Journal of Library and Information Studies 19:1 (June 2021) pp.1-17
doi: 10.6182/jlis.202106_19(1).001

# Adoption of Data Mining Methods in the Discipline of Library and Information Science

**Marie Katsurai[1], Soohyung Joo[2]**

## Abstract

The purpose of this paper is to explore the recent trends of data mining method adoption in the library and information science (LIS) discipline. Bibliographic records from the data mining and LIS fields were collected respectively from the Scopus database. A dictionary of data mining method terms was constructed based on a rule-based textual analysis. Using the dictionary, this study investigated a range of prevalent data mining methods utilized in recent LIS studies. The findings of this study reveal different areas of data mining methods employed in LIS, such as big data, machine learning, text mining, information retrieval, and dimension reduction. The study also confirms the recent popularity of machine learning techniques in LIS research.

Keywords: Library and Information Science; Text Mining; Vocabulary Construction; Bibliometric Analysis; Computational Methods

## 1. Introduction

Library and Information Science (LIS) has been a distinct academic discipline that has applied a wide variety of theories and practices of librarianship, information technologies, scholarly communication, information use behaviors, and social context surrounding information eco-systems (Hjørland, 2000; Timakum, Kim, & Song, 2018). The nature of LIS has been interdisciplinary, and accordingly, collaboration has been active in the research domain of LIS (Han et al., 2014; Hildreth & Aytac, 2007; Jabeen, Yun, Rafiq, Jabeen, & Tahir, 2015; Walters & Wilder, 2016). Various sub-fields are present in the LIS discipline, ranging from librarianship, information seeking, information retrieval, knowledge organization, information literacy, management, digital preservation, digital libraries, and to others (Koufogiannakis, Slater, & Crumley, 2004; Tuomaala, Järvelin, & Vakkari, 2014). In an effort to address various research questions from such diverse subordinate fields in LIS, researchers have adopted a wide variety of evolving research methods (Hider & Pymm, 2008).

In recent decades, computational methods have begun to be adopted as compelling research tools to analyze different types of data across various disciplines (Cioffi-Revilla, 2014; Hox, 2017). Researchers in LIS have also benefited from emerging computational methods including data mining techniques. Recently, the LIS discipline has undergone a substantial change in research scopes and methods to properly respond to the development of information technology, changes in information ecosystems,

[1] Department of Intelligent Information Engineering and Sciences, Doshisha University, Kyoto, Japan
[2] School of Information Science, University of Kentucky, Lexington, Kentucky, USA
* Corresponding Author: Soohyung Joo, Email: soohyung.joo@uky.edu

and increased interdisciplinary (Aharony, 2012; Timakum et al., 2018). The community of LIS researchers has begun to recognize the importance of embracing emerging research techniques as well as interdisciplinary collaborations (Risso, 2016; Togia & Malliari, 2017). Computational methods can further broaden and diversify the methodologies in LIS (Bowker, 2018).

This study investigates to which extent data mining methods have been adopted in LIS research in the recent decade. A significant number of prior literature have explored the uses of different types of research methods in LIS (Chu & Ke, 2017; Malliari & Togia, 2016). However, less research has focused on the data mining methods in the investigation of research methods in LIS. To better understand the adoption trends of data mining methods in LIS research, we built a taxonomy of data mining methods based on a rule-based textual analysis. Then, we investigated which data mining methods were employed in LIS research in recent years.

## 2. Literature Review

A wide variety of research methods have been applied in LIS for various research topics. In an attempt to understand the nature of the LIS domain, researchers have identified types of research methods utilized in LIS research. For example, Chu (2015) investigated a total of 1,162 research articles in three LIS journals published between 2001 and 2010, including *Journal of the Association for Information Science and Technology* (JASIST), *Journal of Documentation* (JD), and *Library and Information Science Research* (LISR). She found that content analysis, surveys, and experiments were among the top choices of research methods in LIS. Chu and Ke (2017) further identified types of methodological strategies in LIS research. Their findings revealed that experiments, bibliometrics, questionnaires, content analysis, theoretical analysis, and interviews were widely employed in LIS. Similarly, Ferran-Ferrer, Guallar, Abadal, and Server (2017) examined types of research methods applied in LIS studies published in Spanish journals. They conducted content analysis of 394 research articles published in seven top-tier LIS journals in Spain. The findings of their study uncovered that qualitative and quantitative approaches were employed in similar proportions. Hildreth and Aytac (2007) explored research activities and behaviors of library practitioners based on the content analysis of a sample of 23 LIS journals. They found that library practitioners preferred qualitative methods over quantitative methods. Particularly, qualitative methods were more often used for research involving textual, verbal, or pictorial data. Morris and Cahill (2017) investigated research methods in the area of school librarianship research. After examining over 200 research articles, they found that most of studies in that area relied on qualitative methods while quantitative methods were limitedly used for descriptive analysis.

Survey has been one of the most popular methodological approaches in LIS. In their investigation of LIS research trends, Malliari and Togia (2016) found that survey was most widely utilized as a procedure to collect data in LIS research. Togia and Malliari (2017) analyzed the content of research articles published in five representative LIS journals between 2011

and 2016. Their findings indicated that the most frequently applied research strategy was categorized as a survey method, accounting for approximately 37%. Similarly, Ullah and Ameen (2018) found that survey was chosen as one of the primary research methods in LIS and descriptive statistics was most often used to analyze survey data.

