Analysis of Research Data in Information Science Using the Topic Modeling Method[1]

Sompejch Junlabuddee[a] and Kulthida Tuamsuk[a*]

[a]*Faculty of Humanities and Social Sciences, Khon Kaen University*
Khon Kaen 40002, Thailand
[*]*Corresponding Author. Email: kultua@kku.ac.th*

Received: January 6, 2021
Revised: March 9, 2021
Accepted: April 19, 2021

Abstract

Most research data in the modern world are in digital format, and there is therefore a need to develop high-efficiency tools that can provide access to and an understanding of these data. Computerization technology based on natural language processing with a capacity for topic extraction and categorization would enable us to identify new topics and future directions for research in several fields of study. The aim of this research was to analyze and categorize information science research data obtained from journals listed in an international database between 2013 and 2019. The research methodology applied here was data analysis based on the topic modeling method, a technique used to locate word groups or topics from a corpus containing complicated and difficult works. This method yields reliable and high-accuracy outcomes. The data analyzed here were drawn from research articles published in information science journals, the names of which were listed in the Scimago Journal and Country Rank between 2013 and 2019. Only journals in the Web of Science and articles written in English were included. A total of 30,571 research articles obtained from 677 volumes of 99 journals were analyzed using the topic modeling method, and topics were assigned by experts in the field. The findings revealed that over the past seven years, research was carried out on 30 topics in information science. The five most frequently researched topics were competency development, data management, social media analytics, public and community services, and bioinformatics. A comparison with other research data analyzed in the field of information science over the past five years using other techniques showed clear differences and a tendency of the research

[1] This paper is a partial fulfilment of the requirements for the degree of Doctor of Philosohy in Information Studies, Graduate School, Khon Kaen University, Thailand.

topics to change. The results of this research can greatly benefit the identification of research directions for the future.

Keywords: research data, information science, data analytics, topic modeling

Introduction

Research in the area of information is relevant to many disciplines, for example information science, information studies, library science, information technology, information management, mass media management, business management, and education technology. Research has also been conducted on information in science-related fields such as engineering and the health sciences, although the research issues differ based on the aspects of information that are of interest in each field. Nevertheless, information has been seen in fields that emphasize information as one of the core elements under investigation. These elements are connected to the modern technologies used in information management, and are advantageous in many areas in which users or people are involved. This field of study is commonly referred to as Information Science. When the iSchools Organization was founded in 2005, as an international organization in which the members were institutions offering programs in information science or relevant fields around the world, the necessity of performing research in these fields and the importance of the role of information in institutions, communities, and peoples' ways of living became clearer (Claver, González, and Llopis, 2000; Larsen, 2008; iSchools Inc., 2015).

Information science is classified as a multidisciplinary field that involves the analysis, compilation, categorization, modification, storage, retrieval, and dissemination of information (Stock and Stock, 2013). In 1968, under the aegis of the American Documentation Institute (the name of which was later changed to the American Society for Information Science), Borko (1968) defined information science as the science developed from library science in which the properties and behaviors of information, use and transmission of information, and information processing are studied, so that access to and use of

information can provide optimal benefits. However, information science is a constantly changing discipline, due to the impacts of the fast-growing information and communication technologies, and studies have therefore been conducted to analyze the scope of information science using various approaches, in order to obtain guidelines that can assist in determining the curricula of courses, developing the competency of information science professionals, and setting research topics that are correlated to the solution of problems and development of various aspects that rely on information as the key decision-making tool.

A review of prior research work reveals a large number of information-related research analyses. However, the most frequently cited study in this field is by Maceviciute and Wilson (2002), who conducted an analysis of 150 research articles on information management that had been published in the six top international journals. Thirteen information research topics were identified: artificial intelligence, economics of information, education for information management, information management, information networks, information professionals, information systems, information technology, information use and users, knowledge management, organization, telecommunication industry, and theory and research methods. Later, Hawkins, Larson, and Caton (2003) performed analyses and categorized 3,004 records published in *Information Science Abstracts* between 1998 and 1999, and developed a taxonomy of information science in which information content was classified into 11 categories, as follows: information science research, knowledge organization, information profession, social issues, the information industry, publishing and dissemination, information technology, information systems and services, electronic information systems and services, subject-specific sources and applications, libraries and library services, and government and legal information and issues. There have also been research studies over the past five years, such as those by Luo and McKinney (2015), Togia and Malliari (2017), and Liu and Yang (2019), which offered analyses of research information in journal articles whose databases are widely accepted in information science. However, these studies were

conducted using manual methods based on content analysis, coding, clustering, and classification; there has been no research work to date that has applied an analytical technique to big data, and particularly the use of the topic modeling method, which computerizes the results of machine learning and analyzes information research data.

The aim of the present research was to analyze and categorize research data in the field of information science that appeared in journals listed in the international database between the years 2013 and 2019. This was a big dataset and therefore required analysis with an appropriate high-accuracy technique. The topic modeling method was selected, as this technique is generally used to find groups of words or topics in a corpus of difficult and complicated works (Meriam, 2012; Xie and Xing, 2013). This approach can give reliable and valid results (Blei et al., 2003), and is popular for use in analyses of research data in fields such as bioinformatics and health informatics, and for data drawn from various social media (Liu et al., 2016; Alabawi, Yeap, and Benyoucef, 2020). The data clusters found in this way can then be classified and used in education and research on different topics. We therefore expected that the use of the topic modeling method to analyze research data in the field of information science would enable us to identify new research topics and future research directions as well as to gain benefits in terms of instruction and research in a wide circle of information science institutions.

