grouping noun phrases in co-occurrence with a high relevance together into clusters. Each cluster can be viewed as a separate topic
- Mapping and clustering of the terms
- The results of the mapping and clustering are visualized.

3. RESULTS AND DISCUSSION

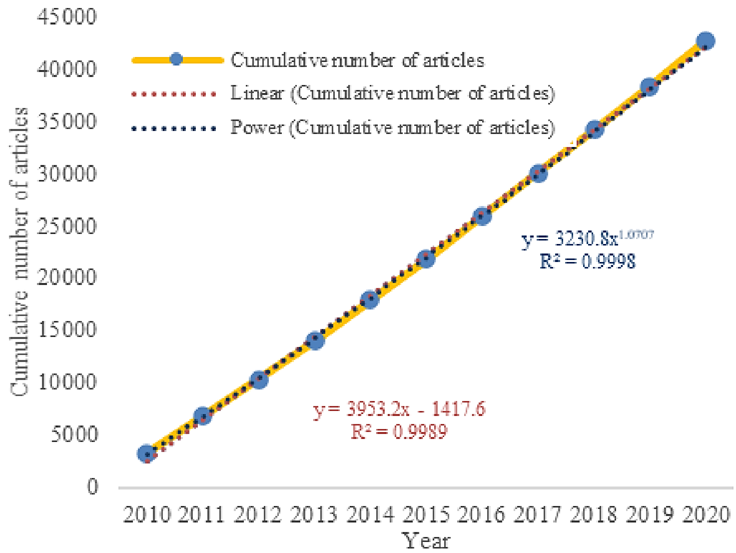
3.1. Bibliometric analysis

3.1.1. Publication analysis

The number of publications in information science has increased significantly over the last decade, and in the coming years, this pattern is expected to continue. As illustrated in Figure 2(a), there is a noticeable increase from 2010-2015 due to growing concerns about information science research. The difference in trends between the two sub-periods, 2010-2015 and 2016-2020, is notable. As a result, the relationship between the annual cumulative number of articles and the publication year for the two sub-periods was described using linear and power models, respectively. Figure 2(b) depicts the overall trend in the number of articles. The linear curve fitting result is $y = 3953.2x - 1417.6$, and the power curve fitting result is $y = 3230.8x^{1.0707}$, where y stands for the cumulative number of articles and x stands for the publication year. Both curves fit the observed data points well with high correlation coefficients ($R^2=0.9989$ for 2010-2015, and $R^2=0.9998$ for 2016-2020). Since 2015, the power model has shown a rapid increase in the number of articles in information science.



(a)



(b)

Figure 2. The trend of (a) number of articles and (b) cumulative number of articles from 2010 to 2020

The trend also spans many general information science and technology journals and more specialized outlets. As Table 2 shows, the most significant number of relevant articles published by general information science and technology appears in *Scientometrics* [28] (3175), followed by *Journal of the American Medical Informatics Association* [29] (1853), *Qualitative Health Research* [30] (1645), and *Journal of Health Communication* [31] (1273). Among the top five publication outlets emphasizes the importance of health for domain-specific informatics scholars. Furthermore, publications in *Telecommunications Policy* [32], *Government Information Quarterly* [33], and *Journal of Documentation* [34] suggest context-specific considerations of information science research.

Table 2. Top 20 publication outlets and their respective number of articles on information science