Prior literature has also explored statistical analysis uses in LIS research. Zhang, Zhao, and Wang (2016) investigated the adoption of statistical methods in six major journals in LIS, such as *Library Quarterly* (LQ), JASIST, JD, LISR, *Information Processing & Management*, and *Journal of Information Science*. From the content analysis of 5,175 articles published between 1999 and 2013, they found that approximately 28.9% employed any kind of statistical analysis. Their findings revealed a growing trend of statistical analysis use in LIS over the years. Zhang, Wang, Zhao, and Cai (2018) further examined types of statistical analyses in LIS research. Most common techniques used in LIS research included t-tests, correlation analysis, analysis of variance, and chi-square tests. These statistical techniques were more likely to be used in the sub-fields of information retrieval and information search behaviors in LIS. Hildreth and Aytac (2007) specifically investigated research activities of library practitioners. They observed that descriptive statistics was dominantly used in research articles contributed by library practitioners.

Textual analysis techniques have been also utilized in LIS. Bowker (2018) recognized potential benefits of computer-based corpus linguistics, particularly its methodological implications for LIS research. She claimed that corpus-based linguistics could complement LIS research by broadening the scope and capacity of research methods in LIS. Timakum et al. (2018) investigated research trends in LIS using text mining techniques, such as co-word analysis, text summarization, and topic modeling. Their study analyzed full-text of research articles from six top-tier LIS journals, and identified interdisciplinarity in LIS research over the past decade. Primary research topics in LIS extracted from topic modeling ranged from academic libraries, digital libraries, information retrieval, digital information, and to others. Joo, Choi, and Choi (2018) surveyed the domain of knowledge organization, which is one of the distinctive research areas in LIS, based on text mining. They observed that topics related to domain analysis and ontologies received increased attention recently.

Analyzing topics and trends of publications has been an important issue in the field of computer sciences (CS) as well, and several computational methodologies have been exploited. In the machine learning and natural language processing communities, a variety of topic models have been developed. For example, Datta, Lakdawala, and Sarkar (2018) analyzed corpora of research publications across four sub-domains of CS using a topic model to show domain-specific topics. Since it is not always easy to interpret the meanings of topics produced by topic models, word statistics has still been recognized as an effective bibliographic analysis tool. Liu et al. (2014) revealed research trends in the field of Human-Computer Interaction using keywords of conference papers. Salatino, Osborne, and Motta (2017) analyzed co-occurrence relationships between CS topics using keywords of three million papers, which were extracted from Scopus.

However, these studies focused on analyzing CS publications only, and relationships between CS and other fields have not been widely investigated.

As shown in this literature review, existent literature has examined various types of research methods employed in LIS. However, there is little research that specifically focused on data mining methods. The discipline of LIS has been interdisciplinary in nature (Chang, 2018), and recently the iSchool movement has expanded collaboration efforts between LIS and other disciplines (Shu & Mongeon, 2016). However, there is no convincing evidence to which extent data mining methods have been employed in LIS. This study fills the gap in existent literature by investigating the adoption of data mining methods in LIS.

## 3. Research Questions

The objectives of this study are two-folded: (1) to build a dictionary of data mining methods and (2) to investigate the adoption of data mining methods in LIS. The following research questions guided this investigation:

RQ 1 – What are the recent research methods used for data mining and analysis?

RQ 2 – What are the data mining methods frequently applied in LIS research?

RQ 3 – Are there any changes in the use of data mining methods in LIS between 2009 and 2018?

## 4. Research Methods

### *4.1. Data collection*

To answer the research questions, we analyzed bibliographic data collected from both LIS and data mining related publications. Bibliographic records, particularly titles and abstracts, typically include information about research methods applied in the study. Two sets of bibliographic data were collected for this study. We first selected representative publication venues for "LIS" and "data mining and analysis" respectively from the Google Scholar Metrics. Google Scholar Metrics is one of the widely accepted altmetrics in scholarly communications, and it provides the ratings of publications in certain disciplines based on the *h*-index (Google Scholar, 2020). Two sub-categories of research areas were chosen from Google Scholar Metrics: "Social Sciences – Library & Information Science" and "Computer Science – Data Mining & Analysis" respectively. Table 1 presents a list of journals and proceedings that we chose for this study. We extracted bibliographic records from these journals/proceedings, including titles and abstracts, published between 2009 and 2018 from the Scopus Abstract and Citation database. The queries were made using ISSN, and the following limiters were further applied: (a) publication years between 2009 and 2018; (b) document type: articles or conference papers; and (c) language: English. As this study focuses on research methods, we collected only articles or conference papers while excluding reviews, letters, editorials, and other types of documents. For some of the CS proceedings, we were not able to find an ISSN. For those cases, conference names were