Topic Modeling Method

Most of the data in the modern world are in digital format, and there is therefore a need to develop high-efficiency tools that can provide accessibility to and understanding of these data. As a result, computerization technology based on natural language processing has been developed that involves pre-feeding sets of algorithms and the necessary data into the system as well as analytical rules in the correct order, to allow for the memorization of various patterns in the corpus or the use of machine learning algorithms (Albalawi, Yeap, and Benyoucef, 2020). Some programs have the capacity, through the use of natural language processing, to categorize data clusters and extract topics and relationships from documents. Topic modeling is a technique that allows for access to the data and the extraction of keywords and topics in a document. It enables the retrieval of specific thematic structures hidden in a large number of documents (Gerrish and Blei, 2011; Hussey et al., 2012; Farzindar and Inkpen, 2015).

Topic modeling (TM) is an approach that can be used to find groups of words in a corpus that has been constructed as a tool for text mining, corresponding to difficult and complicated works. This involves searching for topics in a corpus containing a large amount of data. It can also be used to extract topics from both short and long documents, such as articles or digital books (Meriam, 2012; Xie and Xing, 2013), and yields reliable and valid results when used with methods of analysis such as probabilistic latent analysis (PLSA), latent semantic analysis (LSA) and latent Dirichlet allocation (LDA) (Cheng et al., 2014). LDA was invented by Blei et al. (2003), and is a popular TM algorithm for extracting a set of topics from a large documentary corpus owing to its accuracy and correctness.

LDA-based topic modeling may focus on a topic or a theme hidden in the document that is to be extracted. The specificity of these topics is that they are latent or hidden in the document (in some texts, the topics are conveniently referred to as latent variable or hidden variable), unseen by both humans and computers. This means that a TM analysis can only tell us whether the words belong to the first, second, or third latent topic, and the proportion of the latent topic that is contained in each document; it is left to humans to determine the content of each latent topic, in order to obtain the topic that correlates most closely with the content (Blei et al., 2003). For example, the first latent topic may deal with politics, the second with sports, and the third with social factors.

TM is a probabilistic model that can be used to explain data related to a topic in any statement, and involves the following key components: (i) topics, i.e. the fundamental concepts or themes that

represent a statement (for instance, in a corpus of newspaper articles, there may be fundamental topics related to finance, climates, politics, sports, news, etc.); (ii) a probability distribution, which explains an article by distributing the probability in order to identify the similarity. Topic modeling can be used to draw conclusions, enable retrieval and order files based on topics rather than words (Steyvers and Griffiths, 2007). There are three steps in topic modeling: (i) finding the hidden topic in each document; (ii) labeling the topic to show what topic it is related to; and (iii) grouping the words found in the document under each topic. For the chance and probability that occurs, topic modeling can indicate how many topics there are and which words there are in each topic, as in the example illustrated in Figure 1. Blei (2012) reported the following results from the use of LDA to process a document entitled Seeking Life's Bare (Genetic) Necessities: (i) the proportions and assignments of the topics (as shown in the bar graph on the right) represent the probability of the topics in each document based on a discontinuous probability distribution; for example, this document consists of Topics 1 to 4 with probabilities of 0.42, 0.38, 0.50 and 0.35, respectively; (ii) the figure on the left shows that this document consists of four topics, the keywords are related to each topic and what their probabilities are (the number of topics is a hyper-parameter that requires us to set how many topics there are); (iii) when the output is obtained, an interpretation may be necessary in order to understand what the key words in each topic are related to. For example, Topic 1 (top box) contains the words gene, DNA, and genetics, and is therefore related to genetics, while Topic 4 (bottom box) contains the words data, number, and computer, and is therefore related to computer science. This process therefore requires an expert opinion to ensure the accuracy of the results.

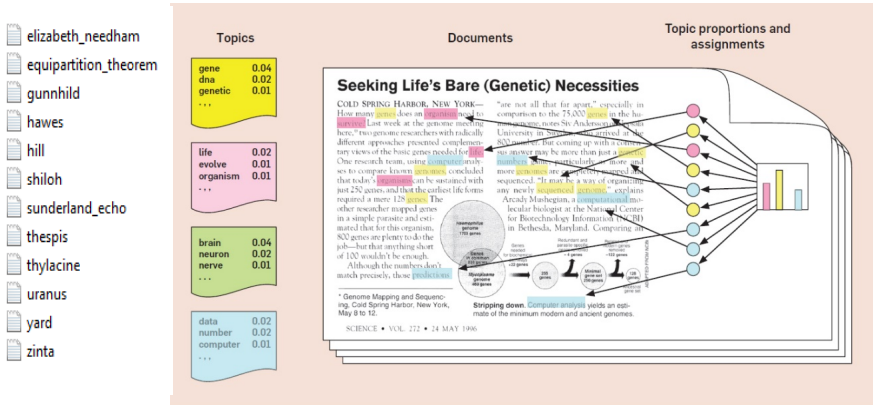

Figure 1 Example of results from topic modeling using the LDA method on a document entitled Seeking Life's Bare (Genetic) Necessities (Blei, 2012)

LDA is a technique for creating a simulated topic. It is widely used for extracting a topic from statements or words from a high-probability topic, or words that frequently appear together, usually within the first 10 to 15 words of a statement. LDA is used to interpret and label the meaning of a topic. However, LDA presents the results of many topics in the form of a set. A low K value means that there are few topics, or that the text is vast, while a high K value indicates that the topics cannot be interpreted and should be grouped together. Determining an appropriate K value is therefore an important step in a topic simulation algorithm based on LDA (Syed and Spruit, 2017). The efficiency of topic modelling can be measured by several methods, and in the present work, this was done based on the values of the coherence and perplexity. A high value of coherence and a low value of perplexity means that the topic model under analysis has high efficiency (Waal and Barnard, 2008; Röder, Both, and Hinneburg, 2015).

