Tampere University

Kaisa Ylikruuvi

AUTOMATING THE DISCIPLINE ANALYSIS WITH LATENT DIRICHLET ALLOCATION

A Case Study on 30 Core Journals of Library and Information Science Published in 2015

Faculty of Information Technology and Communication Sciences
Master's Thesis
May 2023

ABSTRACT

Kaisa Ylikruuvi: Automating the Discipline Analysis with Latent Dirichlet Allocation: A Case Study on 30 Core Journals of Library and Information Science Published in 2015
Master's Thesis
Tampere University
Master's Degree Programme in Computational Big Data Analytics
May 2023

Discipline analysis is an interesting and important research area, especially in the interdisciplinary and multidisciplinary fields of science, such as library and information science (LIS). Discipline analysis helps to identify the current trends and evolution of the research topics and the main methodologies employed within a field of study. In this thesis, discipline analysis is conducted by building a topic model on library and information science articles. The latent Dirichlet allocation (LDA) algorithm is employed in the set of LIS articles, which has been previously classified intellectually by LIS researchers. The thesis aims to compare the LDA model to the result of the intellectual content analysis, previous LDA models of LIS, and the co-citation analysis model of the same data set.

The data consists of 1 440 articles and conference papers published in 30 core journals of LIS in 2015. The selection of journals, and the decision to use only titles, abstracts, and keywords in the analysis, are the same as in the intellectual content analysis. Most of the data could be fetched via Scopus API and the rest were downloaded from ProQuest or collected manually from the journals' homepages. The data preprocessing phase included the correction of errors caused by optical character recognition and XML encoding, the removal of platform-specific metadata, numbers, stopwords, and extra whitespaces, and lemmatization. The data were analysed in R with package topicmodels to perform latent Dirichlet allocation. The quality assessment values of perplexity and topic coherence were calculated with functions from packages topicmodels and topicdoc, respectively. The final LDA model consists of 14 topics: **Impact Indicators**, **Education in LIS Studies and Education as LIS Service**, **Academic Libraries**, **Information Retrieval**, **Computation-Assisted Analysis** (analysis method), **Scientific Collaboration**, **Public Libraries**, **Interactive Information Retrieval**, **Knowledge and Patent Management**, **Bibliometrics** (analysis method), **Open Access**, **Information History**, **Social Media**, and **User Behaviour in Digital Environment**.

The LDA model is of good quality and it succeeds to describe the different aspects of LIS well. The model compares well to the content analysis, which was conducted using the same data set, and to previous topic models of LIS. The LDA model outperforms the result of co-citation analysis, which was performed on the same data set, and which selects labels automatically for its clusters from the titles in the data. LDA topic modelling is a suitable method for pursuing discipline analysis. Further development is still recommended to automate the process more by developing a comprehensive preprocessing framework and especially by implementing high-quality automatic topic labelling for various platforms.

Keywords: Topic modelling, Latent Dirichlet allocation, Co-citation analysis, Discipline analysis, Library and information science, CiteSpace

Contents

1 Introduction

Discipline analysis or autoanalysis means conducting research on an academic field in order to find out what the essence of the field is and how it has evolved over time: what are the research topics and how their popularity has changed, and which research methods are used? Discipline analysis is very time-consuming when it is conducted traditionally by intellectual and manual content analysis, and it requires expertise in the field of science in question. The workload is also growing exponentially year by year. The growth rate of scientific publications depends on the chosen time frame and it varies among the fields of science. In a recent study, the growth rate of scientific publications was estimated to have been 5.08% between 1952 and 2018 with a doubling time of approximately 14 years (Bornmann, Haunschild, and Mutz 2021). There is an interest to automate the discipline analysis with computational methods.

The analysis method in this thesis is latent Dirichlet allocation (LDA), which is a topic modelling algorithm. Topic modelling is a machine learning technique, which can be employed to identify the latent topics of a text corpus, such as research topics of a discipline. Topic modelling is unsupervised, which makes serendipitous findings possible. It is also an approach, which is still evolving and its algorithms are being developed. There are many different algorithms for performing topic modelling of which latent Dirichlet allocation was chosen because of its popularity and generality.

In this thesis, LDA is utilized to construct a topic model that represents the research topics within the field of library and information science (LIS). There are previous studies, which have analysed the intellectual structure of library and information science by fitting a topic model on LIS papers. Most of these studies have used latent Dirichlet allocation as the topic modelling algorithm (e.g. Figuerola, García Marco, and Pinto 2017; Han 2020; Miyata et al. 2020) but also, for example, the author-topic model has been employed (Sugimoto et al. 2011). LIS journal articles have been the most popular genre of LIS papers as a data source, but there are differences in the scope of journal selection and the methods used for making those selections. Topic modelling has been performed also on North American doctoral LIS dissertations to broaden the view of the field by analysing the papers also from another scientific genre (Sugimoto et al. 2011).

The previous studies have aimed to investigate the evolution of LIS research topics and to identify current trends. The studies have also included the evaluation of the consistency of their results with other research and the general consensus within the history of LIS. However, there is no previous study, which compares a topic model of LIS to the result of manually conducted content analysis on the same data set. The objective of this thesis is to fill this gap by comparing the LDA topic model of LIS to the intellectual classification scheme in Järvelin and Vakkari (2021). The data consists of 1 440 articles and conference papers from 30 core LIS journals

published in 2015. Also, a co-citation model with automatic topic labelling is built as a comparative computational method for topic modelling.

The research questions of the thesis are:

1. How the result of the LDA model compares to the LIS research topics in Järvelin and Vakkari (2021)?

2. How the result of the LDA model compares to previous topic models of information science?

3. How the result of the LDA model compares to the co-citation model?

The structure of the thesis is as follows. Chapter 2 provides the theoretical background. It first introduces the analysis methods used in this thesis, followed by the literature review of discipline analysis on LIS. Chapter 3 describes the data, and implementation of the research. The findings and results of the analysis are presented in Chapter 4. Lastly, Chapter 5 presents the conclusions of the thesis.

2 Theoretical Background

The main theoretical background is introduced in this chapter. It begins with presenting the upper-level concept of text mining in Section 2.1. The essential text preprocessing step is presented in Subsection 2.1.1. Topic modelling, which is the specific text mining technique used in the analysis of this study, is described in Subsection 2.1.2. Co-citation analysis, which is a bibliographic method, is presented in Section 2.2. It is a comparative method used in this thesis as being a more traditional discipline analysis method. The literature review is in Section 2.3. The former studies regarding the discipline analysis of library and information science and specifically the studies, which are conducted with topic modelling, are presented in Subsections 2.3.1 and 2.3.2, respectively.

2.1 Text Mining

Text mining (also known as text data mining) is a form of data mining, which uses natural language texts as data. Text mining and natural language processing (NLP) are partly overlapping terms but their mutual hierarchy is often presented so that text mining uses natural language processing techniques and algorithms to reach its goal. According to Oxford Reference, text mining is "The automated process by which large volumes of unstructured, natural-language text are analysed in order to *pinpoint and extract user-specified information* [emphasis added]." ("Text Mining", 2022) and NLP is "The area of computer science that *develops systems that implement natural language understanding* [emphasis added]. It is a sub-discipline of artificial intelligence and of computational linguistics." ("Natural-Language Processing", 2022).

Important terms, which are used in the context of text data are token, document, and corpus. A token is a basic unit of textual data. In topic modelling, tokens are words or stems of words. Other applications may use, for example, characters, punctuation marks, subwords, or even whole sentences as tokens. A document is a sequence of N tokens denoted by $\mathbf{w} = (w_1, w_2, ..., w_n)$, where w_n is the nth word in the sequence. A corpus is a collection of M documents denoted by $D = (\mathbf{w}_1, \mathbf{w}_2, ..., \mathbf{w}_m)$, where $\mathbf{w}_m$ is the mth document of the corpus. (Blei, Ng, and Jordan 2003, p. 995.)

2.1.1 Text Preprocessing

Text preprocessing is an essential part of text mining and it has an enormous impact on the quality of the results. There are several methods, which can be used, and a suitable mixture and order of them have to be always chosen for the data and the task at hand. For that reason, there does not exist any universal preprocessing framework, which would be employed by all the researchers and experts in the field, but they share similarities. For example, Mayo (2017) divides preprocessing methods into three

categories of tokenization, normalization, and noise removal and Aggarwal (2018, pp. 17–18) divides them almost similarly into three main steps of platform-centric extraction and parsing, preprocessing of tokens, and normalization.

Tokenization means that the text is divided into tokens (~words). Tokenization is not as trivial as using whitespaces as token limits. The handling of, for example, apostrophes and hyphens has to be considered. Segmentation can be used as a term for dividing the text into clauses, sentences, or paragraphs.

Normalization of text data reduces the dimensionality of the data set and makes the resulting document term matrix less sparse. It covers several things: converting all letters to lowercase, stripping whitespaces, stemming or lemmatization, and removing numbers, punctuation, and stop words. There are readymade functions to perform all of these but especially stemming and lemmatization functions are very helpful because their algorithms are not as simple as those of others.

Stemming means removing suffixes (and sometimes also prefixes) from stems which are the roots or the main parts of words. Snowball stemming algorithm is the most prevalent stemming algorithm. It was previously called Porter2 because it is built on Porter algorithm, which was the previous most popular stemming algorithm. Snowball (and Porter) have five steps of simple rules, which are performed consecutively. For example, if the word ends with -ily, the ending is replaced with -ili but suffix -ly is removed. Lemmatization is more advanced than stemming. In lemmatization, the lemma, which is the word's dictionary form, is returned and as a result, all the different inflected forms of a word can be analysed as one.

Stop words are words, which are considered uninformative in text analysis. Stop words are frequent but carry little information to differentiate between the meaning of documents, such as pronouns, articles, and prepositions. There are readymade lists that can be used, for example, stop word list of 147 words from Snowball project.

HTML and JSON formatting, systematic errors, and typos are examples of noise in text data. Noise removal is more data and task dependent than methods in the other two categories and it is usually handled with regular expressions by removing or replacing strings. As an example, removable noise in the data set of this thesis is copyright texts in the abstracts and abstract formatting words (Purpose, Design, Findings...), and correctable noise is due to systematic optical character recognition errors and XML encoding errors. Unique typos are usually not systemically looked for and are left uncorrected.

2.1.2 Topic Modelling

Topic modelling is a group of unsupervised machine learning techniques with the exception of some supervised methods[1], which aim is to find the latent topics of a text corpus given k, which is the prefined number of topics. The main idea that is shared by all topic modelling techniques is that each topic is a distribution over words and each document is a distribution over topics. The task is to create k topics and to

[1] e.g., supervised latent Dirichlet allocation (Blei and McAuliffe 2007)

find the distribution of topics that documents belong to based on their words. The model is optimized by choosing a good k and other parameters for the algorithm.

There are several methods, which have been created to reach the goal of finding the latent topics. The most prevalent basic method for fitting a topic model is latent Dirichlet allocation (Blei, Ng, and Jordan 2003). According to its creators, "Latent Dirichlet allocation is a generative probabilistic model of a corpus" (Blei, Ng, and Jordan 2003, p. 996). LDA is used also in this thesis because of its popularity and because there exist publicly available implementations for it. It is a suitable method to use with this data set and for this purpose also according to Vayansky and Kumar (2020, p. 14) who have built a decision tree for choosing a method for topic modelling. Their decision tree leads to LDA after choosing "Average number of words per document ≥ 50: Yes" and "Complex topic relationships are of interest: No".

LDA consists of three steps for each document $\mathbf{w}$ in a corpus D and the following algorithm is its implementation from "topicmodels: An R Package for Fitting Topic Models" (Grün and Hornik 2011, pp. 3–4), which is employed in this thesis:

1. Determine the token distribution for each topic by

$$\beta \sim \text{Dirichlet}(\delta).$$

2. Determine the proportions θ of the topic distribution for the document by

$$\theta \sim \text{Dirichlet}(\alpha).$$

3. For each w_i of N tokens of a document

   (a) Choose a topic $z_i \sim \text{Multinomial}(\theta)$.