**Table 1. A List of Publication Venue**

| Social Sciences - Library & Information Science | | Computer Science - Data Mining & Analysis | |
|---|---|---|---|
| Publication name | h5-index | Publication name | h5-index |
| Journal of the Association for Information Science and Technology | 60 | ACM SIGKDD International Conference on Knowledge Discovery and Data Mining | 86 |
| Scientometrics | 57 | IEEE Transactions on Knowledge and Data Engineering | 77 |
| Journal of Informetrics | 39 | International Conference on Artificial Intelligence and Statistics (2016-2018) | 52 |
| The Journal of Academic Librarianship | 33 | ACM International Conference on Web Search and Data Mining | 51 |
| Online Information Review | 30 | ACM Conference on Recommender Systems (2010-2018) | 45 |
| Journal of Information Science | 29 | IEEE International Conference on Data Mining Workshop (2009-2017) | 44 |
| College & Research Libraries | 28 | Data Mining and Knowledge Discovery | 40 |
| Journal of Documentation | 25 | Knowledge and Information Systems | 38 |
| Portal: Libraries and the Academy | 24 | ACM Transactions on Intelligent Systems and Technology (2010-2018) | 38 |
| The Electronic Library | 24 | SIAM International Conference on Data Mining | 36 |
| Aslib Journal of Information Management | 23 | Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery (2011-2018) | 35 |
| Information Development | 23 | IEEE International Conference on Big Data (2003-2005) | 33 |
| Learned Publishing | 21 | European Conference on Machine Learning and Knowledge Discovery in Databases | 30 |
| Journal of the Medical Library Association: JMLA | 21 | Journal of Big Data (2014-2018) | 27 |
| Library & Information Science Research | 21 | ACM Transactions on Knowledge Discovery from Data (TKDD) | 27 |
| Library Hi Tech | 21 | Social Network Analysis and Mining (2011-2018) | 25 |
| Journal of Librarianship and Information Science | 20 | Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD) | 23 |
| Information Research | 20 | Advances in Data Analysis and Classification | 22 |
| New Library World (Information and Learning Sciences) | 20 | IEEE International Conference on Data Science and Advanced Analytics (2014-2018) | 20 |
| Library Philosophy and Practice | 19 | | |

used alternatively to construct a query. Not all years of data were collected for some of the CS proceedings. For example, we obtained only partial data between 2016 and 2018 for the proceedings of *International Conference on Artificial Intelligence and Statistics*. In addition, we excluded *Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis* (ranked at 20th) from data collection because it was not searchable. In this way, two distinct sets of corpora were constructed: a set of Data Mining (DM) papers $\Omega_{DM}$ and a set of LIS papers $\Omega_{LIS}$.

### *4.2 Constructing a vocabulary of computational analysis method terms*

To analyze how data mining methods are adopted in other fields, we first needed to identify research method terms that are specific to data mining analysis. In this study, we propose a novel approach to constructing a vocabulary of data mining methods based on the rule-based content analysis of research articles. Below, we provide a detailed description of our three-step approach.

*Step 1*. Constructing *n*-grams from the selected DM paper titles. Research methods employed in the selected DM papers are typically presented in their titles. We first divided a title of each paper into multiple phrases using prepositions and punctuations, which are listed in Figure 1, as separators. For example, suppose the title of paper $p \in \Omega_{DM}$ is "*Layered Hidden Markov Models for Real-Time Daily Activity Monitoring Using Body Sensor Networks*." After applying the Porter stemming, the separators produce the following phrases: "layer hidden markov model," "real-tim daili activ monitor," and "bodi sensor network." Then, from the phrases extracted, we constructed all possible bigrams, trigrams, and four-grams. We removed an *n*-gram if it ends with a stop word or a half of the *n*-gram corresponds to stop words. We empirically found that the latter rule effectively filters unnecessary terms such as "the above thing" and "we find that." A set of the resulting *n*-grams ($n = 2, 3, 4$) is denoted by $V_{DM\_all}$.

*Step 2*. Extracting method terms from all *n*-grams on the basis of clue words and word frequency. Because the terms included in $V$ are not necessarily directly related to data mining methods (e.g., the term "body sensor networks" does not belong to any research method but devices), we needed to discard those terms that do not indicate research methods. To extract only data mining method terms, we extracted from $V_{DM\_all}$ *n*-grams whose last words correspond to the following clue words: "model," "method," "techniqu" (the stemmed version of "technique"), or "theori" (the stemmed version of "theory"). A set of the n-grams extracted here is denoted by $V_{method}$. In the previous example, the *n*-gram terms "layer hidden markov model," "hidden markov model," and "markov model" are regarded as research method terms. This approach might also need to extract uninformative terms, such as "learning method," "data mining method," and "novel model." To resolve this problem, we

| . / , / : / ; / for / on / to / by / using/ in / from / based on / with / via / through |
|---|

**Figure 1. Prepositions and punctuations that were used to divide each paper title into phrases**

evaluated the informativeness of an *n*-gram by investigating the frequency of each single word of the *n*-gram. Specifically, we calculated the document frequency of all single words using all abstracts in $\Omega_{DM}$. Preparing a threshold value $T$, if a word appeared in more than $T$ abstracts of the whole dataset, the word is stored in a stop word set $C$. In experiments, we experimentally set $T$ to 10% of the size of $\Omega_{DM}$. Then, if all words that compose a target *n*-gram belong to $C$, we judged that *n*-gram as uninformative. A set of method terms that were considered as to be informative is denoted by $V'_{method}$.

*Step 3*. We further defined method-related adjectives and utilized them to expand the vocabulary. Some method terms were not detected from those predefined clue words. For example, a support vector machine, which is one of the most popular methods in machine learning, does not contain any predefined clue word. To solve this problem, we extracted *n*-grams having any one of clue words from $V'_{method}$ and used their first (*n*-1)-grams as technical adjectives. For example, if the *n*-gram is "hidden markov model," then we regarded "hidden markov" as a technical adjective. Finally, *n*-grams in $V_{DM\_all}$, which have one of the technical adjectives, are collected to construct an additional set of method names, $V^a_{method}$.