Research Methodology

This research involved data analytics using the topic modeling method. A model was created of research data in the field of information science that appeared in journals listed in an international database between 2013 and 2019, based on the following steps:

1. Selection of research articles for analysis: The names of the information science journals were retrieved from the Scimago Journal and Country Rank [https://www.scimagojr.com/ journalrank.php] from between 2013 and 2019. The selection was limited to journals listed in the Web of Science and published in English. Ninety-nine journals and 677 volumes were obtained, from which 30,571 articles were selected (Table 1).

Table 1 Research articles published in the selected journals between 2013 and 2019

Year	No. of articles
2019	5,327
2018	4,409
2017	4,229
2016	3,293
2015	4,160
2014	4,174
2013	3,743
Total	**30,571**

2. Collection and preparation of the data: Bibliographic data were collected from the articles in comma-separated value format. The data to be used in topic modeling were then chosen, including the names of the authors, the title of the article, the name of the journal, the year of publication, and the abstract.

3. Data grouping using the topic modeling method: The 30,571 articles were categorized into clusters based on probabilistic values.

Thirty topics were obtained, with the most frequent topics being Topic 13, with 4,467 records; Topic 1, with 2,937 records; and Topic 16, with 2,915 records. Only nine records were found for Topic 3 in this cluster (Figure 2).

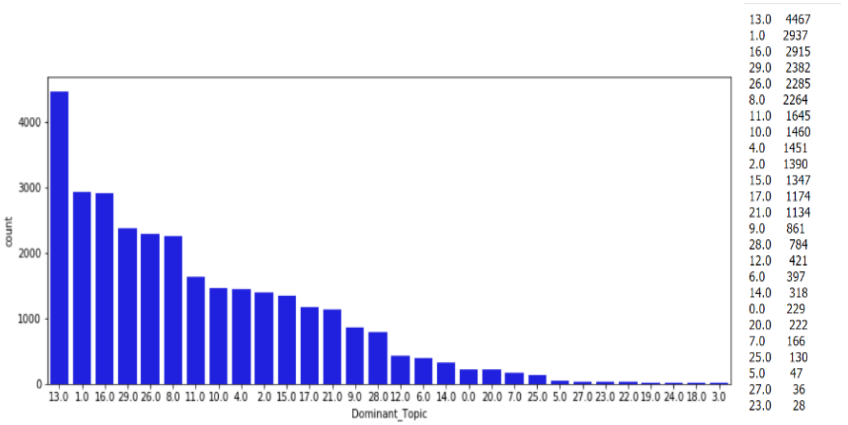

Figure 2 Numbers of research articles clustered by topic, with a total of 30 topics

4. Clustering of keywords for each topic: This was carried out by drawing the first 15 keywords with the highest weight values in each topic, since by using LDA technique a high-probability topic where the words frequently appear together usually fall within the first 10 to 15 words of a statement (Syed and Spruit, 2017). It was found that Topic 13 contained the following keywords: student, support, project, program, practice, work, learn, learning, skill, develop, experience, professional, design, provide, and management. Topic 1 contained the following keywords: technology, change, digital, health, area, place, local, partnership, global, historical, environment, sustainability, national, future, and environmental. Topic 16 contained the words: online, social medium, survey, participant, content, type, group, seek, question, experience, attitude, website, include, respondent, and perception (Table 2).

Table 2 Top 15 keywords found in each of the 30 topics in information science research.