No	Name of Journal	Year of Publication										
		10	11	12	13	14	15	16	17	18	19	20	Total
1	Scientometrics	207	215	214	249	299	343	291	357	350	265	385	3175
2	Journal of the American Medical Informatics Association	107	144	172	200	200	157	169	148	182	165	209	1853
3	Qualitative Health Research	137	133	137	140	143	136	162	172	166	155	164	1645
4	Journal of Health Communication	93	107	111	114	112	180	155	118	105	85	93	1273
5	International Journal of Geographical Information Science	89	96	110	122	125	102	120	111	106	108	98	1187
6	Journal of the Association for Information Science and Technology	-	-	-	-	184	185	215	191	114	101	103	1092
7	Professional De La Information	74	86	82	66	71	88	90	113	113	120	168	1071
8	International Journal of Information Management	55	59	58	87	75	70	99	85	100	142	203	1033
9	Information Processing and Management	56	64	80	89	50	63	72	73	72	140	237	996
10	Telematics and Informatics	-	23	37	38	53	78	93	185	172	93	91	863
11	Journal of Informetrics	59	60	70	93	82	80	79	83	82	71	77	836
12	Journal of Academic Librarianship	56	56	43	71	73	95	84	63	96	78	104	819
13	Journal of Knowledge Management	57	57	54	53	62	68	64	78	81	100	95	769
14	Information & Management	41	44	36	60	86	75	65	80	75	78	102	742
15	Library Journal	59	72	63	79	70	74	68	78	78	54	43	738
16	Telecommunications Policy	60	76	61	72	62	60	65	74	61	55	84	730
17	Journal of the American Society for Information Science and Technology	176	184	178	180	-	-	-	-	-	-	-	718
18	Government Information Quarterly	47	50	70	60	70	46	67	53	67	67	71	668
19	Electronic Library	56	50	48	48	51	69	59	69	66	63	51	630
20	Journal of Documentation	39	43	41	40	53	61	59	65	69	75	67	612

3.1.2. Co-occurrence keywords analysis

Keyword search terms are vocabularies that can locate an article in abstracting and indexing databases. Trends and topics of interest can be discovered using keyword analysis. The *VOSviewer* analyzed all keywords in documents, including author's and index keywords, at a five frequently used keywords threshold. The most used keywords were mapped out as shown in Figure 3. Keywords with similar colors were grouped. The size of each circle in the cluster represented the proportion of citations for that subject's keywords. Larger circles and map labels signified greater relevance and significance.

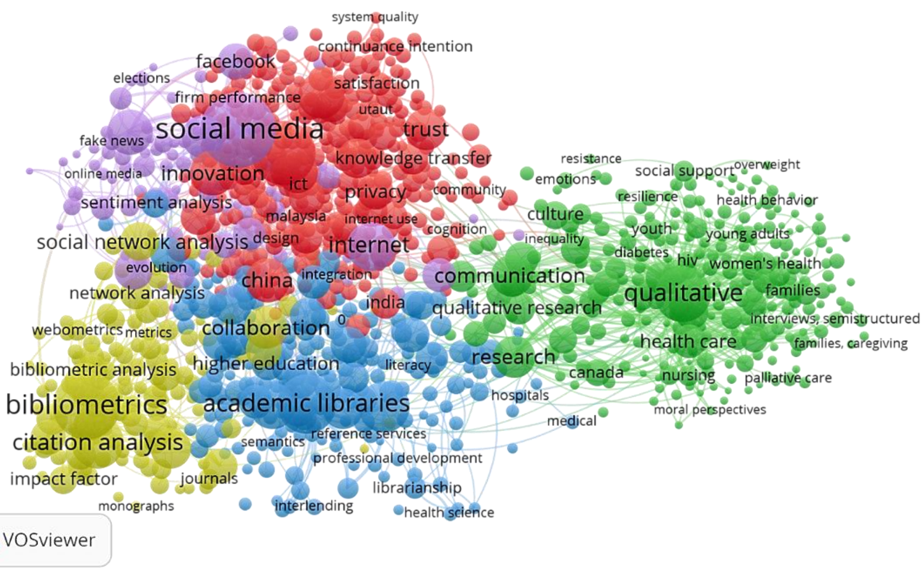


Figure 3. Co-occurrence network map of most frequently used keywords in information science research

As shown in Figure 3, clusters were differentiated by five colors: red, green, blue, yellow, and purple. The central cluster in red included the keywords "knowledge management," "innovation," "knowledge sharing," "information technology," and "big data." The keywords made up the second cluster, highlighted in green "qualitative," "communication," "decision making," "health care," and "culture." The most commonly used keywords in the third blue cluster were "academic libraries," "information literacy," "machine learning," "information retrieval," and "natural language processing." The fourth cluster in yellow included the keywords of "bibliometrics," "citation analysis," "social network analysis," "h-index," and "research evaluation." The fifth cluster shown in purple consisted of the keywords "social media," "social networks," "content analysis," "web 2.0," and "Twitter."