Throughout these three steps, we obtained the final vocabulary of data mining method terms as $W=\{V'_{method} \cup V^a_{method}\}$. Using the abstracts of LIS articles in $\Omega_{LIS}$ as documents, we counted the document frequency (*df*) of the method terms in $W$. The set of all *n*-grams extracted from $\Omega_{LIS}$ is denoted by $V_{LIS\_all}$.

## 5. Results

The number of all *n*-grams in $V_{DM\_all}$ was 93,142. At *Step 2* of the proposed method, the number of *n*-grams in $V_{method}$ was 2,620. The size of stop word set $C$ (i.e., the number of words that were regarded as less informative) was 177. Examples of the words in $C$ are shown in Figure 2. Using $C$, $V_{method}$ was reduced to $V'_{method}$, in which $|V_{method}|=2{,}408$. The number of the methodological adjectives was 1,689, which produces an additional vocabulary, $V^a_{method}$. The number of *n*-grams in $V^a_{method}$ was 6,758. Finally, the number of method terms in $W$ was 9,166 (2,408 + 6,758).

The number of all *n*-grams in $V_{LIS\_all}$ was 1,163,709. Using the abstracts of $\Omega_{LIS}$, we counted the document frequency of each method-related *n*-gram listed in $W$. The number of *n*-grams that were included in both $W$ and $V_{LIS\_all}$ turned out to be 699.

We divided all documents in $\Omega_{DM}$ into two temporal subsets: Period 1 (2009–2013) and Period 2 (2014–2018). Table 2 shows the top 30 method terms in $\Omega_{DM}$ for each time period. Popular method terms are identified as follow: (1) for both the periods, common popular topics include

| |
|---|
| data, propos, paper, us, result, algorithm, method, show, model, problem, approach, base, perform, set, learn, inform, gener, experi, effect, also, present, provid, differ, applic, demonstr, new, two, dataset, effici, howev, studi, time, evalu, network, exist, user, system, larg, novel |

**Figure 2. Examples of frequent words that were stored in the stop word set C**

**Table 2. Top 30 Method Terms**

| Period 1: 2009—2013 (6,764 titles) | | Period 2: 2014—2018 (9,265 titles) | |
|---|---|---|---|
| Method term | *df* | Method term | *df* |
| matrix factor | 69 | big data | 167 |
| collabor filter | 61 | **neural network** | 103 |
| active learn | 61 | matrix factor | 100 |
| topic model | 56 | topic model | 80 |
| big data | 52 | anomali detect | 75 |
| **support vector** | 45 | commun detect | 65 |
| outlier detect | 39 | collabor filter | 57 |
| **vector machin** | 38 | gaussian process | 54 |
| **bayesian network** | 38 | active learn | 49 |
| dimension reduct | 37 | sentiment analysi | 48 |
| **support vector machin** | 35 | link predict | 47 |
| semi-supervis learn | 34 | outlier detect | 46 |
| gaussian process | 31 | transfer learn | 44 |
| link predict | 30 | **graphic model** | 41 |
| anomali detect | 29 | support vector | 38 |
| transfer learn | 29 | **heterogen inform network** | 36 |
| **text classif** | 28 | **multi-task learn** | 33 |
| **privaci preserv** | 28 | vector machin | 31 |
| **markov model** | 28 | nearest neighbor | 30 |
| **document cluster** | 27 | support vector machin | 29 |
| commun detect | 26 | reinforc learn | 28 |
| sentiment analysi | 25 | metric learn | 27 |
| reinforc learn | 24 | **differenti privat** | 26 |
| mixture model | 24 | mixture model | 26 |
| neural network | 22 | semi-supervis learn | 25 |
| metric learn | 21 | **convolute neural network** | 24 |
| **subspace cluster** | 20 | bayesian network | 23 |
| **hidden markov** | 19 | **gradient descent** | 22 |
| **hidden markov model** | 19 | **recurrent neural network** | 22 |
| nearest neighbor | 19 | dimension reduct | 21 |

*Note*. Terms that are specific to the time period are highlighted in bold.

collaborative filtering, matrix factorization, big data, and anomaly/outlier detection; (2) support vector machines and methods for text classification/document clustering were amongst popular topics in Period 1; (3) recently popular methods are likely to be associated with machine learning, e.g., neural network, multi-task learning, and metric learning. We further examined the top 10 terms each year (Table 3). We observed that neural network analysis related stemmed terms (e.g., "neural network" and "reinforc learn") appeared among the top 10 terms most recently. This reveals the recent popularity of neural network analysis in the data mining field.

Next, we investigated the adoption of data mining methods in LIS. Table 4 lists the top 30 method terms appeared in LIS articles for each time period. We found that the most frequent terms for both two periods were related to structural equation modeling. However, the number of DM papers ($\Omega_{DM}$) that include the term "structural equation" was only two. In Period 1, popular data mining methods were related to information retrieval, machine learning, regression, among others. In Period 2, we observed that machine learning, big data, and text mining are highly ranked. We further explored top 10 terms for each year as shown in Table 5. We found that structural equation modeling was popular consistently across the investigation period. The term "big data" was ranked among the top ten since 2015, revealing the increased attention on big data analysis. Information retrieval was popular in earlier years particularly in 2009 and 2010.