Topic	Keywords in each topic
0	0.068*"technology"+0.052*"change"+0.047*"digital"+0.035*"health"+0.015*"area"+0.012*"place"+0.011*"local"+0.010*"partnership"+0.010*"global"+0.009*"historical"+0.009*"environment"+0.009*"sustainability"+0.009*"national" +0.008*"future"+0.008*"environmental"
1	0.028*"algorithm"+0.027*"base"+0.022*"feature"+0.020*"performance"+0.014*"technique"+0.012*"compare"+0.011*"retrieval"+0.011*"term"+0.011*"experimental"+0.011*"classification" +0.010*"task"+0.010*"apply"+0.009*"set"+0.009*"evaluate"+ 0.009*"accuracy"
2	0.027*"problem"+0.019*"code"+0.019*"function"+0.015*"case"+0.011*"sequence"+0.011*"set"+0.010*"time"+0.009*"space"+0.009*"museum"+0.009*"base"+0.009*"operation"+0.008*"property"+0.008*"constraint"+0.008*"provide"+0.008*"construct"
3	0.044*"classification" + 0.029*"assignment" + 0.022*"art" + 0.021*"alignment" + 0.021* "undergraduate_ student" + 0.019*"audience" + 0.017*"orientation" + 0.014*"transfer" +0.013*"engineering" + 0.013*"shape" + 0.013*"diffusion" + 0.012*"pose" + 0.012*"italian" + 0.012*"preserve" + 0.012*"knowledge_transfer"
4	0.041*"article"+0.023*"analysis"+0.020*"literature"+0.019*"identify"+0.018*"topic"+0.018*"review"+0.017*"field"+0.015*"term"+0.015*"language"+0.015*"work"+0.014*"subject"+0.013*"include"+ 0.011*"title"+0.011*"text"+0.010*"focus"
5	0.046*"event"+0.035*"risk"+0.030*"llc"+0.014*"marketing"+0.014*"budget"+0.013*"location"+0.013*"agent"+0.011*"crisis"+0.010*"increase"+0.010*"treatment"+0.009*"brand"+0.009*"time"+0.008*"year"+0.008*"person"+ 0.007*"reduce"
6	0.116*"search"+0.096*"network"+0.041*"query"+0.024*"web"+0.022*"pattern"+0.021*"structure"+0.018*"node"+0.017*"link"+0.013*"graph"+0.012*"term"+0.010*"keywords"+0.009*"database"+0.008*"time"+0.008*"user"+0.008*"identify"
7	0.093*"university"+0.059*"document"+0.028*"educational"+0.022*"libraries"+0.021*"institution"+0.017*"academic"+0.017*"open_access"+0.015*"lis"+0.014*"india"+0.014*"international"+0.012*"woman"+0.012*"gender"+0.012*"instruction"+ 0.009*"college"+0.009*"undergraduate"
8	0.054*"journal"+0.047*"publication"+0.038*"citation"+0.027*"article"+0.021*"publish"+0.020*"country"+0.019*"impact"+ 0.018*"researcher"+ 0.015*"academic"+0.015*"year"+0.013*"number"+ 0.013*"field"+0.011*"high"+0.011*"discipline"+ 0.011*"indicator"
9	0.038*"evaluation"+0.035*"quality"+0.029*"measure"+0.020*"level"+0.017*"evaluate"+0.015*"item"+0.015*"test"+ 0.014*"group"+ 0.014*"compare"+0.013*"score"+0.013*"assess"+0.013*"assessment"+0.009*"criterion"+0.009*"difference"+0.009*"high"
10	0.082*"user"+0.020*"factor"+0.018*"internet"+0.016*"model"+0.013*"personal"+0.012*"service"+0.011*"recommendation"+0.011*"trust"+0.010*"behaviour"+0.010*"behavior"+0.010*"security"+0.009*"mobile"+0.009*"individual"+0.008*"consumer"+ 0.008*"adoption"
11	0.022*"number"+0.016*"size"+0.015*"distribution"+0.015*"estimate"+0.013*"term"+0.013*"sample"+0.012*"class"+0.012*"measure"+0.012*"rate"+0.010*"matrix"+0.010*"variable"+0.010*"large"+0.009*"small"+0.009*"function"+0.009*"analysis"
12	0.251*"data"+0.061*"record"+0.026*"metadata"+0.020*"repository"+0.016*"preservation"+0.015*"access"+0.015*"management"+0.014*"database"+0.012*"share"+0.011*"researcher"+0.010*"archive"+0.009*"provide"+0.009*"africa"+0.008*"include"+ 0.008*"request"
13	0.049*"student"+0.016*"support"+0.014*"project"+0.014*"program"+0.012*"practice"+0.012*"work"+0.012*"learn"+0.012*"learning"+0.011*"skill"+0.011*"develop"+0.011*"experience"+0.010*"professional"+0.010*"design"+0.010*"provide"+ 0.010*"management"
14	0.276*"model"+0.016*"develop"+0.015*"problem"+0.014*"base"+0.011*"solution"+0.010*"representation"+0.010*"test"+0.009*"solve"+0.009*"modeling"+0.008*"fit"+0.008*"complexity"+0.008*"apply"+0.007*"structure"+0.007*"hybrid"+0.007*"framework"

Table 2 Top 15 keywords found in each of the 30 topics in information science research. (cont.)