   (b) Choose a token w_i from a multinomial probability distribution conditioned on the topic z_i : $p(w_i|z_i, \beta)$, where β is the term distribution of topics.

The Dirichlet distribution (2.1) is a multivariate generalization of the beta distribution. The Dirichlet distribution is also the conjugate prior for the multinomial distribution (2.2) in Bayesian statistics. The multinomial distribution is a generalization of the binomial distribution. The probability mass function of the multinomial distribution can be expressed also by using the gamma function (2.3) and it shows a resemblance with the Dirichlet function.

$$\begin{aligned} f(x_1, x_2, \ldots, x_k) &= \frac{\Gamma(\alpha_1 + \alpha_2 + \cdots + \alpha_k)}{\Gamma(\alpha_1)\Gamma(\alpha_2)\cdots\Gamma(\alpha_k)} x_1^{\alpha_1 - 1} x_2^{\alpha_2 - 1} \cdots x_k^{\alpha_k - 1} \\ &= \frac{\Gamma(\sum_{i=1}^{k} \alpha_i)}{\prod_{i=1}^{k} \Gamma(\alpha_i)} \prod_{i=1}^{k} x_i^{\alpha_i - 1} \end{aligned} \tag{2.1}$$

$$
\begin{aligned}
f(x_1, x_2, \ldots, x_k) &= P\{X_1 = x_1, X_2 = x_2, \ldots, X_k = x_k\} \\
&= \frac{n!}{x_1! x_2! \cdots x_k!} p_1^{x_1} p_2^{x_2} \cdots p_k^{x_k} \\
&= \frac{n!}{\prod_{i=1}^{k} x_i!} \prod_{i=1}^{k} p_i^{x_1}, \text{ when } x_1 + x_2 + \cdots + x_k = n.
\end{aligned} \tag{2.2}
$$

$$
f(x_1, x_2, \ldots, x_k) = \frac{\Gamma(\sum_{i=1}^{k} x_i + 1)}{\prod_{i=1}^{k} \Gamma(x_i + 1)} \prod_{i=1}^{k} p_i^{x_i}. \tag{2.3}
$$

The resulting topic model represents documents as random mixtures over latent topics. A document may be presented as an example document of the topic, which it consists most of, but it is important to keep in mind that the document is still a mixture of topics. The topic model represents latent topics as the lists of their top n tokens and labelling of the topic is left for the user, possibly with the help of experts in the field. Often the ten most frequent tokens are used to describe and label the topics but it is also possible to use shorter or longer lists of top tokens.

Previous popular topic modelling methods have been latent semantic indexing (LSI) (Deerwester et al. 1990) and probabilistic latent semantic analysis (PLSA) (Hofmann 1999). LSI (or latent semantic analysis, LSA, outside the field of information retrieval) was developed to improve the traditional word matching information retrieval results, which are weakened by, for example, synonymy and polysemy issues. LSI uses singular value decomposition to build a semantic space where associative documents and terms are close to each other, query terms are used to find a point in the space, and the documents near that point are returned. (Deerwester et al. 1990, p. 392.) PLSA (or probabilistic latent semantic indexing, PLSI, in information retrieval) was developed to be a probabilistic variant of LSI. It is based on a mixture decomposition derived from a latent class model instead of linear singular value decomposition, and compared to LSI it has a sounder foundation in statistics and it defines a proper generative model of the data. (Hofmann 1999, p. 289.) The weaknesses of PLSA are that it does not generalize to new documents outside the training set and that it is prone to overfitting because the number of the model's parameters grows linearly as the size of the corpus grows. LDA does not suffer from these: it succeeds to easily generalize to unseen documents and it does not have a similar risk for overfitting as PLSA. (Blei, Ng, and Jordan 2003, p. 100.)

Researchers have improved and extended the basic LDA in numerous ways. Next, a few examples are presented. Blei himself has continued to work with LDA and created correlated topic model (CTM) for highly correlated data (Blei and Lafferty 2006a), and dynamic topic model (DTM) for long term sequential data (Blei and Lafferty 2006b). Compared to LDA, CTM uses the logistic normal distribution instead of Dirichlet distribution to draw the topic mixture proportions. The former distribution allows the latent topics to be correlated with each other while the latter

assumes that the topics are independent and the probability of their presence is not correlated. (Blei and Lafferty 2006a, p. 148.) DTM divides the documents into time slices and takes into account that the topics of a time slice evolve from the topics of the previous time slice. CTM on the other hand does not use the information about the order of the documents but assumes that the documents are drawn exchangeably from the same set of topics. (Blei and Lafferty 2006b, p. 114.)

Li and McCallum (2006) created Pachinko allocation model (PAM) for data with complex structural relationships. The limitation of CTM is that it allows only pairwise topical correlations but PAM is able to represent and learn arbitrary arity, nested, and sparse topic correlations. The structure of PAM is more complex than that of many other topic models because its topics are distributions not only over tokens but also over other topics. (Li and McCallum 2006, p. 578.) Mixture of unigrams was introduced before LDA but it is still valid and used for short texts. When the documents are very short, such as SMS messages or Google Maps reviews, it is rare that they consist of the distribution of topics but they refer to one topic only. (Nigam et al. 2000.)

Several hierarchical classification systems for different topic modelling algorithms have been created but none of them is exhaustive and they lag behind as new algorithms are presented. Possibly the most recent classification is from Chauhan and Shah (2021, pp. 7–21), whose upper level division of topic modelling algorithms is *plain topic models*, *hierarchical topic models*, *multilingual topic models*, *topic models in distributed environment*, *topic model with prior embedded information*, *topic modeling for short text*, and *modeling topics over non-textual data*. In Vayansky and Kumar (2020, pp. 3–10), the upper level of the hierarchy consists of *basic approaches*, *topic models with advanced topic relationships*, *time-based topic models*, *short text optimized topic models*, and *other significant topic model designs*.

Topic modelling has been used as a tool for working with different kinds of data sets and tasks. According to Chauhan and Shah (2021, p. 25), the four most popular domains that comprise together two-thirds of the papers employing topic modelling are research papers (21%), news articles (17%), Wikipedia (15%), and microblogs (14%). The remaining domains are software, legal records, reviews, images, video, audio, and "other", each of which having a share of 3–6%. Topic modelling has been used for the tasks of text mining, image retrieval, social network analysis, opinion/sentimental analysis, geo-referencing, incident/anomaly detection, and analysing bioinformatics, social, environmental, health, and education data (Mulunda, Wagacha, and Muchemi 2018; Kherwa and Bansal 2019; Vayansky and Kumar 2020).

Griffiths and Steyvers (2004) have extracted scientific topics from the corpus consisting of Proceedings of the National Academy of Sciences of the United States of America (PNAS) abstracts from 1991 to 2001. Layman et al. (2016) applied topic modelling on a corpus of NASA space system problem reports to extract problem trends within and across different space missions. In this thesis, the aim of topic modelling is to perform a disciplinary analysis in the field of library and information science. There have been some similar kinds of studies earlier with different kinds

of LIS related document types and they will be presented in Subsection 2.3.2.

Inference Techniques for LDA The main problem in all the topic modelling methods is how to learn the posterior distributions for the latent variables from the data. Latent variables represent underlying concepts that are not directly observable. In the context of topic modelling, latent variables are the latent topics that are inferred from the data. Posterior inference is intractable to be done by computing but it can be approximated. Two main inference techniques to approximate the posterior distribution for latent Dirichlet allocation are variational expectation maximization (VEM) (Blei, Ng, and Jordan 2003) which was employed in the original LDA algorithm paper and Gibbs sampling (Darling 2011; Heinrich 2005) which was presented as an alternative to VEM in Grün and Hornik (2011). Both of these techniques are used in this thesis.

VEM is a deterministic algorithm. It tends to get stuck at a local optimum instead of reaching the global one when converging. VEM is faster than Gibbs sampling. Gibbs sampling is a stochastic Markov chain Monte Carlo algorithm. It finds the global optimum but the convergence reaches only an approximation of the posterior distribution because the number of runs is limited. Gibbs sampling is better for small data sets than VEM.

Evaluation of the Quality of the Model Topic modelling is usually a form of unsupervised learning. Because there is no annotated document test set to use as a golden standard nor a ready-made list of topics ideally to be found from the corpus, and because topic modelling is not about predicting one single topic for documents, the classification evaluation metrics such as accuracy or F1 score cannot be used. Instead, the quality of the topic model can be assessed by checking how well the modelled topics perform as input to certain computational tasks, by human evaluation, or by using metrics that are suitable for the purpose. Topic models have been evaluated by measuring their performance on measurable tasks, such as document classification, information retrieval, and sentiment analysis (Wallach et al. 2009, p. 1105; Boyd-Graber, Mimno, and Newman 2014, p. 237). Evaluation is more challenging when the task is qualitative, such as exploring semantic themes, as in this thesis.

Chang, Gerrish, et al. (2009) proposed two tasks for humans to perform in order to evaluate how semantically coherent the topics are and how well the mixture of topics assigned to each document associates with the document. In the word intrusion task, one random word is added to the list of the top 5 words of a topic, and the topic is seen as coherent if people can choose the intruder word. In the topic intrusion task, the title of the document and a snippet from it along with the word lists of its three top topics and one random low-probability topic word list are presented. The quality of document-topic assignments is good if people can choose the intruder topic. (Chang, Gerrish, et al. 2009, pp. 3–4.)

The topic evaluation by Chang, Gerrish, et al. (2009) proved to work very well but it is very time and money consuming and thus not applicable in many cases. In this thesis, two traditional quantitative metrics are used to optimize the parameters and to find a good topic model: perplexity and topic coherence.

The perplexity or held out (log-)likelihood measures how well the model can predict new data. The data set is divided into training set and test set. The parameters are optimized and the model is built with the training data set and the model is then tested with the held out test data set. The lower the perplexity, the less perplexed the model is by the test data. Grün and Hornik (2011, p. 7) define mathematically perplexity as being equivalent to the geometric mean per-word likelihood.

$$\text{Perplexity}(w) = \exp\left\{ -\frac{\log(p(w))}{\sum_{d=1}^{D}\sum_{j=1}^{V} n^{(jd)}} \right\}$$

$$\log(p(w)) = \sum_{d=1}^{D}\sum_{j=1}^{V} n^{(jd)} \log\left[\sum_{K=1}^{k} \theta_K^{(d)} \beta_K^{(j)} \right]$$

$$\text{Perplexity}(w) = \exp\left\{ -\frac{\sum_{d=1}^{D}\sum_{j=1}^{V} n^{(jd)} \log\left[\sum_{K=1}^{k} \theta_K^{(d)} \beta_K^{(j)} \right]}{\sum_{d=1}^{D}\sum_{j=1}^{V} n^{(jd)}} \right\},$$

where D denotes the corpus, V denotes the vocabulary, $n^{(jd)}$ denotes how often the jth term occurred in the dth document, K denotes the topic, $\theta_K^{(d)}$ denotes the topic weight and $\beta_K^{(j)}$ denotes the token weight. (Grün and Hornik 2011, pp. 7–8.)

Topic coherence measures the conditional likelihood of the co-occurrence of tokens within a topic. There are several variations for the formula and the one, which is used by Friedman (2019, pp. 7–8) and thus also in this thesis, is from Mimno et al. (2011, p. 265). It measures how often the n (the default is 10) top words of a topic co-occur together. The log of the probability that a document contains both words w_j and w_i is calculated for each possible word pair in top n words and these values are added up to present the topic coherence for the topic in question. For each topic we calculate

$$\sum_{i}\sum_{j<i} \log \frac{D(w_j, w_i) + \beta}{D(w_i)},$$

where $D(w_i)$ is the number of documents that contain at least one w_i, and $D(w_j, w_i)$ is the number of documents that contain at least one w_j and one w_i. β parameter (the default is 1) is added to avoid logarithm zero errors. Topic coherence scores are negative because they are logarithm probabilities. More coherent the topic is, more closer to zero is the topic coherence score. The mean of topic coherence scores describes how coherent the model is on average.