## 6. Discussion

This study investigated what kinds of data mining methods have been employed in LIS in the past decade. We constructed a dictionary of data mining methods based on the rule-based analysis of selected DM papers. Then, we matched those data mining method terms with the abstracts of LIS research articles from representative journals.

First, we constructed a dictionary of data mining method terms by analyzing the titles of 16,029 research papers in the area of data mining and analysis. Not surprisingly, highly ranked method terms are relevant to machine learning, for example, matrix factorization, machine learning, support vector, among others. Text mining turned out to be another popular method in the data mining and analysis field, such as topic models, text classification, sentiment analysis, and document clustering. In addition, we compared the two time periods, i.e., Period 1 (2009-2013) vs. Period 2 (2014-2018) to examine the change of popular methods over time. In Period 1, the top methods include matrix factorization, collaborative filtering, active learning, topic modeling, and others. We noticed that text mining was more popular in Period 1 than Period 2. For example, the phrases of topic models, text classification, and document clustering were highly ranked in Period 1. In Period 2, we observed machine learning related terms among the top terms. In particular, we found that artificial intelligence-based methods are more often observed in Period 2, such as neural network, reinforcement learning, convolutional neural network, and recurrent neural network.

**Table 3. Top 10 Method Terms in DM Paper Titles Published in each Year**

| Year | Paper num. | Method terms (*df*) |
|---|---|---|
| 2018 | 1,712 | neural network (34) / big data (22) / matrix factor (18) / **reinforc learn** (15) / anomali detect (14) / commun detect (13) / outlier detect (12) / topic model (11) / link predict (11) / gaussian process (11) |
| 2017 | 1,877 | neural network (37) / big data (25) / matrix factor (17) / topic model (16) / anomali detect (15) / outlier detect (14) / collabor filter (14) / commun detect (13) / sentiment analysi (12) / transfer learn (12) |
| 2016 | 1,913 | **neural network** (24) / big data (20) / matrix factor (19) / topic model (17) / anomali detect (13) / gaussian process (13) / commun detect (11) / collabor filter (10) / **graphic model** (10) / outlier detect (10) |
| 2015 | 2,075 | big data (76) / matrix factor (27) / anomali detect (17) / topic model (17) / commun detect (14) / sentiment analysi (13) / **multi-task learn** (12) / support vector (11) / vector machin (10) / support vector machin (10) |
| 2014 | 1,688 | big data (24) / topic model (19) / matrix factor (19) / activ learn (17) / anomali detect (16) / collabor filter (16) / commun detect (14) / gaussian process (14) / link predict (11) / transfer learn (10) |
| 2013 | 1,826 | **big data** (52) / matrix factor (20) / collabor filter (18) / topic model (17) / outlier detect (13) / support vector (13) / activ learn (11) / vector machin (11) / **transfer learn** (11) / gaussian process (10) |
| 2012 | 1,410 | activ learn (15) / matrix factor (15) / topic model (12) / collabor filter (10) / **metric learn** (10) / **commun detect** (9) / **sentiment analysi** (7) / support vector (7) / text classif (7) / outlier detect (7) |
| 2011 | 1,371 | matrix factor (16) / collabor filter (15) / activ learn (14) / topic model (11) / support vector (11) / bayesian network (9) / text classif (9) / vector machin (9) / support vector machin (9) / **privaci preserv** (9) |
| 2010 | 1,177 | **collabor filter** (16) / dimension reduct (11) / matrix factor (11) / **support vector** (10) / **mixtur model** (10) / **gaussian process** (10) / **vector machin** (9) / activ learn (8) / **support vector machin** (8) / **link predict** (8) |
| 2009 | 980 | activ learn (13) / bayesian network (11) / topic model (9) / dimension reduct (9) / matrix factor (7) / markov model (7) / outlier detect (6) / text classif (6) / document cluster (6) / anomali detect (6) |

*Note*. Emerging terms are highlighted.

Second, we investigated what kinds of computational method terms occurred in LIS articles. Interestingly, "equat model" and "structure equat model," which indicate structural equation modeling (SEM), turned out to be the most frequent method terms in LIS. Although SEM related terms appeared in the data mining method dictionary, it is considered more of a social science method. SEM has been popular in LIS to statistically test complex research models with multiple variables. Despite the popularity of SEM in LIS, it is not part of the mainstream methods

**Table 4. Top 30 Data Mining Method Terms Appeared in LIS Articles**

| Period 1: 2009—2013 (5,141 abstracts) | | Period 2: 2014—2018 (7,328 abstracts) | |
|---|---|---|---|
| Method term | *df* | Method term | *df* |
| equat model | 64 | equat model | 131 |
| structur equat model | 62 | structur equat model | 128 |
| inform retriev system | 38 | **sampl techniq** | 97 |
| regress model | 35 | least squar | 70 |
| conceptu model | 32 | mix method | 60 |
| least squar | 26 | **big data** | 59 |
| busi model | 26 | partial least squar | 59 |
| theoret model | 25 | text mine | 49 |
| logist regress | 21 | regress model | 48 |
| partial least squar | 19 | logist regress | 45 |
| text mine | 19 | **topic model** | 45 |
| **keyword search** | 17 | theoret model | 45 |
| **text classif** | 16 | sentiment analysi | 41 |
| **princip compon** | 16 | conceptu model | 36 |
| support vector | 15 | inform retriev system | 34 |
| sentiment analysi | 14 | correl analysi | 34 |
| **princip compon analysi** | 14 | sampl method | 34 |
| support vector machin | 14 | statist model | 24 |
| vector machin | 14 | support vector | 24 |
| **vector space** | 14 | support vector machin | 23 |
| correl analysi | 13 | vector machin | 23 |
| linear regress | 13 | queri expans | 19 |
| **local commun** | 13 | busi model | 18 |
| **visual inform** | 13 | **neural network** | 18 |
| hierarch cluster | 12 | **integr model** | 18 |
| **resour manag** | 12 | hierarch cluster | 17 |
| queri expans | 11 | role of social | 17 |
| sampl method | 11 | analyt method | 16 |
| **document retriev** | 11 | linear regress | 16 |
| topic model | 11 | **commun channel** | 16 |