Topic	Keywords in each topic
15	0.056*"knowledge"+0.031*"music"+0.021*"relationship"+0.018*"organization"+0.018*"company"+0.017*"innovation"+0.017*"business"+0.014*"performance"+.013*"knowledge_management"+0.013*"management"+0.012*"factor"+0.011*"firm"+0.010*"product"+0.010*"employee"+0.010*"industry"
16	0.020*"online"+0.017*"social_medium"+0.014*"survey"+0.014*"participant"+0.012*"content"+0.011*"type"+0.011*"group"+0.010*"seek"+0.010*"question"+0.008*"experience"+0.008*"attitude"+0.008*"website"+0.008*"include"+0.007*"respondent"+ 0.007*"perception"
17	0.030*"tool"+0.023*"design"+0.020*"application"+0.018*"develop"+0.016*"software"+0.015*"provide"+0.014*"framework"+0.011*"ontology"+0.011*"platform"+0.011*"concept"+0.011*"archival"+0.011*"base"+0.010*"standard"+0.009*"support"+ 0.009*"knowledge"
18	0.152*"cluster"+0.038*"clustering"+0.029*"privacy"+0.023*"metaphor"+0.020*"analysis"+0.019*"protection"+0.018*"fuzzy"+0.017*"regulatory"+0.017*"expression"+0.016*"roll"+0.013*"mutual"+0.012*"food"+0.011*"fusion"+0.011*"thematic"+0.011*"broadcast"
19	0.133*"theory"+0.065*"child"+0.056*"family"+0.049*"identity"+0.042*"construction"+0.030*"mobile_phone"+0.022*"memory"+0.019*"machine"+0.017*"checklist"+0.017*"convergence"+0.016*"hierarchy"+0.016*"film"+0.015*"appraisal"+ 0.015*"structural_equation_model"+ 0.013*"operational_society"
20	0.064*"communication"+0.064*"scheme"+0.052*"law"+0.044*"channel"+0.039*"legal"+0.031*"message"+0.025*"tag"+0.023*"feedback"+0.021*"state"+0.020*"protocol"+0.014*"capacity"+0.012*"transmission"+0.011*"service_quality"+0.011*"strategy"+ 0.011*"secure"
21	0.222*"library"+0.070*"service"+0.047*"collection"+0.042*"resource"+0.023*"provide"+0.021*"access"+0.018*"user"+0.017*"material"+0.010*"space"+0.009*"include"+0.009*"usage"+0.009*"article"+0.009*"archive"+0.007*"website"+0.007*"offer"
22	0.063*"school"+0.028*"netherlands"+0.026*"reference"+0.018*"article_discuss"+0.017*"workshop"+0.015*"educator"+0.014*"hospital"+0.013*"canada"+0.012*"intellectual"+0.012*"module"+0.012*"article_explore"+0.012*"discussion"+0.011*"computer"+0.011*"character"
23	0.180*"source"+0.133*"book"+0.062*"patent"+0.030*"reader"+0.029*"edition"+0.027*"south_africa"+0.012*"newspaper"+0.011*"press"+0.010*"technological"+0.009*"cultural"+0.009*"century"+0.008*"commercial"+0.007*"market"+0.007*"digital_age"+0.007*"volume"
24	0.044*"patient"+0.036*"nigeria"+0.036*"manuscript"+0.026*"catalog"+0.024*"answer"+0.020*"movement"+0.018*"portal"+0.018*"establishment"+0.017*"day"+0.014*"devote"+0.013*"division"+0.012*"campaign"+0.012*"grade"+0.012*"user_satisfaction"+ 0.011*"job"
25	0.051*"image"+0.025*"digital_library"+0.021*"video"+0.018*"visual"+0.016*"strategy"+0.014*"time"+0.013*"campus"+0.011*"cost"+0.008*"control"+0.008*"attack"+0.008*"acquisition"+0.007*"copy"+0.007*"presentation"+0.006*"dynamic"+0.006*"mode"
26	0.019*"interaction"+0.013*"identify"+0.013*"structure"+0.013*"analysis"+0.012*"data"+0.010*"gene"+0.010*"protein"+0.010*"compound"+0.009*"include"+0.008*"biological"+0.008*"potential"+0.008*"simulation"+0.007*"complex"+0.007*"predict"+ 0.007*"human"
27	0.096*"map"+0.060*"region"+0.020*"series"+0.019*"youth"+0.018*"printing"+0.017*"display"+0.016*"ensemble"+0.016*"african"+0.015*"digital_collection"+0.013*"mapping"+0.012*"sampling"+0.012*"obstacle"+0.012*"land"+0.011*"taiwan"+0.010*"ligand"
28	0.028*"collaboration"+0.023*"concept"+0.020*"science"+0.019*"policy"+0.016*"scientific"+ 0.016*"analysis"+0.016*"work"+0.016*"knowledge"+0.015*"researcher"+0.015*"teach"+0.015*"activity"+0.012*"collaborative"+0.011*"development"+ 0.011*"framework"+0.010*"field"
29	0.019*"community"+0.019*"practice"+0.015*"role"+0.013*"public_library"+0.013*"development"+0.012*"article"+0.012*"archive"+0.011*"social"+0.011*"government"+0.011*"public"+0.010*"context"+0.009*"work"+0.008*"country"+ 0.007*"society"+ 0.007*"focus"

5. Clustering the research articles by labeling: This was performed to determine the topic to which each article belonged to. It was achieved by creating a function called the dominant topic and calculating previously processed keywords that appeared in the abstract with the weights of the keywords appearing in the topic clusters that were obtained from LDA modeling (Figure 3).

Dominant_Topic	Perc_Contribution	Topic_Keywords	Abstract
10	0.324600011	user, factor, internet, model, personal	Prior works offer compelling evidence that, on the demand side of t
13	0.2667	student, support, project, program, pr	Capacity is the maximum short-run output with capital in place und
13	0.508000016	student, support, project, program, pr	This study examines the team-level effects of pair programming by
16	0.193100005	online, social_medium, survey, partic	Online reviews offer consumers the indirect experience of products
0	0.187600002	technology, change, digital, health, ar	Given the cost of electronic health records (EHRs) to society and he
10	0.581399977	user, factor, internet, model, personal	The rise of online shopping cart-tracking technologies enables new
15	0.372000009	knowledge, music, relationship, organ	Protecting organizational information is a top priority for most firms.
10	0.165099993	user, factor, internet, model, personal	Live chat tools have emerged as a channel for fostering synchronous
10	0.347600013	user, factor, internet, model, personal	Abundant empirical evidence supports the overall efficacy of social
15	0.486499995	knowledge, music, relationship, organ	Many e-commerce platforms offer editor-curated recommendations
16	0.290600002	online, social_medium, survey, partic	Digital infrastructures are a result of individual yet interdependent sy
10	0.310699999	user, factor, internet, model, personal	WeChat, an instant messaging app, is considered a mega app becaus
25	0.417299986	image, digital_library, video, visual, st	From an upset stomach to a life-threatening foodborne illness, gettin
29	0.516900003	community, practice, role, public_libr	In this research commentary, we argue that the current digital era c
15	0.274699986	knowledge, music, relationship, organ	Although there are both process-related and human-related grounds
25	0.240199998	image, digital_library, video, visual, st	A cloud service agreement entails the provisioning of a required set
15	0.730300009	knowledge, music, relationship, organ	Strategic alliances have become popular organizational forms in the
10	0.474700004	user, factor, internet, model, personal	The open internet is plagued by congestion that restricts the develop
13	0.336600006	student, support, project, program, pr	Collaboration through open superposition describes the dominant w

Figure 3 Examples of dominant topics, identified based on groups of words in the abstract of each document

6. Measuring the efficiency of the topic model: This was achieved by using the coherence and perplexity values obtained from the research analytics in the field of information science. It was found that the models of 30 topics had a coherence value of 0.4109 and a perplexity value of -24.0031, which supported the efficiency, appropriateness, and validity of topic clustering. The models could therefore be used to assign the topics in information science.