2.2 Co-Citation Analysis

Citation analysis methods are used to describe the citations within a collection of documents with, for example, citation counts and graphs, and they belong to the wider group of bibliometric methods. Bibliometrics studies printed publications in a quantitative manner, especially in scientific communication. It examines for example the most cited publications, the number of scientific publications, and their mutual references.

Co-citation analysis is one of the citation analysis methods. Two documents have a co-citation relationship when they are cited by the same document. Other citation analysis methods are direct citation and bibliographic coupling. Two documents are bibliographically coupled when both of them cite a third document. (Chang, Huang, and Lin 2015, p. 2072.)

2.3 Discipline Analysis

Discipline analysis has been popular, especially in the field of library and information science. The roots of LIS are in library science and librarianship. It has grown to include also other aspects of information collection and dissemination and the focus has shifted from libraries to information (Figuerola, García Marco, and Pinto 2017, p. 1508). The popularity of autoanalysis in LIS is partly related to the discussion of the relationship of library science and information science and especially the role of library science in LIS as the proportion of research on library related topics has diminished (Sugimoto et al. 2011, p. 185; Tuomaala, Järvelin, and Vakkari 2014, p. 1446; Onyancha 2018, p. 467). Partly it may be explained by the fact that the bibliometric methods, which are often used in discipline analysis, are developed by LIS.

Two traditional research strategies, which are used in autoanalysing LIS, are content analysis and bibliometric methods. Model-based approaches, such as topic modelling, are new and still evolving. Content analysis in this context means that the research papers are intellectually and manually categorized along different features (e.g. the research topic or the data collection method) so that the analysis can be performed quantitatively. There are several bibliometric research methods of which citation analysis is the most popular subgroup of methods when the interest is in the knowledge structure and the evolution of research topics of a discipline (Hou, Yang, and Chen 2018, p. 870). Direct citation has been the most popular citation analysis method (Chang, Huang, and Lin 2015, p. 2072).

Some papers on autoanalysis of LIS are presented in subsection 2.3.1, and the ones that employ topic modelling are presented separately in subsection 2.3.2.

2.3.1 Discipline Analysis in Library and Information Science

Content Analysis Järvelin and Vakkari (1990, 1993, 2021) have done groundbreaking work in the content analysis side of the field. Their first study in the series created a taxonomy for classifying the topic, the approach, and the research method of articles along 7 variables. The data consisted of 833 research and professional articles published in 1985 in 37 English and German core journals in LIS. Primarily only the abstracts of the articles were used in the classification. The introduction was used in the absence of an abstract and an adequate number of first pages if there was no introduction either. The most popular main topics in the research articles were *Information storage & retrieval* (29.2%) and *Library & information service activities* (27.2%). All other research fields of LIS had a share of less than 10%.

A few years later they continued and broadened the analysis to describe the evolution of LIS research through the years 1965 (data from Huusko, 1992, cited in Järvelin and Vakkari 1993), 1975 (data from Kumpulainen, 1991, cited in Järvelin and Vakkari 1993) and 1985. Differing from their earlier study, Järvelin and Vakkari analysed only research articles and excluded professional articles. The data consisted of 142, 359, and 449 research article abstracts respectively from altogether 40 English and German journals but none of the yearly data sets included articles from all of the journals. The main finding was that the foci of LIS had not changed much from 1965 to 1985. The most popular main research topics in 1985 had been the most popular throughout the whole time period: *Information storage & retrieval* (26.2%–32.4%) and *Library & information service activities* (25.4%–27.2%).

In 2014, Tuomaala revised the 1993 study with Järvelin and Vakkari. First, they analysed LIS articles published in 2005 and then they examined the evolution of library and information science research field through the years 1965, 1985, and 2005. The German sources were left out and the year 2005 data consisted of 718 English research article abstracts from 29 core LIS journals. Some updates to the topic classification system were made to better describe the status and the development of the field. For example, topic subclasses of *Digital information resources*, *Interactive information retrieval* and *Webometrics* were added. In 2005 *Information storage & retrieval* (30.1%) was still the most popular research topic but *Scientific and professional communication* (24.3%) had reached the second position ahead of *Library & information service activities* (17.0%). *Information seeking* (12.3%) was the fourth big main class.

Latest in the series, Järvelin and Vakkari (2021) added the content analysis of the journal articles of the year 2015 to their and Tuomaala's earlier studies and thus covered fifty years of library and information research. The new data consisted of 1 210 journal abstracts from 33 English LIS journals. In 2015 the order of the most popular research topics changed compared to 2005 with *Scientific and professional communication* (37.4%) leaving behind *Information storage & retrieval* (22.9%), and *Library & information service activities* (13.9%) and *Information seeking* (13.9%) sharing the third place. The most popular sub-topics were *Other*

aspects of communication (S&PC, 14.0%), *Scientific or professional publishing* (S&PC, 12.9%), *Citation patterns and structures* (S&PC, 7.6%), *Digital information resources* (IS&R, 5.0%), and *Classification and indexing* (IS&R, 4.0%).

Bibliographic method Chen, Ibekwe-SanJuan, and Hou (2010), Hou, Yang, and Chen (2018) and Li, Yang, and Wang (2019) used CiteSpace application to perform document co-citation analysis, which is a bibliographic method. Chen, Ibekwe-SanJuan, and Hou (2010) used 10 853 articles from 12 journals published in 1996–2008 as they introduced the new version of their multiple-perspective co-citation analysis tool, CiteSpace. It combines network visualization, spectral clustering, automatic cluster labelling, and text summarization. The resulting five largest document co-citation clusters were *Interactive information retrieval*, *Academic Web*, *Information retrieval*, *Citation behavior*, and *H-index*.

In 2018, Hou, Yang, and Chen continued with the previous methods and analysed 7 574 articles from 10 LIS journals published in 2009–2016. The articles included 20 960 references. They noticed clear changes in the most popular research topics in LIS between the time periods of 1996–2008 and 2009–2016. The seven largest clusters in the 2nd period were *Triple helix*, *Hirsch index*, *Citation performance*, *Citation count*, *Intellectual structure*, *Bibliometric analysis*, and *Information behavior*. Information retrieval, webometrics, and citation behaviour from the 1st period had thus been replaced by scientometric indicators, citation analysis, scientific collaboration, and information behaviour in the 2nd period. Another finding in the study was that using CiteSpace in automatic topic labelling reduces subjectivity compared to manual labelling.

Li, Yang, and Wang (2019) had 88 304 papers published in 1989–2018 in their data set. The papers were fetched from Web of Science and they had 1 445 168 references. The clustering label words were taken from the titles of the cited documents. Based on their analysis, they predicted that from the prevalent research topics *Social media, Information system,* and *Scientific evaluation* will be widely researched also in the future but the research on traditional *Information retrieval, Information behaviour, Bibliometrics and webometrics*, and *Knowledge management* is diminishing as three new social media influenced theme areas *Metrology*, *Open government* and *Big data* are growing in popularity.

Model-based approach Liu et al. (2015) proposed a formal concept analysis (FCA) based method for identifying the intellectual structure of LIS. FCA generates and visualizes the hierarchical concept structure of data sets. Their final data set for analysis consisted of 60 core author names and 99 standardized high frequency keywords from articles of 16 LIS journals published in 2001–2013. As a visual result of the analysis, the keywords and the authors formed a concept lattice. The concept lattice was further analysed and nine main topics (keywords) as well as the relationships between the topics, between the scholars, and between the topics and the scholars were found. The main topics were *Bibliometrics, scientometrics,*

and informetrics, *Citation analysis*, *Information retrieval*, *Information behavior*, *Libraries*, *User studies*, *Social network analysis*, *Information visualization*, and *Webometrics*.

Research methods In addition to analysing the research topics in the field of library and information science, there are also studies that have focused on the research methods that have been employed. Aforementioned Järvelin and Vakkari's (1990, 1993, 2021) studies and also Tuomaala, Järvelin, and Vakkari's (2014) included research methods along with several other features of research in addition to the research topics. Their main findings were that the LIS researchers were utilizing predominantly quantitative research methods and that there was not much variation in the methods that were chosen. Ma and Lund's (2020) quite recent data from years 2006, 2012, and 2018 with only six years' increments showed a modest increase in the use of qualitative methods as they replicated Järvelin and Vakkari's (1990, 1993, 2021) and Tuomaala, Järvelin, and Vakkari's (2014) studies.

Chu's (2015) study focused solely on the research method selection and application. Her data was from three LIS journals published in 2001–2010 and it showed that the LIS researchers were choosing the research method from a more broad selection than before and that the previously popular questionnaire survey and historical method were losing their dominant position to content analysis, experiment, and theoretical approach. Ullah and Ameen (2018) conducted a meta-analysis and combined the results of 58 studies on methods and methodology in LIS published in 1980–2016. Empirical, descriptive, and quantitative were the most popular methodologies used and survey was the most dominant research method. They also noticed that it would be important for the whole research and higher education community to define the terms and taxonomies which are used to describe the various aspects of the research methods and methodology. The lack of these standardized definitions resulted to inconsistencies, overlapping and ambiguity in the terms that were used in their data.

2.3.2 Discipline Analysis in Library and Information Science Using Topic Modelling

Sugimoto et al. (2011) were the first ones utilizing topic modelling as a research method for the evaluation of library and information science. More precisely they applied the author-topic model which extends the latent Dirichlet allocation. The data consisted of 3 121 titles and abstracts from doctoral dissertations completed between 1930 and 2009 at North American LIS programs. The data was divided into five time periods and for each time period, five topics were formed based on 20 top words. The most recent data set consisted of 766 dissertations from the years 2000–2009 and it resulted to topics of *Information use*, *Internet*, *Information seeking behavior*, *Information retrieval/user-centered*, and *Information retrieval/classification*. The topics were compared to other analyses, which were conducted using the data of

the same time periods. In these other analyses the data usually consisted of journal articles and the applied method was content analysis or a bibliographic method. There were some discrepancies between the results acquired by topic modelling and more traditional methods but the conclusion was that topic modelling is a suitable method for finding latent topics of LIS data.

Figuerola, García Marco, and Pinto (2017) based their research on Sugimoto et al.'s (2011) and used latent Dirichlet allocation to quantitatively identify the main topics and categories of research in library and information science. Their data included the peer-reviewed documents from LISA (Library and Information Science Abstracts) database published in 1978–2014. The corpus consisted of 92 705 documents and they resulted to 19 topics: *LIS profession and education*, *LIS and social development*, *Information behaviour*, *Legal and ethical aspects of LIS*, *Document preservation*, *Communication networks*, *Advanced statistics applications*, *Automatic information processing*, *Online search services*, *Library management*, *Reference services*, *Cataloging and library co-operation*, *Historical sources*, *Informetrics*, *Health information*, *Media communication*, *Education and learning*, *Business management*, and *Knowledge management*.

Han (2020) performed recently an autoanalysis of LIS using topic modelling. They applied LDA on 14 035 journal articles, which were published in 1996–2019. The time period was divided into five smaller time periods and the nine most influential journals for each time period were chosen with an advanced data selection method. The topic number was set to 10 for each time period. The corpus of the time period 2011–2015 had 3 840 papers and the latent topics found were *Bibliometrics analysis*, *Research performance*, *Citation analysis/measurement*, *Scientific collaboration*, *Citation analysis/impact factor*, *Information management*, *Government*, *Online/community*, *Organizational information activities*, and *Ranking research*.

Miyata et al. (2020) applied LDA to identify the topics of LIS from two time periods: 2000–2002 and 2015–2017. The 15 years break between the time periods allowed them to analyse the development and change of the field also. Their data was 1 648 full texts from five core LIS journals and the predefined k as the number of topics was 30. The data set for 2015–2017 had 1 087 articles. Five categories were formed from the topics: *Information Retrieval*, *Information Search and User*, *Library*, *Scholarly Communication*, and *Tweet Analysis*. Categorization was done by placing the topics on a 2D map and combining topics into categories based on their location and characteristics.