*Note*. Terms that are specific to the time period are highlighted in bold.

**Table 5. Top 10 Method Terms in LIS Articles for each Year**

| Year | Paper num. | Method terms (*df*) |
|---|---|---|
| 2018 | 1,571 | equat model (29) / structur equat model (29) / sampl techniqu (27) / mix method (20) / least squar (19) / partial least squar (18) / big data (16) / regress model (13) / support vector (12) / correl analysi (12) |
| 2017 | 1,464 | equat model (32) / structur equat model (30) / sampl techniqu (17) / theoret model (16) / big data (16) / topic model (15) / text mine (13) / mix method (10) / least squar (10) / inform retriev system (9) |
| 2016 | 1,465 | equat model (27) / structur equat model (27) / sampl techniqu (26) / least squar (17) / conceptu model (13) / partial least squar (13) / mix method (12) / big data (11) / regress model (11) / sentiment analysi (10) |
| 2015 | 1,367 | equat model (22) / structur equat model (22) / regress model (13) / sampl techniqu (13) / **big data** (12) / **mix method** (12) / least squar (11) / text mine (10) / partial least squar (10) / topic model (9) |
| 2014 | 1,461 | equat model (21) / structur equat model (20) / **sampl techniqu** (14) / least squar (13) / logist regress (11) / partial least squar (9) / inform retriev system (7) / **topic model** (7) / correl analysi (7) / commun channel (7) |
| 2013 | 1,106 | equat model (23) / structur equat model (22) / regress model (12) / inform retriev system (11) / conceptu model (8) / **support vector** (6) / logist regress (6) / theoret model (6) / least squar (6) / **partial least squar** (6) |
| 2012 | 950 | equat model (11) / structur equat model (10) / regress model (8) / busi model (7) / **link analysi** (6) / conceptu model (6) / least squar (6) / **queri expans** (5) / text mine (5) / vector space (4) |
| 2011 | 1,018 | equat model (13) / structur equat model (13) / **theoret model** (9) / least squar (8) / regress model (7) / **logist regress** (6) / **hierarch cluster** (6) / busi model (6) / keyword search (5) / **video retriev** (5) |
| 2010 | 1,051 | inform retriev system (10) / conceptu model (8) / **text mine** (7) / equat model (6) / structur equat model (6) / **busi model** (5) / regress model (4) / text classif (4) / **sampl method** (4) / **vector space** (4) |
| 2009 | 1,016 | equat model (11) / structur equat model (11) / inform retriev system (9) / text classif (9) / conceptu model (7) / correl analysi (6) / least squar (5) / keyword search (4) / ir model (4) / regress model (4) |

*Note*. Emerging terms are highlighted.

in data mining. The results reveal that machine learning has been popular in LIS. We found that machine learning related terms are highly ranked, such as support vector machine and logistic regression. Some method terms can be categorized as statistical analysis, such as regression model and correlation analysis. We observed that some of the top terms were associated with information retrieval research, for example, information retrieval systems, vector space, and query expansion. Text mining related phrases also appeared among the frequent terms, such as topic modeling, text mining, and text classification. One of the noteworthy observations is that several of

the top methods can be categorized as statistical analysis, which are also widely utilized across social sciences, rather than typical data mining techniques. Therefore, not all method terms appeared among the top terms can be claimed to be data mining methods.

We further examined the trends of data mining method adoption in LIS. Method terms related to information retrieval were observed more often in Period 1 than Period 2, such as "information retriev system" and "vector space." In Period 1, we also observed two distinct stemmed terms indicating principal component analysis, including "princip compon" and "princip compon analysi." In Period 2, the term "big data" distinctly appeared among top 30 terms and also ranked at fifth. This implies that big data analysis received increased attention lately in LIS. However, the adoption of data mining methods was likely to be limited to the areas of scientometrics, informetrics, data analytics, or information retrieval. Those highly ranked method terms were more often appeared in informetrics or technology-focused journals, such as JASIST, *Scientometrics*, and *Journal of Informetrics*. Contributions to LIS publications have also been made by faculty in disciplines other than LIS, including computer science researchers (Chang, 2019; Walters & Wilder, 2016). On the contrary, library-context research has less utilized data mining methods yet. Library focused journals, such as *College & Research Libraries* and LISR, were more likely to rely on traditional social science methods (Malliar & Togia, 2016). Particularly, library practitioners, who have significantly contributed to the library focused journals, preferred using qualitative methods (Hildreth & Aytac, 2007).