7. Assignment of research topics by experts: As mentioned earlier, TM involves the assignment of probable topics depending on groups of keywords that appear in the document. However, the assignment of research topics based on a review of the words that appear depends on human experience and expertise. In this research, a snowball technique was applied, and assignment was carried out by five experts

in the field of information science, from both Thailand and abroad, who had published at least 10 articles over five years in the Scopus or Web of Science databases. The research topics were first assigned by one expert, followed by the next, who added more topics and so on, until the fifth expert completed the task.

Findings and Discussion

Research data in the field of information science that appeared in an international database between 2013 and 2019 were analyzed using the topic modeling method, and the topics were assigned by five experts within the field of information science. These research topics were categorized into 30 groups, and the five most frequent were: (i) competency development; (ii) data management; (iii) social media analytics; (iv) public and community services; and (v) bioinformatics (as shown in Table 3).

Table 3 Research topics in information science, as assigned by experts based on the results of topic modeling

Topic	IS research topic	Description
13	Competency development	Research into the development of competency in human resources in the information professions in different forms, and especially the competency development of students in the field of information science through, for example, training, practicum, experience training, learning management, project implementation, etc.
1	Data management	Studies of data management, techniques, methods, algorithms, experiments, categorization, evaluation, and applications for data
16	Social media analytics	The use of online social media analytics to explain the patterns, content, users, actors, attitudes, perceptions and impacts of online social media
29	Public and community services	Studies of information services for people and communities in different forms, with an emphasis on community participation and a lessening of the digital divide
26	Bioinformatics	Analysis and categorization of bioinformatics data for the benefit of studies in human anatomy and the development of medicine and public health

Table 3 Research topics in information science, as assigned by experts based on the results of topic modeling (cont.)

Topic	IS research topic	Description
8	Scholarly communication	Analysis of the production of academic works, including the impact of works in the field of information science
11	Knowledge organization	Knowledge management, analysis of knowledge through the organization of structures, categories, relations, and other details in the domain of knowledge for the benefits of retrieval and access to knowledge that lacks systematic management
10	User behavior	Studies of user behavior in different dimensions, including factors affecting the behavioral changes of users, to improve service management or organizational operations
4	Data analytics	Analysis of a large amounts of data from a big data source, and especially data on articles and literature, in order to distinguish content, language use, words, and emphasized points
2	Information system	Development of information systems, computerization of information systems, and program writing
15	Knowledge management	Knowledge management of specific content, and knowledge management and innovation for business and industrial organizations
17	Software development	Development of software, including design, tools, framework, platform coding, and various standards
21	Information services	Studies of information services related to collection management, accessibility, website development, area management, and the design of various services that can increase the efficiency of information services
9	Information quality	Analysis of information quality, including quality criteria, indicators, assessment and evaluation, testing, comparison, rating, etc.
28	Knowledge networks	Creation of knowledge networks, including the development of concepts, policies, cooperation, activities, etc.
12	Digital collection management	Digital collection management, including the management of records, creating metadata, storage, exchange, and dissemination
6	Information retrieval	Information retrieval, producing queries, setting words/terms for retrieval, retrieval techniques, connection of retrieval sources, retrieval efficiency
14	Data modeling	Data modeling, analysis and management of data structures, relationships between data and the presentation of data worth utilizing in various forms

Table 3 Research topics in information science, as assigned by experts based on the results of topic modeling (cont.)

Topic	IS research topic	Description
0	Information development	Studies of information for development in social dimensions that are subject to impacts from technological changes, such as impacts on the environment, health, and people's ways of living
20	Information governance	Studies of information governance, with an emphasis on correctness, safety and observing laws when managing different types of data that have impacts on organizations
7	Academic library management	Studies of academic library administration from various perspectives related to instructions, serving foreign students, collection management, and new challenges in library administration
25	Digital libraries	Studies of digital libraries, management of different forms of digital resources, management of provision, accessibility, control and copying (patents, rights)
5	Information economy	Studies of the information economy, such as analyses of budgets, risks, crises, marketing, and branding
27	Knowledge mapping	Development of knowledge maps involving interesting content at the group, organization, area, and community levels, in order to demonstrate the connection of knowledge of a topic
23	Digital resources	Studies of digital or knowledge resources that may be in various forms and may be connected with reading, access, and dissemination
22	Information education	Studies of information education, references, discussions, meetings, seminars, and intellectual property
19	Information literacy	Studies of information literacy, explanations of information literacy, analysis of the behaviors of different groups of people to evaluate or develop information literacy
24	Information marketing	Information marketing, information management where importance is placed on customers, campaigning, making catalogs, promotion, and the assessment of customer satisfaction
18	Information security	Studies of information security and privacy, safety measures, avoidance and control of fuzzy data
3	Knowledge transfer	Studies of knowledge transfer in the context of instruction, orientation, assignment, and arranging activities