3.2. Topic modeling analysis

3.2.1. Topic identification

Statistics of terms that frequently appear in a collection of abstracts can provide a primary picture of a research field. The latent intellectual concepts in the literary corpus were discovered using the LDA model. Table 3 summarizes the LDA results generated by the TMT. Each topic's top 10 frequent terms for the ten years are organized in descending order according to their probability values.

Table 3. Top 10 frequent terms according to their probability values

Topic	Potential topic	List of topics
0	Information research design	information research design-based paper process findings show international study data results in level related language government network implementation potential evaluate
1	Information health-based	information health-based digital library systems result in authors related found knowledge articles research examine attention discussed patient publications community aims
2	Model data public	model data public library methodology system paper survey context factors mobile information approaches academic sharing methods order risk analyzed results
3	Study information studies	study information studies paper research data Elsevier technology tools source topic results users high methods influence groups measure common capital
4	Analysis effect implications	analysis effect implications information services study case provide future article individual libraries business online applied social records citation theoretical
5	Knowledge support web	knowledge support web information research data practical models based analysis scientific role practice conducted quality significantly rights higher limited cited
6	Paper data research	paper data research management purpose online analysis development libraries significant study years work results performance users two authors decision user
7	Social research study	social research study characteristics result in experience software performance resources education media organizational assessment understand journals studies patterns institutional citation reference
8	Study media information	study media information control processes article sources examined behavior librarians increase quality knowledge significant dimensions documents published factors paper collection
9	Research impact time	research impact time data method social approach analysis findings number information collected results journals knowledge set trust service level

The following are some examples of interpretations from Table 3:

- Topic 0 contains words like "information," "research," "design," "findings," "results," and thus apparently discusses the role of information research design in increasing collaboration implementation. It also includes words like "language," "government," "network." Indicating that novel information science technology, such as natural language processing (NLP), data governance, social media analysis, text mining, social media mining, becomes more critical in the development research strategy of information science.
- Topic 1 focuses on information health-related concerns in information science and the effects of information health-based digital library systems on patient community knowledge and well-being.
- Topic 3 and topic 6 contain words like "paper," "research," "data," "tools source," "online analysis," "topic results," "users," "methods," and "performance." Topics 3 and 6 discuss paper research data issues related to information science. However, different from topic 3, topic 6 focuses on management purpose and online analysis development significant study. An essential issue in this topic is the study results in performance influence users.
- Topic 4 refers to the effects of information analysis on information services and contains words like "analysis," "implications," "information service," and "business online." Topic 4 frequently uses terms like "social records," "social approach analysis," and "citation theoretical."
- Topic 9 also addresses social approach analysis with the highest frequent term of "analysis effect implications." However, unlike topic 4, topic 9 focuses on the research impact time and social approach analysis findings in analysis effect implications. Terms like "research impact," "data method," "social approach," "journal knowledge" are all research impact time. An essential issue in this topic is the analysis findings set trust service level.

– Topics 7 and 8 discuss social research study issues related to study media information. Topic 7 contains words like "characteristics," "social," "research," "study," "education," "media," "experience," "software," "performance," "organizational," "assessment," "patterns," "institutional," "citation," "reference.", and refers to the influences of social research projects on the organizational assessment and pattern institutional citation reference. Topic 8 focuses more on information control and knowledge quality level, thus, employs words like "control," "processes," "sources," "examined," "behavior," "paper collection," "factors," and "document published."