This study yields methodological contributions. Most of prior studies that investigated LIS research methods relied on content analysis based on manual coding (e.g., Malliari & Togia, 2016). Manual content analysis can be effective to categorize types of methods, but it is time-consuming and requires ample amount of human effort. Thus, it might not be ideal to analyze a large-scale data. This study suggests a novel approach to building a dictionary of research methods in a certain discipline based on textual analysis. We devised a rule to extract method-related phrases from the titles of research articles. The dictionary developed in this study covers a comprehensive list of data analysis methods lately used in the area of data mining. The text mining approach enabled us to investigate a large size corpus, i.e., 12,469 articles collected from 20 LIS journals, efficiently. Given the recent popularity of data mining methods, the dictionary produced in this study can be used as a reference to examine data mining method adoption in other disciplines.

## 7. Conclusion

This study is one of the first attempt that explores the trends of data mining methods employed in LIS. We came up with a dictionary specific to data mining methods based on textual analysis. Then, we investigated what kinds of data mining methods have been adopted frequently in LIS. The findings of this paper raise the awareness about the benefits of data mining methods in LIS research. With the increased capability of computational tools, diverse data mining methods have become available for various LIS research agenda (Bowker, 2018). LIS research

can benefit from diversity of research methods. Data mining methods can serve as compelling tools to respond to various research questions in LIS.

This study is not without limitations. We cannot confirm that all method terms were detected through the method suggested herein. Not all computational analysis methods might include those cue words we used in this study. Also, we found that certain method terms extracted from data mining papers would represent statistical analysis in general, rather than data mining techniques. This study did not clearly distinguish data mining methods from the observed method terms. In addition, we only chose 19 publication venues in the area of "Data Mining and Analysis," which is a small part of the entire CS discipline, to build the dictionary. In addition, we plan to develop a more sophisticated method to filter unnecessary words for effective data mining in our future work.

## Acknowledgement

This research was partly supported by JST ACT- X grant number JPMJAX1909.

## References

Aharony, N. (2012). Library and Information Science research areas: A content analysis of articles from the top 10 journals 2007–8. *Journal of Librarianship & Information Science, 44*(1), 27-35. doi: 10.1177/0961000611424819

Bowker, L. (2018). Corpus linguistics is not just for linguists: Considering the potential of computer-based corpus methods for library and information science research. *Library Hi Tech, 36*(2), 358-371. doi: 10.1108/LHT-12-2017-0271

Chang, Y.-W. (2018). Examining interdisciplinarity of library and information science (LIS) based on LIS articles contributed by non-LIS authors. *Scientometrics, 116*(3), 1589-1613. doi: 10.1007/s11192-018-2822-7

Chang, Y.-W. (2019). Are articles in library and information science (LIS) journals primarily contributed to by LIS authors? *Scientometrics, 121*(1), 81-104. doi: 10.1007/s11192-019-03186-w

Chu, H. (2015). Research methods in library and information science: A content analysis. *Library & Information Science Research, 37*(1), 36-41. doi: 10.1016/j.lisr.2014.09.003

Chu, H., & Ke, Q. (2017). Research methods: What's in the name? *Library & Information Science Research, 39*(4), 284-294. doi: 10.1016/j.lisr.2017.11.001

Cioffi-Revilla, C. (2014). *Introduction to computational social science*. London, England: Springer. doi: 10.1007/978-1-4471-5661-1

Datta, S., Lakdawala, R., & Sarkar, S. (2018). Understanding the inter-domain presence of research topics in the computing discipline: An empirical study. *IEEE Transactions on Emerging Topics in Computing, 9*(1), 366-378. doi: 10.1109/TETC.2018.2869556

Ferran-Ferrer, N., Guallar, J., Abadal, E., & Server, A. (2017). Research methods and techniques in Spanish library and information science

journals (2012-2014). *Information Research, 22*(1).

Google Scholar. (2020). *Google Scholar Metrics*. Retrieved from https://scholar.google.com/intl/en/scholar/metrics.html

Han, P., Shi, J., Li, X., Wang, D., Shen, S., & Su, X. (2014). International collaboration in LIS: Global trends and networks at the country and institution level. *Scientometrics, 98*(1), 53-72. doi: 10.1007/s11192-013-1146-x

Hider, P., & Pymm, B. (2008). Empirical research methods reported in high-profile LIS journal literature. *Library & Information Science Research, 30*(2), 108-114. doi: 10.1016/j.lisr.2007.11.007

Hildreth, C. R., & Aytac, S. (2007). Recent library practitioner research: A methodological analysis and critique. *Journal of Education for Library & Information Science, 48*(3), 236-258.

Hjørland, B. (2000). Library and information science: Practice, theory, and philosophical basis. *Information Processing & Management, 36*(3), 501-531. doi: 10.1016/S0306-4573(99)00038-2

Hox, J. J. (2017). Computational social science methodology, anyone? *Methodology: European Journal of Research Methods for the Behavioral & Social Sciences, 13*(S1), 3-12. doi: 10.1027/1614-2241/a000127

Jabeen, M., Yun, L., Rafiq, M., Jabeen, M., & Tahir, M. A. (2015). Scientometric analysis of library and information science journals 2003–2012 using Web of Science. *International Information & Library Review, 47*(3/4), 71-82. doi: 10.1080/10572317.2015.1113602

Joo, S., Choi, I., & Choi, N. (2018). Topic analysis of the research domain in knowledge organization: A latent dirichlet allocation approach. *Knowledge Organization, 45*(2), 170-183. doi: 10.5771/0943-7444-2018-2-170