A comparison of the results of our research with those of other studies carried out over the past five years by Luo and McKinney (2015), Togia and Malliari (2017), and Liu and Yang (2019) (as summarized

in Table 4) showed that there was a clear difference in the research topics found in the first five clusters; only the topics related to social media were found to be similar to the results of the study by Liu and Yang (2019). With regard to the first 10 topics, two more frequent topics in the journals were found to correlate to this research, i.e., scholarly communication and user/information behaviour (Luo and McKinney, 2015; Togia and Malliari, 2017). It should also be noted that research topics related to information literacy were found in the first six results in studies by Luo and McKinney (2015), Togia and Malliari (2017), and Liu and Yang (2019), while in this research, information literacy was found to be in 27th place. Although these previous research works were carried out over almost the same period of time, i.e. between 2015 and 2019, the sources for the content analysis were different. However, a comparison of their findings with those of this research at least confirms the most popular research topics and trends in information science during this period.

An interesting issue that arose in this research was the categorization of keywords identified from research articles using the topic modeling technique. This showed that the following research topics have a tendency to change: (1) those that involve the use of technology to analyze data and information in order to understand the content, including social media analytics, data analytics, bioinformatics, and data modeling; (2) research related to digital data management, including digital collection management, digital libraries, and digital resources; and (3) research related to the quality, security and safety of data, including information quality, information governance and information security. Research issues associated with information services, user behavior, information retrieval, information literacy, information economy, information development and information marketing were consistently popular in information science, while the number of research studies on organization management and library management decreased. Only articles on topics involving the management of university libraries were still published and disseminated in high numbers.

Table 4 Comparison of research topics in information science found in this research and previous studies

Items	This research	Luo and McKinney (2015)	Togia and Malliari (2017)	Liu and Yang (2019)
Sources	Research articles published in ISI journals, ranked in SJR (99 journals selected)	Research articles published in JAL (Journal of Academic Librarianship)	Research articles published in the top five journals with the highest impact factor classified in Ulrich's Serials Directory	Research articles published in Web of Science core journals (41 journals selected)
Coverage	2013-2019	2004-2013	2011-2016	2008-2017
Methods	Topic modeling	Content analysis	Coding/classification	Keyword grouping and clustering
Findings	30 topics	24 topics	18 topics	20 topics
Top 1-5	1. Competency development 2. Data management 3. Social media analytics 4. Public and community services 5. Bioinformatics	1. Information literacy 2. User information behaviour 3. Library personnel 4. Scholarly communication 5. E-resources	1. Information retrieval 2. Information behaviour 3. Information literacy 4. Library services 5. Organization and management	1. Social media 2. Data 3. Web 4. E-government 5. Information retrieval
Top 6-10	6. Scholarly communication 7. Knowledge organization 8. User behaviour 9. Data analytics 10. Information systems 11. Knowledge management 12. Software development 13. Information services 14. Information quality 15. Knowledge network	6. Library collections 7. Organization and management 8. Library reference services 9. Planning and assessment 10. Review and conceptualization of academic libraries 11. Information organization 12. Library website and web services 13. New technologies 14. Innovative or unique library programs/services 15. Digital libraries	6. Scholarly communication 7. Digital libraries and metadata 8. Knowledge organization 9. Library collections 10. Library personnel 11. Research in LIS 12. Social media 13. Spaces and facilities 14. Information/ knowledge management 15. Library information systems	6. Information literacy 7. Government 8. Students 9. Classification 10. Evaluation 11. Collaboration 12. Information seeking 13. Assessment 14. Bibliometrics 15. Knowledge management

Table 4 Comparison of research topics in information science found in this research and previous studies (cont.)

Items	This research	Luo and McKinney (2015)	Togia and Malliari (2017)	Liu and Yang (2019)
	16. Digital collection management 17. Information retrieval 18. Data modeling 19. Information development 20. Information governance 21. Academic library management 22. Digital libraries 23. Information economy 24. Knowledge mapping 25. Digital resources 26. Information education 27. Information literacy 28. Information marketing 29. Information security 30. Knowledge transfer	16. Spaces and facilities 17. System and technical services 18. Collaboration between libraries 19. Legal issues 20. Librarian/faculty relationships 21. Outreach 22. LIS education 23. Librarians' research activities and output 24. Data services	16. LIS theory 17. Informetrics 18. Other (e.g. information history, library cooperation)	16. Scholarly communication 17. User studies 18. Citation analysis 19. Information management 20. Information behavior

Conclusion

Analyses of research data in the field of information science by applying the topic modeling method to articles that appear in journals in the Web of Science (an international database that lists the names of journals with high impact factors in different fields) have led to obtaining of research topics in the field that will benefit the revision of undergoing research work, setting research issues in study programs at graduate level, and guidelines for the development of research topics corresponding to international research directions. This will increase the chances of publication of information science research in high-quality international journals. In addition, this research provides guidelines for the development and extension of future research, as follows: (1) the research topics identified here can be developed into tools for the retrieval of research data, such as ontology, subject headings, and thesaurus, owing to its structure that includes 30 main topics, which can be divided into sub-topics by considering the word group appearing under each topic. Moreover, relationships can be identified between topics containing similar or overlapping word groups, which will greatly benefit future research data management systems; (2) this research presents guidelines for the analysis of research data by drawing on data from the past seven years from a large database containing 30,571 records, with a focus on information science. The topic modeling method can also be used to analyze research data from large data sources in other fields.