3 Implementation of the Research

The research procedure is described in this chapter. In Section 3.1, the data, its collection, and the preprocessing steps before the topic modelling are described. Then, in Section 3.2, the analysis workflow of the basic topic modelling with latent Dirichlet allocation is documented, including the optimization choices that were made. The data collection and preprocessing phases for co-citation analysis with CiteSpace are very simple and they are included in Section 3.3 which describes the co-citation analysis.

3.1 Data

The data consists of 1 440 documents in the field of library and information science that were published in 2015. The initial selection of 31 core library and information science journals (Appendix) is the same as in the previous research of Tuomaala, Järvelin, and Vakkari (2014) and Järvelin and Vakkari (2021). The publication year of the documents is the same as the most recent year in Järvelin and Vakkari (2021). The whole data collection and preprocessing workflow is presented in Figure 3.1.

3.1.1 Data Collection

The data for this study was collected from four different sources. The primary source was Scopus. Scopus API (Application Programming Interface) was used to fetch the documents of 28 journals to R as a data frame. The documents of *Journal of Education for Library and Information Science* and *Information & Culture* were downloaded from ProQuest as .xls files. The documents of *Indexer* were collected from the journal's homepage manually to an .xls file. Also the documents of the issue 35(4) of *Information Services and Use* were collected the same way because the issue in question is not available in Scopus.

All the documents of the available journals were first retrieved from Scopus. The documents that are not classified as articles or conference papers by Scopus were discarded. The document type *Article* in Scopus is defined as an "Original research or opinion." and *Conference paper* is an "Original article reporting data presented at a conference or symposium." The aim was to include only the research articles as was done also in the study of Järvelin and Vakkari (2021) and this filtering gave the closest equivalent to their intellectually selected data set. Also, the documents that do not have an abstract were removed from the data set.

Next, the documents of two journals were retrieved from ProQuest one journal at a time. The resulting two .xls files were merged into one file and only the documents that were used in the research of Järvelin and Vakkari (2021) and that have an abstract were kept. The number of documents was such small that the checking was manually

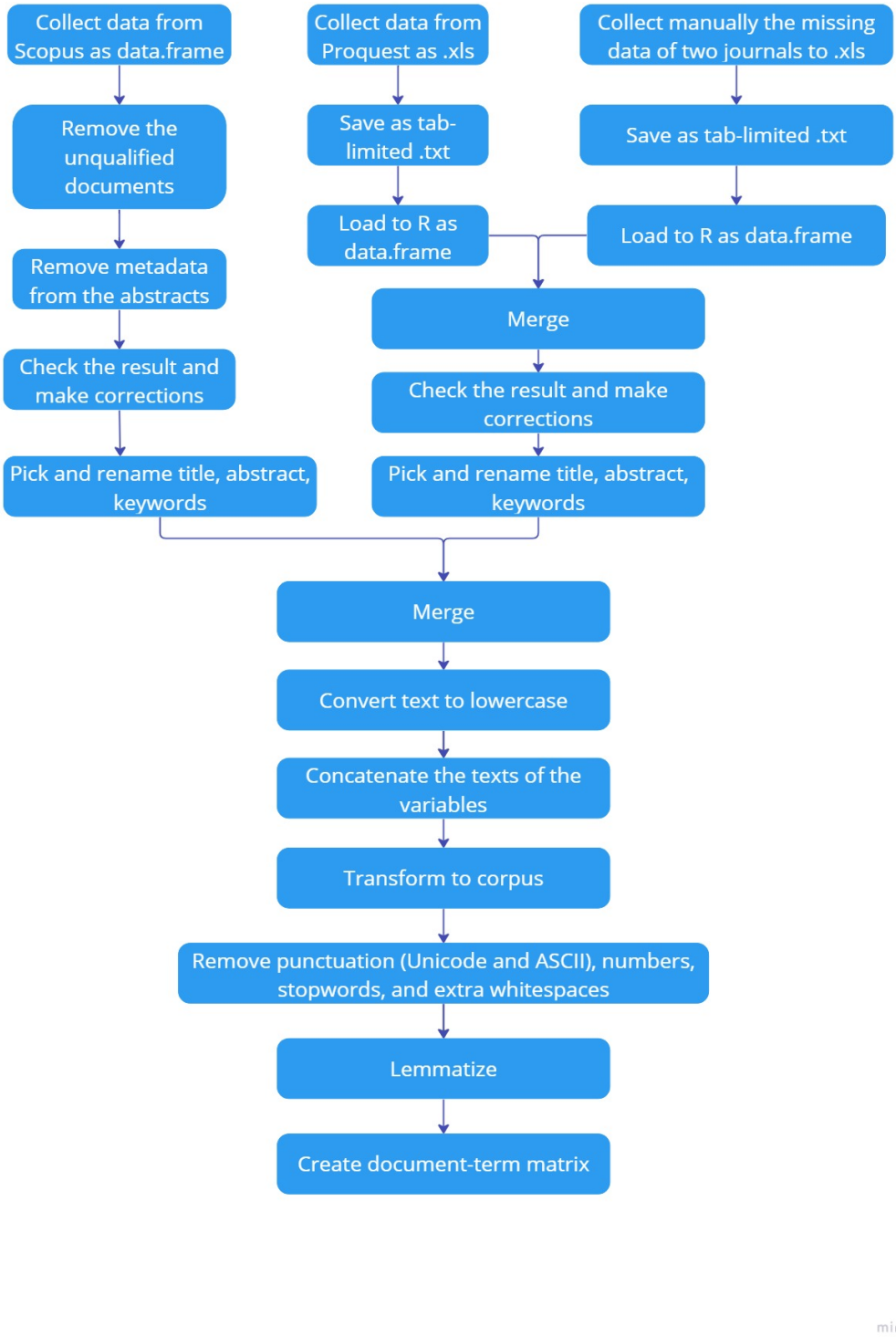


Figure 3.1. The data collection and preprocessing flowchart.

doable. The resulting file was saved as a tab-limited .txt file and loaded to R as a data frame.

The documents of the remaining one journal and one issue were manually collected and the variables were named after the variable names used in ProQuest. Again, the documents that do not have an abstract or had not been qualified to the data set of Järvelin and Vakkari (2021) were discarded. The result was saved as a tab-limited .txt file, loaded to R as a data frame, and merged into the data frame from the previous step.

3.1.2 Description of the Data

After collecting the data and removing the documents that were not research articles or did not have an abstract, there were 1 440 documents left from 30 journals. There were no documents from *Reference and User Services Quarterly* that could be qualified for the final data set. None of its documents in 2015 is classified as *Article*. There are two *Conference paper*s but they do not have an abstract.

The number of documents from each journal in the data set varies between 9 and 70 for other journals but there are two outliers. There are 341 (~23.7%) documents from *Scientometrics* and 171 (~11.9%) documents from *Journal of the Association for Information Science and Technology*. The mean for the number of documents per journal is ~46.5 and the median is 27. The keywords are missing from 247 (~17%) documents. The number of documents from each journal is presented in Figure 3.2.

Below, there is an example of the three features of a document that were used in the analysis after preprocessing. This document is from *Journal of Documentation* 71(6) and its preprocessed version is presented on pages 22–23.

dc:title
Systematic and serendipitous discoveries: a shift in sensemaking

dc:description
© 2015, Emerald Group Publishing Limited. Purpose – The purpose of this paper is to enrich the theoretical understanding of the phenomenon of sensemaking where a conceptual shift was provoked by a serendipitous encounter. Design/methodology/approach – A theoretical framework consisting of three elements of reflexivity: the cognitive, the social, and the normative, all of which support the study. Semi-structured interviews were conducted in the investigation of a serendipitous Episode that occurred in a larger research project. This Episode took place at a meeting between a social welfare officer and a psychologist in which they discussed the treatment of a psychiatric patient. When the psychologist left the meeting for a brief period, the researchers, unexpectedly, were able to interview the social welfare officer alone. Findings – This interview revealed a deviation from the institutionalized patient treatment procedure that was explained to the researchers in earlier interviews. The study shows that shifts in sensemaking are possible when

researchers are open to serendipitous encounters. This shift in sensemaking in this Episode was strategic because it concerned the three most important aspects of the actor's decision making: how to make diagnosis, treatments, and cooperate around the patient. Research limitations/implications – It is recommended that researchers use the theoretical framework of reflexivity to test their sensemaking processes as well as remain open to changes in planned, traditional methodological approaches. Originality/value – The study applies a post-hoc analysis with reflections on serendipitous events that may guide researchers when they encounter unanticipated events and make anomalous discoveries.

authkeywords
Abduction | Qualitative research | Sensemaking process | Serendipitous | Serendipity

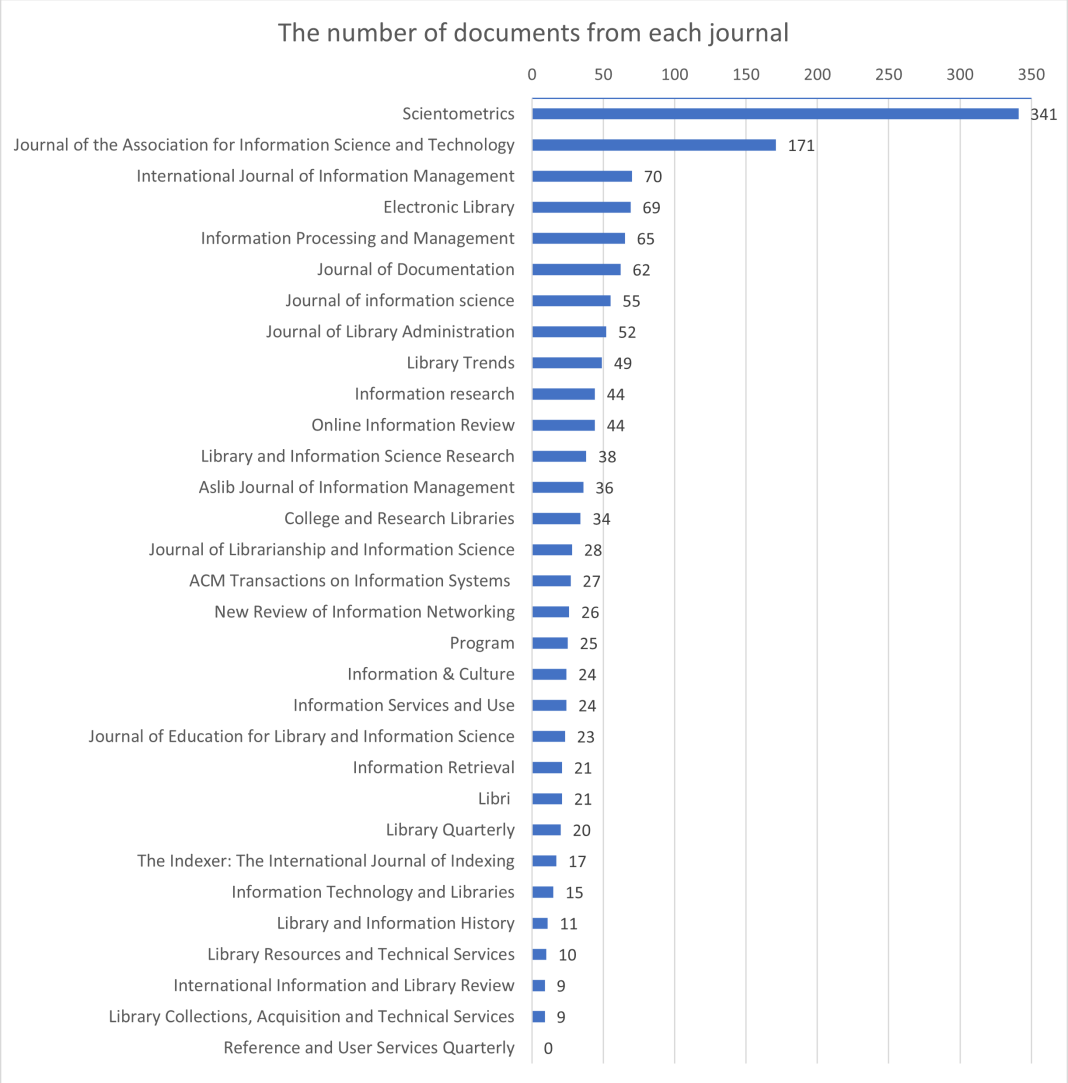


Figure 3.2. The number of documents included in the final data set from each journal.