Koufogiannakis, D., Slater, L., & Crumley, E. (2004). A content analysis of librarianship research. *Journal of Information Science, 30*(3), 227-239. doi: 10.1177/0165551504044668

Liu, Y., Goncalves, J., Ferreira, D., Xiao, B., Hosio, S., & Kostakos, V. (2014). CHI 1994–2013: Mapping two decades of intellectual progress through co-word analysis. In M. Jones & P. Palanque (Chairs), *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI'14)* (pp. 3553-3562). New York, NY: Association for Computing Machinery. doi: 10.1145/2556288.2556969

Malliari, A., & Togia, A. (2016). An analysis of research strategies of articles published in library science journals: The example of library and information science research. *Qualitative & Quantitative Methods in Libraries, 5*(4), 805-818.

Morris, R. J., & Cahill, M. (2017). A study of how we study: Methodologies of school library research 2007 through July 2015. *School Library Research, 20*.

Risso, V. G. (2016). Research methods used in library and information science during the 1970–2010. *New Library World, 117*(1/2), 74-93. doi: 10.1108/NLW-08-2015-0055

Salatino, A. A., Osborne, F., & Motta, E. (2017). How are topics born? Understanding the research dynamics preceding the emergence of new areas. *PeerJ Computer Science, 3*, e119. doi: 10.7717/peerj-cs.119

Shu, F., & Mongeon, P. (2016). The evolution of iSchool movement (1988–2013): A bibliometric view. *Education for Information, 32*(4), 359-373. doi: 10.3233/EFI-160982

Timakum, T., Kim, G., & Song, M. (2018). A data-driven analysis of the knowledge structure of library science with full-text journal articles. *Journal of Librarianship & Information Science, 52*(2), 345-365. doi: 10.1177/0961000618793977

Togia, A., & Malliari, A. (2017). Research methods in library and information science, qualitative versus quantitative research. In S. Oflazoglu (Ed.), *Qualitative versus quantitative research* (pp. 43-64). Rijeka, Croatia: InTech. doi: 10.5772/intechopen.68749

Tuomaala, O., Järvelin, K., & Vakkari, P. (2014). Evolution of library and information science, 1965–2005: Content analysis of journal articles. *Journal of the Association for Information Science & Technology, 65*(7), 1446-1462. doi: 10.1002/asi.23034

Ullah, A., & Ameen, K. (2018). Account of methodologies and methods applied in LIS research: A systematic review. *Library & Information Science Research, 40*(1), 53-60. doi: 10.1016/j.lisr.2018.03.002

Walters, W. H., & Wilder, E. I. (2016). Disciplinary, national, and departmental contributions to the literature of library and information science, 2007–2012. *Journal of the Association for Information Science & Technology, 67*(6), 1487-1506. doi: 10.1002/asi.23448

Zhang, J., Wang, Y., Zhao, Y., & Cai, X. (2018). Applications of inferential statistical methods in library and information science. *Data and Information Management, 2*(2), 103-120. doi: 10.2478/dim-2018-0007

Zhang, J., Zhao, Y., & Wang, Y. (2016). A study on statistical methods used in six journals of library and information science. *Online Information Review, 40*(3), 416-434. doi: 10.1108/OIR-07-2015-0247

(Received: 2020/10/12; Accepted: 2021/1/29)

圖書資訊學刊　第19卷第1期　（2021.6）　頁1-17
doi: 10.6182/jlis.202106_19(1).001

# 資料探勘方法於圖書資訊學領域之運用

# Adoption of Data Mining Methods in the Discipline of Library and Information Science

**Marie Katsurai[1], Soohyung Joo[2]**

## 摘　要

本文探索2009至2018年，圖書資訊學領域研究運用資料探勘方法的趨勢。本研究自Scopus資料庫分別蒐集資料探勘領域和圖書資訊學領域之書目紀錄，並根據基於規則（rule-based）的文字分析法，建構資料探勘方法術語字典；藉由此字典，調查近期圖書資訊學研究中常見之各種資料探勘方法。研究結果發現，圖書資訊學領域運用多元資料探勘法，如大數據、機器學習、文字探勘、資訊檢索以及降維（dimension reduction）等；同時，本研究發現近期流行之機器學習技法（machine learning techniques）的確也被運用於圖書資訊學研究。

關鍵字：圖書資訊學、文字探勘、詞彙建構、書目計量分析、計算方法

[1] 日本京都同志社大學情報資訊工程科學系
Department of Intelligent Information Engineering and Sciences, Doshisha University, Kyoto, Japan
[2] 美國肯塔基大學資訊科學系
School of Information Science, University of Kentucky, Lexington, Kentucky, USA
* 通訊作者Corresponding Author: Soohyung Joo, Email: soohyung.joo@uky.edu

註：本中文摘要由圖書資訊學刊編輯提供。
以APA格式引用本文：Katsurai, M., & Joo, S. (2021). Adoption of data mining methods in the discipline of library and information science. *Journal of Library and Information Studies, 19*(1), 1-17. doi: 10.6182/jlis.202106_19(1).001
以Chicago格式引用本文：Marie Katsurai and Soohyung Joo. “Adoption of data mining methods in the discipline of library and information science.” *Journal of Library and Information Studies* 19, no. 1 (2021): 1-17. doi: 10.6182/jlis.202106_19(1).001