Acknowledgements

This research formed part of a doctoral dissertation supported by the Digital Humanities Research Group, Department of Information Science, Faculty of Humanities and Social Sciences, Khon Kaen University, Thailand.

References

Alabawi, R., Yeap, T. H., and Benyoucef, M. (2020). Using topic modeling methods for short-text data: A comparative analysis. **Frontiers in Artificial Intelligence**, **3**(article 42), 1-22. doi:10.3389/ frai.2020.00042

Blei, D. M. (2012). Latent dirichlet allocation. **Communications of the ACM**, **55**(4), 77-84.

Blei, D. M., Ng, A. Y., and Jordan, M. I. (2003). Latent dirichlet allocation. **Journal of Machine Learning Research**, **3**, 993-1022.

Borko, H. (1968). Information science: What is it? **Journal of the Association for Information Science and Technology**, **19**(5), 3-5.

Cheng, V. C., et al. (2014). Probabilistic aspect mining model for drug reviews. **IEEE Transactions on Knowledge and Data Engineering**, **26**(8), 2002-2013.

Claver, E., González, R., and Llopis, C. (2000). An analysis of research in information systems (1981-1997). **Information and Management**, **37**(4), 181-195.

Farzindar, A. and Inkpen, D. (2015). Natural language processing for social media. **Synthesis Lectures on Human Language Technologies**, **8**(2), 1-166. doi:10.2200/S00659ED1V01Y201508HLT030

Hawkins, D. T., Larson, S. E., and Caton, B. Q. (2003). Information science abstracts: Tracking the literature of information science-Part 2: A new taxonomy for information science. **Journal of the Association for Information Science and Technology**, **54**(8), 771-781.

Hussey, R., Williams, S., and Mitchell, R. (2012). Automatic keyphrase extraction: A comparison of methods. **In The 4th International Conference on Information Process, and Knowledge Management. (**pp. 18-23). Valencia, Spain.

Liu, G. and Yang, L. (2019). Popular research topics in the recent journal publications of library and information science. **Journal of Academic Librarianship**, **45**(3), 278-287.

Liu, L., Tang, L., Dong, W., Yao, S., and Zhou, W. (2016). An overview of topic modeling and its currentapplications in bioinformatics. **Springerplus**, **5**, 1-22. doi:10.1186/s40064-016-3252-8.

Luo, L. and McKinney, M. (2015). JAL in the past decade: A comparative analysis of academic library research. **Journal of Academic Librarianship**, **41**(2), 123-129.

Roder, M., Both, A., and Hinneburg, A. (2015). Exploring the space of topic coherence measures. **In WSDM '15: Proceedings of the Eighth ACM International Conference on Web Search and Data Mining**. (pp. 399-408). Shanghai. doi:10.1145/2684822.2685324.

Steyvers, M. and Griffiths, T. (2007). Probabilistic topic models. In Landauer, T.K., McNamara, D.S., Dennis, S., and Kintsch, W. (Eds.). **Handbook of Latent Semantic Analysis.** (pp. 427-448). New Jersey: Lawrence Erlbaum Associates Publishers.

Stock, W. G. and Stock, M. (2013). **Handbook of Information Science**. Boston, MA: De Gruyter Saur.

Syed, S. and Spruit, M. (2017). Full-text or abstract? Examining topic coherence scores using latent dirichlet allocation. **In IEEE International Conference on Data Science and Advanced Analytics (DSAA)**. (pp. 165-174). Tokyo, Japan. doi:10.1109/DSAA.2017.61.

Togia, A. and Malliari, A. (2017). Research methods in library and information science. In Oflazoglu, S. (Ed.). **Qualitative Versus Quantitative Research**. (pp. 43-64). London: InTechOpen.

Websites

Gerrish, S. M. and Blei, D. M. (2011). Predicting legislative roll calls from text. **The 28th International Conference on Machine Learning**. (pp. 489-496). Bellevue, WA. Retrieved December 15, 2020, from https://icml.cc/2011/papers/333_icmlpaper.pdf.

iSchools Inc. (2015). **About the iSchools Organization**. Retrieved December 21, 2020, from https://ischools.org/About.

Larsen, R. L. (2008). **History of the iSchools**. Retrieved December 21, 2020, from https://ischools.org/ resources/Documents/History-of-the-iSchools-2009.pdf

Maceviciute, E. and Wilson, T. D. (2002). The development of the information management research area. **Information Research, 7**(3). Retrieved December 21, 2020 from http://InformationR.net/ir/7-3/paper133.html.

Meriam, P. (2012). **Very basic strategies for interpreting results from the topic modeling tool**. Retrieved December 21, 2020, from https://miriamposner.com/blog/very-basic-strategies-for-interpreting-results-from-the-topic-modeling-tool/.

Waal, A. and Barnard, E. (2008). Evaluating topic models with stability. **Proceedings of the Nineteenth Annual Symposium of the Pattern Recognition Association of South Africa**. Retrieve November 2, 2020, from http://www.prasa.org/ proceedings/ 2008/prasa08-13.pdf.

Xie, P. and Xing, E. P. (2013). Integrating document clustering and topic modeling. **Proceedings of the Twenty-Ninth Conference on Uncertainty in Artificial Intelligence**. (pp. 694-703). Bellevue, WA. Retrieved November 2, 2020, from https://arxiv.org/ftp/arxiv/papers/1309/1309.6874.pdf.