3.1.3 Preprocessing the Data

The abstract texts from Scopus needed to be cleaned from platform-specific metadata content and it was done by using regular expressions because the removable words varied from journal to journal. For example, copyright texts and abstract formatting words used by journals such as *Purpose* and *Results* were removed.

Minor corrections were made to the documents that were collected from ProQuest. A few errors in the abstracts were due to the Optical Character Recognition errors in ProQuest. There had been also a couple of XML encoding errors in the titles when the data was imported from ProQuest to Excel. There was no need to make any corrections or cleaning to the data that was manually collected.

After cleaning and correcting, the three main variables *dc:title*, *dc:description* and *authkeywords* from the R data frame that was originated from Scopus were picked to form a new data frame, and the variables were renamed as *title*, *abstract* and *keywords*. The three variables *Title*, *Abstract* and *subjectTerms* from the R data frame having the rest of the data set were also picked to a new data frame and renamed to match the primary data set.

At this point, the two data sets with three variables were merged into one data frame. The final preprocessing steps for the data frame included converting text to lower case and concatenating the three variable fields into one. Then the data frame was transformed into a corpus. The data in the corpus was further preprocessed: punctuation (both ASCII and Unicode), numbers, stop words (Snowball project, 174 words), and extra whitespaces were removed and the resulting tokens were lemmatized (Měchura's English Lemmatization Listas dictionary). Finally, the document-term matrix was formed.

Below, there is the preprocessed version of the example document from *Journal of Documentation*. The original version was presented on pages 20–21.

> systematic serendipitous discovery shift sensemaking purpose paper enrich theoretical understand phenomenon sensemaking conceptual shift provoke serendipitous encounter theoretical framework consist three element reflexivity cognitive social normative support study semistructured interview conduct investigation serendipitous episode occur large research project episode take place meet social welfare officer psychologist discuss treatment psychiatric patient psychologist leave meet brief period researcher unexpectedly able interview social welfare officer alone interview reveal deviation institutionalize patient treatment procedure explain researcher early interview study show shift sensemaking possible researcher open serendipitous encounter shift sensemaking episode strategic concern three important aspect actor decision make make diagnosis treatment cooperate around patient recommend researcher use theoretical framework reflexivity test sensemaking process good remain open change plan traditional methodological approach study apply posthoc analysis reflection serendipitous event may guide researcher encounter unanticipated event make anomalous discovery abduction qualita-

tive research sensemaking process serendipitous serendipity

3.2 Latent Dirichlet Allocation

All analyses were performed using statistical software R (v4.0.5; R Core Team 2020). The package topicmodels (v0.2-11; Grün and Hornik 2011) was used to perform the topic modelling and to calculate the perplexities of models and the package topicdoc (v0.1.0; Friedman 2019) was used to calculate the coherences of topics. The code from the Appendix A. of the paper of Grün and Hornik (2011, pp. 28–29) was modified and used to run the analysis. The model was optimized with the following choices and evaluation methods.

10-fold cross-validation was first run on a coarse grid of ks: 10, 20, 30, 40, 50, 100, and 200 with three different inference techniques: VEM with α fixed to the default $k/50$, VEM with α estimated by the model, and Gibbs sampling. The parameters for Gibbs sampling were set according to Grün and Hornik (2011, p. 17). *burnin* was 1 000, which means that the first 1 000 samples were ignored. *iter* was 1 000, which means that there were another 1 000 iterations after the discarded first 1 000 iterations. *thin* was 100, which means that every 100th sample was taken into account. The quality of the models was evaluated with the function `perplexity()`. Gibbs sampling was chosen as the inference technique because it yielded much smaller perplexity (~2 000) values than either version of VEM (~60 000) as seen in Figure 3.3.

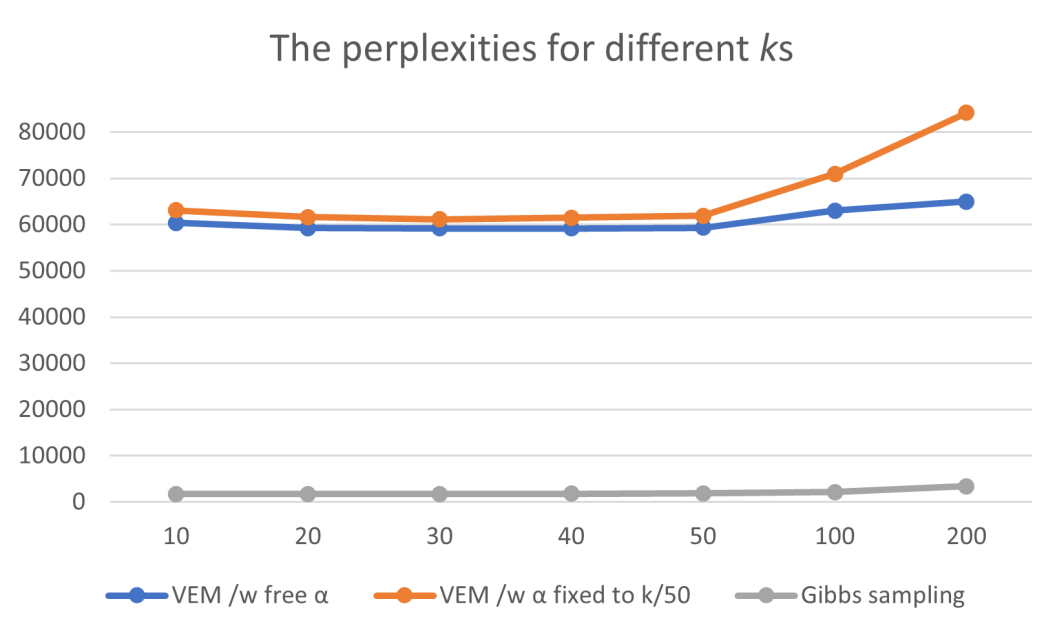


Figure 3.3. The perplexities for different ks and three different inference techniques.

The perplexity was lowest for the $k = 10$, $k = 20$, and $k = 30$ (Figure 3.4) and so the 10-fold cross-validation was run again on a fine grid [10, 30]. Now, the lowest

perplexities were for $k = 15$ (1 735.4), $k = 17$ (1 736.0) and $k = 18$ (1 735.3) (Figure 3.5). k was chosen to be 15 because it is the smallest of the three and thus easiest to interpret and present.

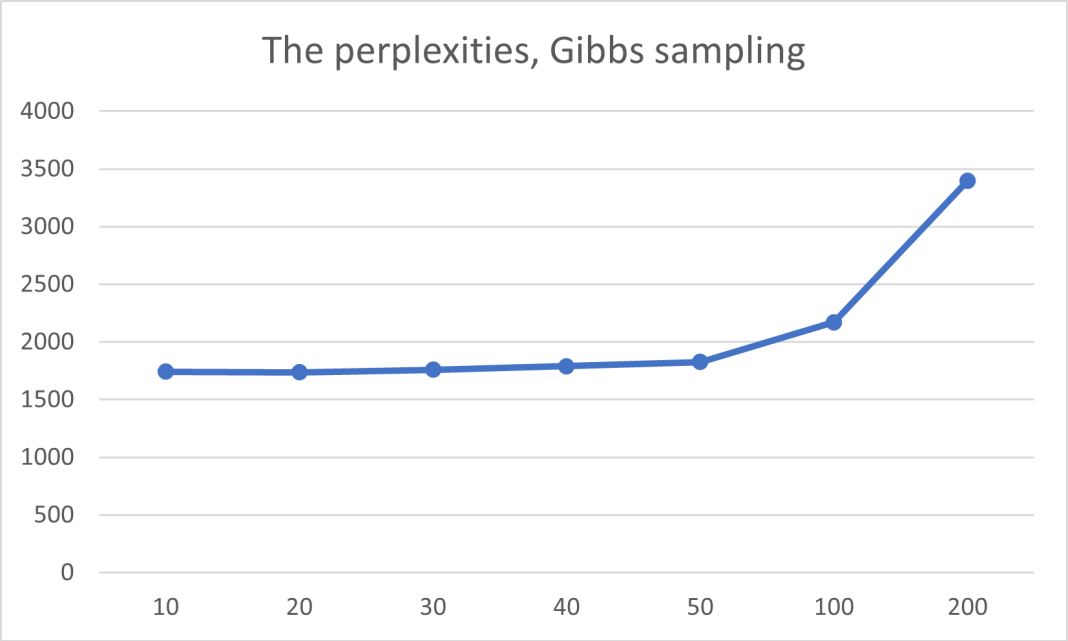


Figure 3.4. The perplexities for ks on coarse grid with Gibbs sampling as the inference technique.

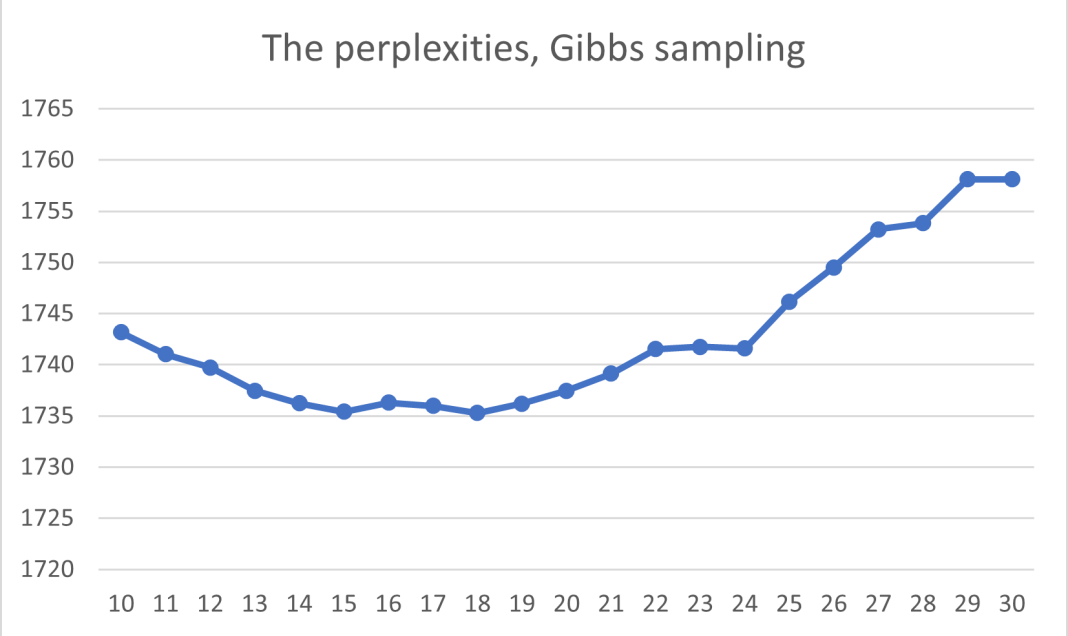


Figure 3.5. The perplexities for ks on fine grid with Gibbs sampling as the inference technique.

So far α that was used was the default $k/50$ with the exception of VEM with free

α. For $k = 15$ the default α was 0.3. After the k was set, a more optimal α was estimated by LDA that was run with VEM with free α as was done in the paper of Grün and Hornik (2011, p. 29). This order of magnitude smaller α (~0.029) was then used when the final model was built. The LDA model was fitted to the whole data set and the mean of topic coherences was calculated with `topic_coherence()` for 15 initializations to find the best model. The mean of topic coherences ranged from ~ -124.0 to ~ -106.7 and the model with the biggest mean was chosen to be the final model.