4. CONCLUSION

A topic modeling-based bibliometric exploration of information science research uses the 42738 articles collected from the SCI-Expanded, SSCI, and A&HCI databases. This investigation's findings provide a comprehensive overview, focusing on information science research topics from 2010 to 2020. From 2020 to 2020, linear and exponential relationships between the annual cumulative number of articles and publication year were obtained for the collected articles, revealing that annual article publications grow constantly. The findings of the co-occurrence map based on the author keywords of the information science research, the keywords *knowledge management, innovation, qualitative, decision making, academic libraries, machine learning, bibliometrics, citation analysis, social media, social networks* were the most co-occurrences and the hot topics in the information science research. This topic analysis shows that information research design issues appeal to scholars more than information studies themselves and that an interdisciplinary trend of information health-based research is emerging from the convergence of health science, social science, and media information. This research contributes to our understanding of information science's academic concerns over the last ten decades. It can be said that information science research is the core of current knowledge and powerfully connects with much other research in related fields. The study's findings have implications for future information policy. The rapid increase in publications indicates a significant demand for information-related research. In addition, the government should provide more funding for this research field in conjunction with the accelerated information development process. Second, because large projects in emerging economics and health science will account for most of the growth in information generation, these should learn from the experience of information science development in these areas. Furthermore, this research provides a comprehensive overview for this purpose.

ACKNOWLEDGEMENTS

This study received no specific funding from the public, private, or non-profit sectors. The authors are grateful to thank the anonymous reviewers for their careful reading of our manuscript and their many insightful comments and suggestions.

REFERENCES

[1] S. Moro, G. Pires, P. Rita, and P. Cortez, “A text mining and topic modelling perspective of ethnic marketing research,” *J. Bus. Res.*, vol. 103, pp. 275–285, 2019, doi: 10.1016/j.jbusres.2019.01.053.
[2] D. M. Blei, A. Y. Ng, and J. B. Edu, “Latent Dirichlet Allocation Michael I. Jordan,” *J. Mach. Learn. Res.*, vol. 3, pp. 993–1022, 2003, doi: 10.5555/944919.944937.
[3] V. Gupta and G. S. L. Professor, “A Survey of Text Mining Techniques and Applications,” 2009, [Online]. Available: www.alerts.yahoo.com
[4] Y. Lee and J. Cho, “Web document classification using topic modeling based document ranking,” *Int. J. Electr. Comput. Eng.*, vol. 11, no. 3, pp. 2386–2392, 2021, doi: 10.11591/IJECE.V11I3.PP2386-2392.
[5] M. Allahyari *et al.*, “A Brief Survey of Text Mining: Classification, Clustering and Extraction Techniques,” vol. 13, 2017, [Online]. Available: http://en.wikipedia.org/wiki/Statistics
[6] F. F. Lubis, Y. Rosmansyah, and S. H. Supangkat, “Topic discovery of online course reviews using LDA with leveraging reviews helpfulness,” *Int. J. Electr. Comput. Eng.*, vol. 9, no. 1, pp. 426–438, 2019, doi: 10.11591/IJECE.V9I1.PP426-438.
[7] B. Subeno, R. Kusumaningrum, and Farikhin, “Optimisation towards Latent Dirichlet Allocation: Its Topic Number and Collapsed Gibbs Sampling Inference Process,” *Int. J. Electr. Comput. Eng.*, vol. 8, no. 5, pp. 3204–3213, 2018, doi: 10.11591/IJECE.V8I5.PP3204-3213.
[8] A. A. Jalal and B. H. Ali, “Text documents clustering using data mining techniques,” *Int. J. Electr. Comput. Eng.*, vol. 11, no. 1, pp. 664–670, 2021, doi: 10.11591/IJECE.V11I1.PP664-670.
[9] U. Chauhan and A. Shah, “Topic Modeling Using Latent Dirichlet allocation: A Survey,” *ACM Comput. Surv.*, vol. 54, no. 7, 2022, doi: 10.1145/3462478.
[10] M. Mariani and R. Baggio, “Big data and analytics in hospitality and tourism: a systematic literature review,” *Int. J. Contemp. Hosp. Manag.*, vol. 34, no. 1, pp. 231–278, Jan. 2022, doi: 10.1108/IJCHM-03-2021-0301/FULL/PDF.
[11] S. Chen, S. Ren, L. Zheng, H. Yang, W. Du, and X. Cao, “A comparison study of educational scientific collaboration in China and the USA,” *Phys. A Stat. Mech. its Appl.*, vol. 585, p. 126330, Jan. 2022, doi: 10.1016/J.PHYSA.2021.126330.
[12] A. Suominen, S. Ranaei, and O. Dedehayir, “Exploration of Science and Technology Interaction: A Case Study on Taxol,” *IEEE Trans. Eng. Manag.*, vol. 68, no. 6, pp. 1786–1801, Dec. 2021, doi: 10.1109/TEM.2019.2923634.

[13] C. C. Silva *et al.*, "Topic modeling in software engineering research," *Empir. Softw. Eng. 2021 266*, vol. 26, no. 6, pp. 1–62, 2021, doi: 10.1007/S10664-021-10026-0.
[14] "GitHub - senderle/topic-modeling-tool: A point-and-click tool for creating and analyzing topic models produced by MALLET." https://github.com/senderle/topic-modeling-tool (accessed Mar. 12, 2022).
[15] N. J. van Eck and L. Waltman, "Text mining and visualization using VOSviewer", [Online]. Available: www.vosviewer.com.
[16] T. Jiang, X. Liu, C. Zhang, C. Yin, and H. Liu, "Overview of Trends in Global Single Cell Research Based on Bibliometric Analysis and LDA Model (2009-2019)," *J. Data Inf. Sci.*, vol. 6, no. 2, pp. 163–178, 2021, doi: 10.2478/JDIS-2021-0008.
[17] X. Han, "Evolution of research topics in LIS between 1996 and 2019: an analysis based on latent Dirichlet allocation topic model," *Scientometrics*, vol. 125, no. 3, pp. 2561–2595, 2020, doi: 10.1007/S11192-020-03721-0/FIGURES/2.
[18] Z. Hu, S. Fang, and T. Liang, "Empirical study of constructing a knowledge organization system of patent documents using topic modeling," *Scientometrics*, vol. 100, no. 3, pp. 787–799, 2014, doi: 10.1007/s11192-014-1328-1.
[19] A. R. Destarani, I. Slamet, and S. Subanti, "Trend Topic Analysis using Latent Dirichlet Allocation (LDA) (Study Case: Denpasar People's Complaints Online Website)," *J. Ilm. Tek. Elektro Komput. dan Inform.*, vol. 5, no. 1, pp. 50–58, 2019, doi: 10.26555/JITEKI.V5I1.13088.
[20] M.-H. Hwang, S. Ha, M. In, and K. Lee, "A Method of Trend Analysis using Latent Dirichlet Allocation," *Int. J. Control Autom.*, vol. 11, no. 5, pp. 173–182, 2018, doi: 10.14257/ijca.2018.11.5.15.
[21] M. Lamba and M. Madhusudhan, "Mapping of topics in DESIDOC Journal of Library and Information Technology, India," *Scientometrics*, vol. 120, no. 2, pp. 477–505, 2019, doi: 10.1007/S11192-019-03137-5.
[22] G. Shobha and S. Rangaswamy, "Computational analysis and understanding of natural languages: Principles, methods and applications Chapter 8 - Machine learning," *Handb. Stat.*, vol. 38, pp. 197–228, 2018.
[23] "Projects senderle/topic-modeling-tool GitHub." https://github.com/senderle/topic-modeling-tool/projects (accessed Mar. 12, 2022).
[24] "Quickstart Guide | Topic Modeling Tool Blog." https://senderle.github.io/topic-modeling-tool/documentation/2017/01/06/quickstart.html (accessed Mar. 12, 2022).
[25] "senderle (Jonathan Scott Enderle) · GitHub." https://github.com/senderle (accessed Mar. 12, 2022).
[26] N. van Eck and L. Waltman, "VOSviewer - Visualizing scientific landscapes," *VOSviewer*. Accessed: Feb. 12, 2022. [Online]. Available: https://www.vosviewer.com//
[27] N. J. van Eck and L. Waltman, "Software survey: VOSviewer, a computer program for bibliometric mapping," *Scientometrics*, vol. 84, no. 2, pp. 523–538, Dec. 2010, doi: 10.1007/S11192-009-0146-3/FIGURES/7.
[28] E. Coşkun, G. Özdağoğlu, M. Damar, and B. A. Çallı, "Scientometrics-based study of computer science and information systems research community macro level profiles," *Proc. 12th IADIS Int. Conf. Inf. Syst. 2019, IS 2019*, pp. 180–188, 2019, doi: 10.33965/IS2019_201905L023.
[29] T. A. Morris and K. W. McCain, "The structure of medical informatics journal literature," *J. Am. Med. Informatics Assoc.*, vol. 5, no. 5, pp. 448–466, Sep. 1998, doi: 10.1136/JAMIA.1998.0050448/2/JAMIA0050448.F08.JPEG.
[30] N. Mays and C. Pope, "Qualitative research in health care: Assessing quality in qualitative research," *BMJ Br. Med. J.*, vol. 320, no. 7226, p. 50, Jan. 2000, doi: 10.1136/BMJ.320.7226.50.
[31] A. F. Hannawa, L. García-Jiménez, C. Candrian, C. Rossmann, and P. J. Schulz, "Identifying the field of health communication," *J. Health Commun.*, vol. 20, no. 5, pp. 521–530, May 2015, doi: 10.1080/10810730.2014.999891.
[32] Y. Kwon and J. Kwon, "An overview of Telecommunications Policy's 40-year research history: Text and bibliographic analyses," *Telecomm. Policy*, vol. 41, no. 10, pp. 878–890, Nov. 2017, doi: 10.1016/J.TELPOL.2017.07.012.
[33] F. Bannister and Å. Grönlund, "Information Technology and Government Research: A Brief History," *Hawaii Int. Conf. Syst. Sci. 2017*, Jan. 2017, Accessed: Apr. 06, 2022. [Online]. Available: https://aisel.aisnet.org/hicss-50/eg/transformational_government/4
[34] D. Ivanović and Y. S. Ho, "Highly cited articles in the Information Science and Library Science category in Social Science Citation Index: A bibliometric analysis," *J. Librariansh. Inf. Sci.*, vol. 48, no. 1, pp. 36–46, Mar. 2016, doi: 10.1177/0961000614537514.