Topic modelling was also performed with $k = 4$ in the interest of comparing the resulting topics with the four major topics in Järvelin and Vakkari (2021). The procedure was simpler than the previous because k was set to 4 and there was no need to calculate the perplexities. The default value for α would have been $4/50 = 0.08$. It is very close to the estimated value from VEM with free α (~0.081) but the estimated value was used as being more accurate. The mean of topic coherences was between ~ -88.6 and ~ -82.5 during the 15 initializations and the model with the biggest mean is presented as the result in Section 4.1.

3.3 Co-Citation Analysis

Co-citation analysis was performed with the CiteSpace software (6.2.R2 Basic; Chen 2006). Another popular co-citation analysis software considered was CitNetExplorer (Van Eck and Waltman 2014). CiteSpace was chosen because it has the advantage of automatically labelling clusters (Chen, Ibekwe-SanJuan, and Hou 2010).

The collection of data for CiteSpace was very simple but the resulting data set was not exactly the same as for topic modelling nor as in the original Järvelin and Vakkari (2021) paper. Only the papers from Scopus were able to be included because the data has to be in a certain format for CiteSpace and it was not possible to add the papers from other sources. The same filters were used as before when the data was collected for the topic modelling analysis: 2015 as the publication year, and original paper or conference paper as the document type. This time the papers without abstracts were not discarded because the abstracts were not essential for the analysis. The data consisted of 1 402 papers from 29 journals.

There was no need to preprocess the data before importing it to CiteSpace. The data needed to be converted to the Web of Science export format before running the analysis but the conversion tool is included in CiteSpace.

There are many choices that can be made before the analysis is run in the software. The default values and options were chosen to be used. Only k as a scale factor for g-index (Egghe 2006) was given the values 50, 75, 100, 125, and 150 in addition to the default value of 25 to increase the number of obtained clusters and to see how it affects the result. It would be also easier to compare the result with LDA model if there was not a big difference in the number of topics. A modified g-index is CiteSpace's default node selection criteria to be used to sample records for the

network:

$$g^2 \leq k \sum_{i \leq g} c_i, k \in Z^+,$$

where c denotes the number of citations for a paper. It is an alternative to the older h-index to differentiate the scientists based on how many times they have been cited. Their difference is that g-index takes better into account the citation scores of an author's highly cited papers. The basic version of CiteSpace did not visualize the result of the analysis with $k \geq 175$ because it exceeded the network size limit. It is worth noting that the scale factor is named with the same letter, k, which is used also as a number of prefined topics in the context of topic modelling.

CiteSpace can take the cluster label from the titles, the abstracts, or the keyword lists of the papers citing the reference cluster or from the combination of all of those. It is also possible to use the titles of the references themselves as a source of labels but the default is to use the titles of the citing papers. Hou, Yang, and Chen (2018, p. 871) explain that the labels should reflect how the references have been cited and it is accomplished by using the citing papers instead of cited papers.

The labels can be chosen by Latent semantic indexing or by Log-likelihood ratio, or user defined cluster labels can be used. The labels chosen by LLR tests tend to reflect the uniqueness of clusters (Chen, Ibekwe-SanJuan, and Hou 2010, p. 1392). LSI, however, aims to find the underlying semantic structure of a document collection which leads the labels to represent the main themes of clusters. CiteSpace was unable to use the keywords of this data set but otherwise, all the possible combinations of the source of the labels and the algorithm were used. The Log-likelihood ratio gave the best results according to Hou, Yang, and Chen (2018, p. 879), and Markscheffel and Schröter (2021, p. 386). Also in this study, the best result was obtained with LLR and having titles of the citing papers as the source of labels. In all of the studies, the best was chosen intellectually by comparing the lists of label options.

Chen (2016, p. 69) recommends aiming to the result of having 7–10 main clusters which are of good quality. The clusters should be big and cohesive enough and it is achieved by having at least 10 nodes with silhouette values > 0.7. Silhouette value is a clustering quality metric that measures how similar a node is to other nodes in the same cluster compared to the nodes in the closest neighboring cluster. The number of clusters, the number of nodes in a cluster, and the silhouette values resulting from the different values of k are presented in Table 3.1. The default value $k = 25$ resulted in 7 clusters of good quality. Also, the results with values 50 and 75 for k were in line with Chen's recommendation. Values 100, 125, and 150 for k resulted in more than 10 clusters but they were of good quality.

Table 3.1. The qualities of the results of the co-citation analysis with different values of scale factor *k*.

k	Clusters	Nodes in a cluster	Silhouette values	Mean of silh. values
25	7	10–23	0.85–0.99	0.90
50	8	14–29	0.83–0.97	0.91
75	10	10–40	0.76–1.00	0.89
100	12	11–43	0.80–1.00	0.91
125	13	11–45	0.79–1.00	0.91
150	12	13–49	0.84–1.00	0.91

The final model, which utilized $k = 100$ and LLR as the algorithm for labelling, was selected based on an intellectual evaluation. Modelling with LLR was more robust than with LSI. There are three cluster labels that are present in the results with LLR for every value of *k*, while LSI models have only one cluster label, which is shared by all of them. All of the labels in the chosen model are related to LIS. Many of the other models include a cluster or two with irrelevant labels, such as *former warsaw pact* or *solar energy*.

4 Results

The result of topic modelling with latent Dirichlet allocation is presented in Section 4.1 and the result of co-citation analysis in Section 4.2. The result of LDA is compared to previous research in Section 4.3 and to the result of co-citation analysis in Section 4.4.

4.1 Latent Dirichlet Allocation

The final LDA model consists of 14 topics and 1 catch-all. The topics and their top 10 word lists are presented in Table 4.1. The topic modelling result is listed in declining order where the uppermost is the topic that covers the biggest proportion of the corpus. The labelling of topics was made by the writer of the thesis while taking into consideration the suggestions of Emeritus Professor Kalervo Järvelin (LIS expert, Tampere University), Emeritus Professor Pertti Vakkari (LIS expert, Tampere University), and Doctoral Researcher Chien Lu (topic modelling expert, Tampere University).

The first topic is labelled **Impact Indicators** and it covers research assessment indicators such as citation analysis and the *h*-index. **Education in LIS Studies and Education as LIS Service** includes both the library and information science education and information literacy and user education that is offered in libraries. The combining factor in the topic of **Academic Libraries** is academic libraries and especially the websites and other online services provided by them. **Information Retrieval** is in the core of information science side of LIS and it covers for example queries and indexing. **Computation-Assisted Analysis** is one of the two "topics" which have an analysis method related label but neither of them is a topic about researching the method but about employing the method. Topic modelling is a good example of computation-assisted analysis methods.

Scientific Collaboration topic is the topic of especially international scientific collaboration and co-authorship. **Public Libraries** is a topic in the core of the Library Science side of LIS concerning also the role of public libraries in their communities. **Interactive Information Retrieval** is a hot topic in LIS and it comes up in this LDA model as a topic of its own and not merged into **Information Retrieval**. **Knowledge and Patent Management** is the topic about the private sector and business related research. The "topic" labelled **Bibliometrics** is about research that employs bibliometric methods whereas the research on bibliometric measures is included in **Impact indicators**.

The topic labelled **General Terms** consists of frequent and non-specific LIS related words. Hence, it is not a proper topic but it is common for topic models to contain one or more overly general topics which are not useful in the task of finding the latent topics of a corpus (Boyd-Graber, Mimno, and Newman 2014, p. 235). The

Table 4.1. The extracted topics of the LDA model, $k = 15$.

Topic	Top 10 Words
Impact Indicators	citation, journal, research, publication, article, impact, paper, science, use, indicator
Education in LIS Studies and Education as LIS Service	student, learn, literacy, school, education, library, faculty, information, academic, study
Academic Libraries	library, use, study, service, academic, datum, university, resource, librarian, web
Information Retrieval	query, retrieval, document, method, search, use, approach, model, propose, term
Computation-Assisted Analysis (analysis method)	model, topic, algorithm, datum, propose, text, use, analysis, result, cluster
Scientific Collaboration	collaboration, network, scientific, research, country, international, publication, coauthorship, university, science
Public Libraries	library, public, information, community, service, lis, librarian, change, role, development
Interactive Information Retrieval	search, user, system, task, information, design, interface, recommendation, use, web
Knowledge and Patent Management	patent, technology, management, knowledge, innovation, business, firm, system, technological, service
Bibliometrics (analysis method)	research, analysis, science, study, paper, publication, bibliometric, journal, trend, field
(General Terms	information, study, use, research, knowledge, paper, analysis, health, process, model)
Open Access	datum, open, research, repository, access, project, management, institutional, paper, service
Information History	book, archive, digital, history, article, preservation, record, index, internet, reader
Social Media	social, medium, user, network, twitter, online, site, community, facebook, study
User Behaviour in Digital Environment	factor, model, use, study, influence, perceive, online, effect, knowledge, quality

topic in question is in parentheses in Table 4.1 and is not considered a real topic in the LDA model. **Open Access** topic covers also the more broad research topics of open science and research data management. **Information History** is about book and information history, and preserving written cultural heritage. **Social Media** topic combines social media related research topics but is mainly about information behaviour in social media and the topology of social media communities. Twitter and Facebook are included in the top 10 word list of **Social Media** by name and they are the only proper nouns in the top 10 word lists of the LDA model. **User Behaviour in Digital Environment** presents factors behind users' behaviour in e-commerce and other digital environments.

The LDA model with four topics is presented in Table 4.2. The most prominent topic is **Scientific and Professional Communication** which covers for example bibliometrics and scientific publishing. Information Seeking and Social Media. **Information Seeking and Social Media** combines the topics of information seeking and use in social media, and in the private sector. **Library and Information Service Activities** is a topic about public and academic libraries and information literacy. **Information Storage and Retrieval** is a topic of information retrieval and developing information retrieval.

Table 4.2. The extracted topics of the LDA model, $k = 4$.

Topic	Top 10 Words
Scientific and Professional Communication	research, citation, journal, analysis, science, study, publication, index, paper, author
Information Seeking and Social Media	information, study, use, social, user, knowledge, model, research, factor, online
Library and Information Service Activities	library, information, datum, use, book, digital, study, academic, student, research
Information Storage and Retrieval	search, base, use, model, user, method, datum, information, document, propose

4.2 Co-Citation Analysis

The final co-citation model consists of 12 clusters and their labels which have been automatically generated in CiteSpace. The visualization of the model is presented in Figure 4.1. The clusters' sizes and silhouette values can be found in Table 4.3.

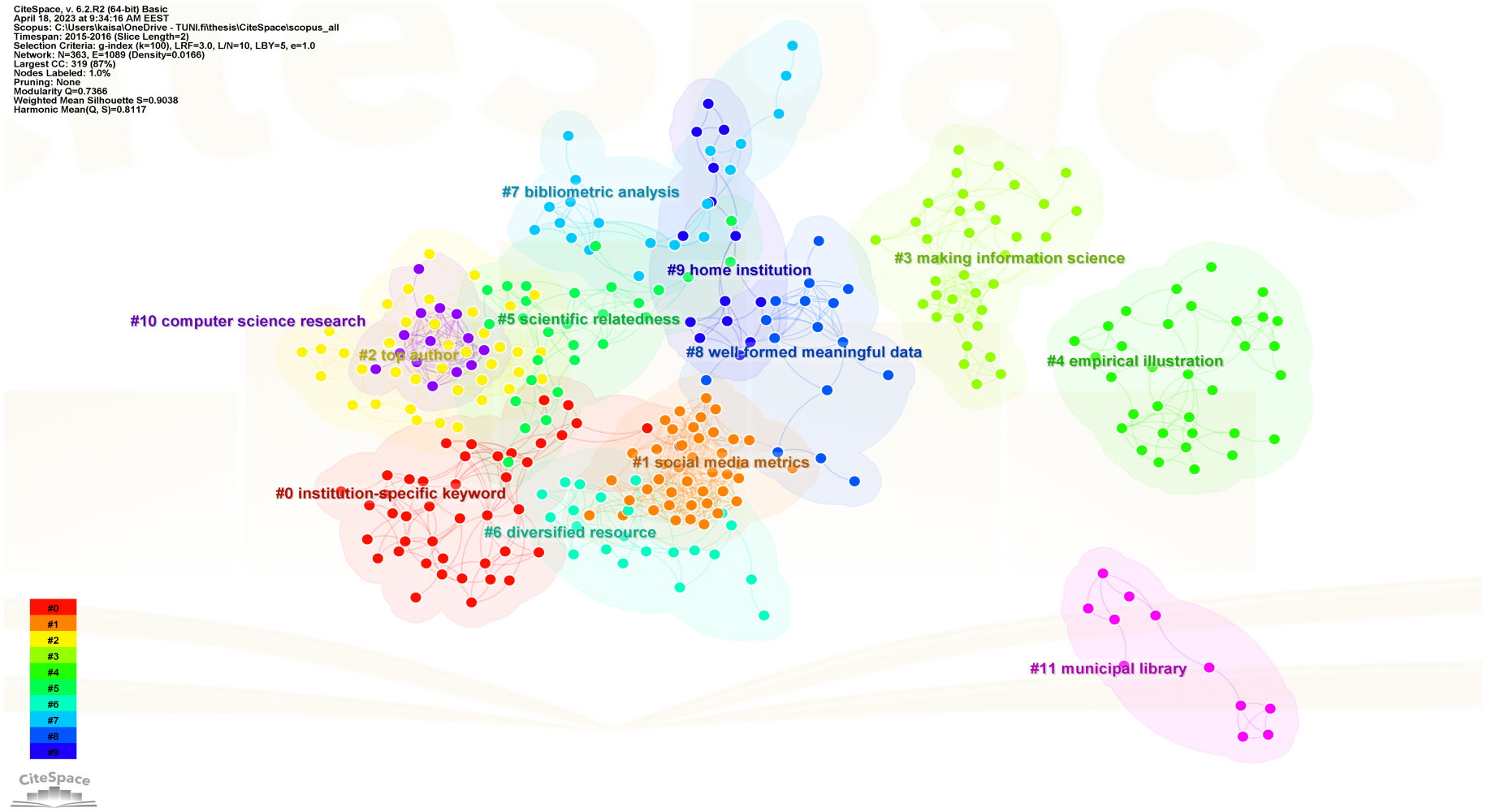
The majority of the labels refer to LIS research interests: **institution-specific keyword** and **well-formed meaningful data** to information retrieval, **social media**

metrics to social media, **top author**, **bibliometric analysis**, and **home institution** to scientific publishing, and **municipal library** to (public) libraries. **Scientific relatedness** and **computer science research** highlight the interdisciplinarity and multidisciplinarity of library and information science. However, the topic labels are very specific and they cannot be used as such to describe the field of library and information science. According to Chen, Ibekwe-SanJuan, and Hou (2010, p. 1406), it is common that algorithmically generated topic labels are more specific than those that have been chosen by humans. The pool of possible terms is also limited to those that have been used by the authors unless Wikipedia or another external source has been used.

Table 4.3. The labels and quality measures of the final co-citation model.

Node	Size	Silhouette	Label
0	43	0.84	institution-specific keyword
1	42	0.90	social media metrics
2	39	0.94	top author
3	37	0.92	making information science
4	33	0.97	empirical illustration
5	28	0.80	scientific relatedness
6	22	0.85	diversified resource
7	20	0.90	bibliometric analysis
8	16	0.95	well-formed meaningful data
9	14	0.95	home institution
10	14	0.90	computer science research
11	11	1.00	municipal library

Figure 4.1. CiteSpace-generated co-citation network of library and information science.



4.3 Results Compared to Previous Research

4.3.1 Järvelin and Vakkari (2021)

The final topic model of 14 topics compares well to Järvelin and Vakkari (2021). All of the four major topics of Järvelin and Vakkari (2021) get covered by the topics of LDA topic model. **Impact Indicators**, **Scientific Collaboration**, and **Open Access** fall under *Scientific and professional communication*. **Information Retrieval** and **Interactive Information Retrieval** are topics from *Information storage and retrieval*. The LDA topics **Knowledge and Patent Management**, and **User Behaviour in Digital Environment** are about *Information seeking*. *Library and information service activities* related research includes **Academic Libraries**, **Public Libraries**, and the latter part of **Education in LIS Studies and Education as LIS Service**.

However, all of the LDA topics do not match those four major topics. Järvelin and Vakkari (2021) have six smallish main topics in their classification scheme in addition to the four major topics and also sub-topics for the major topics. These are considered when comparing the rest of the LDA topics to Järvelin and Vakkari (2021). **Social Media** as a whole is not a topic of its own in Järvelin and Vakkari (2021) but some aspects of it are included in the classification scheme under a few different topics. For example *Social media retrieval* is a sub-topic of *Information storage and retrieval* and "presence in social media sites" is mentioned as an example of *Other types of information-seeking studies*, which is a sub-topic of *Information seeking*.

Information History combines the research of two small main topics from Järvelin and Vakkari (2021): *Library history* and *Publishing and book history*. Also **Education in LIS Studies and Education as LIS Service** is formed from two topics: aforementioned *Library and information service activities* and the small main topic of *Education in LIS*. **Information History** is a coherent topic combining the historical aspect of libraries, publishing, and books which all are closely related to each other. **Education in LIS Studies and Education as LIS Service**, on the contrary, is formed from two different topics. They both are related to the concept of education but one is about the education of LIS professionals and another is about the education provided by those LIS professionals.

Methodology is one of the small main topics but the analysis methods related LDA topics **Computation-Assisted Analysis** and **Bibliometrics** do not fit into that. The research analysis method in the papers forming these "topics" is the unifying factor found by LDA but the actual research topic varies from paper to paper. The LIS papers are classified along many factors in Järvelin and Vakkari (2021) and one of them is Research strategy. The classification scheme has 11 empirical methods of which *Citation analysis* and *Other bibliometric method* refer to **Bibliometrics**. **Computation-Assisted Analysis** would be included in *Other empirical method* as topic modelling and other more recent methods do not have a class of their own.

The analysis with k fixed to four topics was done to see whether it is comparable

to the four major topics in Järvelin and Vakkari (2021). The latent topics of the model are so close to the topic classification scheme of Järvelin and Vakkari (2021) that their topics can be used in the label assignment directly with the exception of broadening *Information Seeking* to **Information Seeking and Social Media** because social media is so noticeable in that topic.

4.3.2 Previous Latent Dirichlet Allocation Models of Library and Information Science Research

The LDA model is subjective to the choice of k as the number of topics and to labelling the topics on the basis of the top word lists. Figuerola, García Marco, and Pinto (2017), and Miyata et al. (2020) made the decision to use a bigger k (19 and 30, respectively) and then group the topics into a few categories. Sugimoto et al. (2011) had a small k (5) to begin with and Han's (2020) choice for k was something in between (10). k is 15 in this study and it is comparable to the 19 topics in Figuerola, García Marco, and Pinto (2017) and 10 topics in Han (2020).

Han's (2020) 10 topics describe the impact indicators, the scientific collaboration, and other aspects of scholarly communication in detail but for example, information retrieval and library related topics are missing from the model. The 19 topics of Figuerola, García Marco, and Pinto (2017) succeed better in covering LIS and they are also making serendipitous discoveries of research topics through topic modelling. *Legal and ethical aspects of LIS* and *Health information* are examples of topics that are not usually seen among the research topic categorizations of LIS.

The topic model of this thesis and the one of Figuerola, García Marco, and Pinto's (2017) are both good in finding the latent research topics of LIS. Their difference is that Figuerola, García Marco, and Pinto (2017) have decided to use more specific labels than the author of this thesis. The advantage of their approach is that their model describes LIS in more detail and gives examples of actual research topics. Conversely, the advantage of more broad labels of this thesis is that they cover the field of LIS better and they are more comparable to other studies.

4.4 Topic Modelling and Co-Citation Analysis Results Compared

The result of topic modelling has better quality than that of co-citation analysis when evaluating intellectually. The topics of the LDA model cover the field of library and information science well, and better than the topics of co-citation analysis. The main reason is that the topics of LDA model are labelled by humans and those of co-citation analysis are automatically labelled. It is still worth noting that the LDA model and its top word lists provide an accurate representation of the field of library and information science and it is not reliant on the chosen labels. The majority of the LDA labels are found or formed directly from the top 10 word lists.

The aim of labelling the LDA topics was to find labels which are broad enough to cover the field of LIS but still specific enough to describe the various research topics

of library and information science. The automatically generated topic labels of the co-citation analysis, however, are too specific and the result draws a granular picture of LIS.

Social Media (~ **social media metrics**) is a topic which is shared by LDA model and the co-citation analysis result but is not in Järvelin and Vakkari (2021) as a separate topic nor otherwise clearly noticeable. Other topics which are present in the results of both analysis methods of this thesis are **Bibliometrics** (~ **bibliometric analysis**) and **Public Libraries** (~ **municipal library**). Both models have also a separate topic for general terms or the broad concept of information science.

5 Conclusions

In this thesis, an LDA topic model was built on articles and conference papers from 30 library and information science journals published in 2015. The topics were labelled by the writer of the thesis while considering the proposals from LIS and topic modelling experts. The goal was to give the topics concise and not too specific labels. The LDA model was compared to the manual and intellectual classification of the same data set in Järvelin and Vakkari (2021), to previous topic models of LIS, and to the co-citation model, which was built with CiteSpace from the same data set.

The final LDA model consists of 14 topics: **Impact Indicators**, **Education in LIS Studies and Education as LIS Service**, **Academic Libraries**, **Information Retrieval**, **Computation-Assisted Analysis** (analysis method), **Scientific Collaboration**, **Public Libraries**, **Interactive Information Retrieval**, **Knowledge and Patent Management**, **Bibliometrics** (analysis method), **Open Access**, **Information History**, **Social Media**, and **User Behaviour in Digital Environment**. There is additionally one catch-all topic labeled **General Terms** but it is not considered as a proper topic. The LDA model's performance is comparable to other topic models and the intellectual classification by Järvelin and Vakkari (2021) in describing the field of LIS, and it outperforms the co-citation model.

The main difference between the topics of the LDA model and those in the classification scheme of Järvelin and Vakkari (2021), is the topic of **Social Media**. It shows up as a distinct research topic in the LDA model, while only some features of it are mentioned explicitly in Järvelin and Vakkari (2021) and otherwise it is merged to the classification scheme the same way as more traditional media. Another big difference is that there are two topics in the LDA model, **Computation-Assisted Analysis** and **Bibliometrics**, describing the analysis method instead of the research topic. Järvelin and Vakkari (2021), on the other hand, had classified LIS articles along several features, one of which being *Research strategy*. The topic of **Bibliometrics** is represented as two separate classes in it, *citation analysis* and *other bibliometric strategy*, but **Computation-Assisted Analysis** falls under *other empirical method*. What is common for **Social Media** and **Computation-Assisted Analysis**, is that they represent recent topical and methodological advancements in LIS.

A classification scheme can be fine-tuned to be highly detailed yet balanced, and a topic model cannot compete with that. A topic model can be expanded by increasing the value of k, but the quality of the model starts to decrease after a certain point, which depends on the data. Topics become less coherent and thus harder to interpret, topics begin to overlap, and the topic model as a whole loses its balance, and becomes too granular when k is too big. Topic modelling's advantage, compared to intellectual content analysis, is that it can provide serendipitous findings and new perspectives to the field. Classification scheme can be more prone to portray the field being studied from a traditional and conventional perspective and show the

development retrospectively.

There are some topics represented by top 10 word lists in the LDA model, which can be misleading if they are not interpreted correctly. **Education in LIS Studies and Education as LIS Service** was initially labeled as **Education**, but it was later given a longer label to reflect the two different research topics of LIS, which LDA had combined to one. Moreover, the topics of **Computation-Assisted Analysis** and **Bibliometrics** are not formed based on the actual research topics in the papers but based on the analysis methods employed.