BIOGRAPHIES OF AUTHORS

Tipawan Silwattananusarn is an Assistant Professor in the Information Management Program, Faculty of Humanities and Social Sciences, Prince of Songkla University, Thailand. She has received her Ph.D. degree in Information Studies from Khon Kaen University. She is involved in research and teaching in Data Mining and Machine Learning Tools, Data Analytics and Visualization, Information Management Systems and Technologies, Information Security Management, Digital Information Management. She is a specialist in data mining and analytics, data mining applications for knowledge management, and IT applications in academic libraries. She is an author and published research and textbooks in Data Mining and Data Analytics & Visualization. Her latest book, "Practical Data Mining with WEKA" (written in Thai), was published in June 2021. She can be contacted at email: tipawan.s@psu.ac.th.

Pachisa Kulkanjanapiban is an Academic Librarian with a Professional Level in the Khunying Long Athakravisunthorn Learning Resources Center, Prince of Songkla University, Thailand. She holds an M.Sc. degree in Management Information Technology (MIT) from the Prince of Songkla University. She is a specialist in scientometrics, bibliometrics, database and information retrieval, and predictive modeling. Her expertise is Institutional Repository (IR), especially running the "PSU Knowledge Bank." Also, she is responsible for IT project and IT service management, knowledge and data management in her organization. She is fascinated by data science, data mining and machine learning, digital management systems, data analytics and visualization, business intelligence, digital transformation, database analysis and design, computing in social sciences, arts and humanities, and data management topics. She can be contacted at email: pachisa.ku@psu.ac.th.