Almost $^1/_4$ of the papers in the data set are from *Scientometrics*. It is a journal, which publishes research on scientific publications, citations, and other bibliometric data. It is debatable whether this big share of papers from *Scientometrics* causes bias in the discipline analysis. While it can be said that scientometrics is overrepresented in the data, it also reflects the substantial proportion of scientometrics research within the discipline. Tuomaala, Järvelin, and Vakkari (2014) made the decision to present two versions of the main results: one including the data from *Scientometrics* and one excluding it. However, Järvelin and Vakkari (2021) defend including the papers from *Scientometrics* in their data set by stating that scientometrics is part of LIS and it cannot be reasonably justified to select only a proportion of papers. **Impact Indicators** is possibly the most prominent topic in the LDA model because of the data from *Scientometrics*.

Computational discipline analyses have usually used bigger data sets than the one in this research. Also, one year's data alone is not as meaningful as it would be having several years' data and seeing the evolution of LIS. It would be interesting to see how the approach and workflow of this thesis would work in analysing the development of LIS by creating the LDA models also for other analysis years in Järvelin and Vakkari (2021): 1965, 1985, and 2005.

Major limitation for this thesis is that there is not available any high-quality automatic topic labelling package for R. Such a high-quality algorithm would use, for example, Wikipedia as a pool of possible labels, and not be restricted to using words, which are present in the data set.

Topic modelling is an objective statistical technique but the selection of journals in the context of discipline analysis, the preprocessing phase, and the interpretation of the results involve some subjectivity. Further development is recommended to automate the process by developing a comprehensive preprocessing framework and especially by implementing an automatic topic labelling tool for various platforms. A high-quality automated process would not only be effective but also reduce subjectivity and better enable serendipitous findings.

List of References

Aggarwal, C. C. (2018). "Text Preparation and Similarity Computation". In: *Machine Learning for Text*. Springer International Publishing, pp. 17–30.

Blei, D. M. and Lafferty, J. D. (2006a). "Correlated topic models". In: *Advances in neural information processing systems* 18, pp. 147–154.

— (2006b). "Dynamic topic models". In: *Proceedings of the 23rd international conference on machine learning*, pp. 113–120.

Blei, D. M. and McAuliffe, J. D. (2007). "Supervised Topic Models". In: *NIPS*, pp. 121–128.

Blei, D. M., Ng, A. Y., and Jordan, M. I. (2003). "Latent Dirichlet Allocation". In: *The Journal of Machine Learning Research* 3, pp. 993–1022.

Bornmann, L., Haunschild, R., and Mutz, R. (2021). "Growth rates of modern science: a latent piecewise growth curve approach to model publication numbers from established and new literature databases". In: *Humanities & social sciences communications* 8.1, pp. 1–15.

Boyd-Graber, J., Mimno, D., and Newman, D. (2014). "Care and Feeding of Topic Models: Problems, Diagnostics, and Improvementes". In: *Handbook of Mixed Membership Models and Their Applications*. CRC Press LLC, pp. 225–254.

Chang, J., Gerrish, S., et al. (2009). "Reading tea leaves: How humans interpret topic models". In: *Advances in neural information processing systems*, pp. 288–296.

Chang, Y.-W., Huang, M.-H., and Lin, C.-W. (2015). "Evolution of research subjects in library and information science based on keyword, bibliographical coupling, and co-citation analyses". In: *Scientometrics* 105.3, pp. 2071–2087.

Chauhan, U. and Shah, A. (2021). "Topic Modeling Using Latent Dirichlet allocation: A Survey". In: *ACM Computing Surveys (CSUR)* 54.7, pp. 1–35.

Chen, C. (2006). "CiteSpace II: Detecting and visualizing emerging trends and transient patterns in scientific literature". In: *Journal of the American Society for Information Science and Technology* 57.3, pp. 359–377.

— (2016). *CiteSpace : a practical guide for mapping scientific literature*. Nova Science Publishers.

Chen, C., Ibekwe-SanJuan, F., and Hou, J. (2010). "The structure and dynamics of cocitation clusters: A multiple-perspective cocitation analysis". In: *Journal of the American Society for Information Science and Technology* 61.7, pp. 1386–1409.

Chu, H. (2015). "Research methods in library and information science: A content analysis". In: *Library & information science research* 37.1, pp. 36–41.

Darling, W. M. (2011). "A theoretical and practical implementation tutorial on topic modeling and gibbs sampling". In: *Proceedings of the 49th annual meeting of the association for computational linguistics: Human language technologies*, pp. 642–647.

Deerwester, S. et al. (1990). “Indexing by latent semantic analysis”. In: *Journal of the American Society for Information Science* 41.6, pp. 391–407.

Egghe, L. (2006). “Theory and practise of the g-index”. In: *Scientometrics* 69.1, pp. 131–152.

Figuerola, C. G., García Marco, F. J., and Pinto, M. (2017). “Mapping the evolution of library and information science (1978–2014) using topic modeling on LISA”. In: *Scientometrics* 112.3, pp. 1507–1535.

Friedman, D. (2019). *topicdoc: Topic-Specific Diagnostics for LDA and CTM Topic Models*. R package version 0.1.0. URL: https://CRAN.R-project.org/package=topicdoc.

Griffiths, T. L. and Steyvers, M. (2004). “Finding scientific topics”. In: *Proceedings of the National Academy of Sciences - PNAS* 101.Suppl 1, pp. 5228–5235.

Grün, B. and Hornik, K. (2011). “topicmodels: An R Package for Fitting Topic Models”. In: *Journal of Statistical Software* 40.13, pp. 1–30. DOI: 10.18637/jss.v040.i13.

Han, X. (2020). “Evolution of research topics in LIS between 1996 and 2019: an analysis based on latent Dirichlet allocation topic model”. In: *Scientometrics* 125.3, pp. 2561–2595.

Heinrich, G. (2005). *Parameter estimation for text analysis*. Tech. rep. Technical report.

Hofmann, T. (1999). “Probabilistic Latent Semantic Analysis”. In: *UAI ’99: Proceedings of the Fifteenth Conference on Uncertainty in Artificial Intelligence, Stockholm, Sweden, July 30 - August 1, 1999*, pp. 289–296.

Hou, J., Yang, X., and Chen, C. (2018). “Emerging trends and new developments in information science: a document co-citation analysis (2009–2016)”. In: *Scientometrics* 115.2, pp. 869–892.

Järvelin, K. and Vakkari, P. (1990). “Content analysis of research articles in library and information science”. In: *Library & information science research* 12.4, pp. 385–421.

— (1993). “The evolution of library and information science 1965–1985: A content analysis of journal articles”. In: *Information Processing & Management* 29.1, pp. 129–144. DOI: 10.1016/0306-4573(93)90028-C.

— (2021). “LIS research across 50 years: content analysis of journal articles”. In: *Journal of documentation* 78.7, pp. 65–88.

Kherwa, P. and Bansal, P. (2019). “Empirical Evaluation of Inference Technique for Topic Models”. In: *Progress in Advanced Computing and Intelligent Engineering*. Vol. 713. Advances in Intelligent Systems and Computing. DOI: 10.1007/978-981-13-1708-8_22.

Layman, L. et al. (2016). “Topic Modeling of NASA Space System Problem Reports: Research in Practice”. In: *2016 IEEE/ACM 13th Working Conference on Mining Software Repositories (MSR)*, pp. 303–314.

Li, P., Yang, G., and Wang, C. (2019). “Visual topical analysis of library and information science”. In: *Scientometrics* 121.3, pp. 1753–1791.

Li, W. and McCallum, A. (2006). “Pachinko allocation: DAG-structured mixture models of topic correlations”. In: *Proceedings of the 23rd international conference on machine learning*, pp. 577–584.

Liu, P. et al. (2015). “Detecting the intellectual structure of library and information science based on formal concept analysis”. In: *Scientometrics* 104.3, pp. 737–762.

Ma, J. and Lund, B. (2020). “The evolution of LIS research topics and methods from 2006 to 2018: A content analysis”. In: *Proceedings of the Association for Information Science and Technology* 57.1.

Markscheffel, B. and Schröter, F. (2021). “Comparison of two science mapping tools based on software technical evaluation and bibliometric case studies”. In: *Collnet journal of scientometrics and information management* 15.2, pp. 365–396.

Mayo, M. (2017). *A General Approach to Preprocessing Text Data*. Accessed 2022-03-08. URL: https://www.kdnuggets.com/2017/12/general-approach-preprocessing-text-data.html.

Mimno, D. et al. (2011). “Optimizing Semantic Coherence in Topic Models”. In: *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pp. 262–272.

Miyata, Y. et al. (2020). “Knowledge structure transition in library and information science: topic modeling and visualization”. In: *Scientometrics* 125.1, pp. 665–687.

Mulunda, C. K., Wagacha, P. W., and Muchemi, L. (2018). “Review of Trends in Topic Modeling Techniques, Tools, Inference Algorithms and Applications”. In: *2018 5th International Conference on Soft Computing & Machine Intelligence (ISCMI)*, pp. 28–37. DOI: 10.1109/ISCMI.2018.8703231.

Natural-language processing (2022). Accessed: 2022-03-08. URL: https://www.oxfordreference.com/search?q=natural+language+processing&searchBtn=Search&isQuickSearch=true.

Onyancha, O. B. (2018). “Forty-Five Years of LIS Research Evolution, 1971–2015: An Informetrics Study of the Author-Supplied Keywords”. In: *Publishing research quarterly* 34.3, pp. 456–470.

R Core Team (2020). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. Vienna, Austria. URL: https://www.R-project.org/.

Sugimoto, C. R. et al. (2011). “The shifting sands of disciplinary development: Analyzing North American Library and Information Science dissertations using latent Dirichlet allocation”. In: *Journal of the American Society for Information Science and Technology* 62.1, pp. 185–204.

Text mining (2022). Accessed: 2022-03-08. URL: https://www.oxfordreference.com/search?q=text+mining&searchBtn=Search&isQuickSearch=true.

Tuomaala, O., Järvelin, K., and Vakkari, P. (2014). “Evolution of library and information science, 1965–2005: Content analysis of journal articles”. In: *Journal of*

the Association for Information Science and Technology 65.7, pp. 1446–1462. DOI: 10.1002/asi.23034.

Ullah, A. and Ameen, K. (2018). "Account of methodologies and methods applied in LIS research: A systematic review". In: *Library & information science research* 40.1, pp. 53–60.

Van Eck, N. J. and Waltman, L. (2014). "CitNetExplorer: A new software tool for analyzing and visualizing citation networks". In: *Journal of Informetrics* 8.4, pp. 802–823. DOI: https://doi.org/10.1016/j.joi.2014.07.006.

Vayansky, I. and Kumar, S. A. P. (2020). "A review of topic modeling methods". In: *Information Systems* 94, pp. 1–15. DOI: 10.1016/j.is.2020.101582.

Wallach, H. et al. (2009). "Evaluation methods for topic models". In: *Proceedings of the 26th Annual International Conference on machine learning*, pp. 1105–1112.

Appendix: The Journals

ACM Transactions on Information Systems
Aslib Journal of Information Management
College and Research Libraries
Electronic Library
The Indexer : The International Journal of Indexing
Information & Culture
Information Processing & Management
Information Research : An International Electronic Journal
Information Retrieval
Information Services & Use
Information Technology and Libraries
International Information & Library Review
International Journal of Information Management
Journal of Documentation
Journal of Education for Library and Information Science
Journal of Information Science
Journal of Librarianship and Information Science
Journal of Library Administration
Journal of the Association for Information Science and Technology
Library & Information History
Library & Information Science Research
Library Collections, Acquisition, & Technical Services
Library Quarterly
Library Resources and Technical Services
Library Trends
Libri : International Journal of Libraries and Information Studies
New Review of Information Networking
Online Information Review
The Program
Reference & User Services Quarterly
Scientometrics