Topic evolution of digital library research: Using word embedding and clustering techniques to understand research topics

Zach Coble
School of Information Science & Learning Technologies
University of Missouri
Columbia, MO 65201
zccw9d@umsystem.edu

Abstract

This paper employs unsupervised learning methods to analyze the evolution of topics in the digital library literature. As a sub-field of library and information science, research on digital libraries sits at the intersection of librarianship, computer science and information retrieval, and the economics of publishing. While bibliometric studies have examined productivity metrics, such as citation statistics, there is no existing research on what the research is about or how it has evolved over time. Data was gathering from abstracts published from 1996-2023 and indexed in the Library and Information Science Source database. Word embedding was performed on the abstracts and the resulting vectors were sub-divided into 4-year increments. After dimension reduction, the sub-corpora were clustered and topics extracted. Content analysis was used to label the topics and discuss their evolution over time. Quality clusters emerged along with important topical changes, most notably in the clusters corresponding to digital library software development.

1 Introduction

Since emerging in the 1990s as a sub-field of library and information science, research on digital libraries has advanced rapidly. Arms defines a digital library as “a managed collection of information, with associated services, where the information is stored in digital formats and accessible over a network”[1]. This includes a wide variety of applications, from digital renderings of materials from libraries, archives, and museums, to online journals, to research materials provided by commercial database vendors[2]. Research in digital libraries brings together a unique combination of related disciplines, including librarianship, computer science and information retrieval, and legal and economic aspects of publishing[1].

Simultaneously, library and information science (LIS) researchers have developed several approaches for assessing the output of research literature using bibliometrics. Bibliometric methods can be used to examine the intellectual structure and development of disciplines through techniques such as co-citation analysis, co-word analysis, and topic models. Recent technical advances in natural language processing have created opportunities for computational model-based bibliometric analysis, such as topic modeling and clustering. These model-based approaches can be applied in various ways, allowing researchers to analyze a larger amount of text that was possible with content analysis and other bibliometric methods[3]. While bibliometric studies have examined research output and scholarly productivity in digital libraries, there has not yet been a study of the evolution of research topics within the field. This study examines how research topics in the digital library literature have evolved from 1996-2023, as represented in article abstracts indexed in the Library and Information Science Source database.

2 Literature review

This review describes bibliometric analyses of the LIS literature that employ model-based methods. Model-based methods have not yet been applied to the digital library literature, so this review describes studies across the wider field of LIS. The review concludes with a summary of bibliometric studies on the subfield of digital libraries, and is arranged by the type of model used.

2.1 Latent Dirichlet allocation

Latent Dirichlet allocation (LDA) is a commonly used model-based technique for topic modeling[4]. LDA uncovers existing topics in a corpus using a Bayesian approach to reveal topics through a generative statistical model examining word co-occurrences in the same document and among all documents in the corpus[5]. LDA has proven successful in creating meaningful clusters of topics from large corpora [3][6][7]. The ability to discover thematic structure from a large number of documents is due to the method's two-way estimation of both documents and words[8].

LDA has been widely applied to the academic literature to understand the evolution of research themes[9-12]. Within LIS, Sugimoto et al. applied LDA to uncover themes in 3,121 doctoral dissertations from 1930 to 2009[13]. Figuerola et al. analyzed titles and abstracts from 92,705 documents published between 1978–2014 from the database LISA to identify 19 dominant topics and four main areas[14]. Han applied an LDA model to articles published in 20 high impact factor journals between 1996–2019 to identify changes in LIS research topics[3].

One drawback of LDA is that it relies on a term-document matrix, which provides a simple word count for each document in the corpus (e.g. document 1 contains the term "digital" three times, the term "library" four times, etc). This approach does not consider the order that terms appear in (e.g. does "digital" appear before or after "library"). In contrast, word embeddings contain information on both word counts as well as the order, or context, that words appear in.

2.2 Word Embeddings

Word embedding models examine the relationship between words' semantic and syntactic features[15]. While LDA analyzes the co-occurrence frequency of words in and across documents, word embedding models go further by analyzing the contextual relationship among words, which gives a key advantage of being able to uncover the semantic regularities of documents[16][17][18].

Fewer studies have applied word embeddings to the LIS literature that LDA. While word embedding provides a more sophisticated analysis, it is computationally intensive and requires combination with other techniques, such as clustering or topics models, in order to discover latent topics. Nieminen et al. established that when applied to abstracts from the academic literature, output from the combination of word embedding and clustering corresponds to topics[24]. Gao et al. combined topic modeling and word embedding techniques to achieve a more nuanced analysis of topics in the LIS literature, from 66,830 abstracts from documents published between 2000-2019 gathered from the Web of Science database[19].

The literature shows that model-based methods are effective at extracting topics from the scholarly literature. While LDA is an established and reliable method, the more recent approach of word embedding and clustering also has proven capable of discovering topics due to its complex representation of the corpus. Across LIS, several studies have used LDA and a few have used word embedding, although no studies have applied a model-based method to analyze the sub-field of digital libraries.

2.3 Digital Library Literature

Among bibliometric analyses of the digital library literature, Ahmad et al. examines productivity metrics from 4,206 documents from Web of Science published between 2002-2016[20]. They found that the most productive year of publication was 2016, the number citations per article have increased rapidly, and that the U.S. dominates in terms of research output. Singh et al. analyzed 1,000 digital library research articles from 1998-2004 from the LISA Plus database[21]. They found that 61 percent are single-authored, D-Lib Magazine has published the most articles, author productivity is not in agreement with Lotka's Law, and that distribution of articles nearly follows Bradford's Law. Although these studies provide an important aspect of the research output, there is not an adequate understanding of the major topics in digital library research and how they have evolved during the digital era.

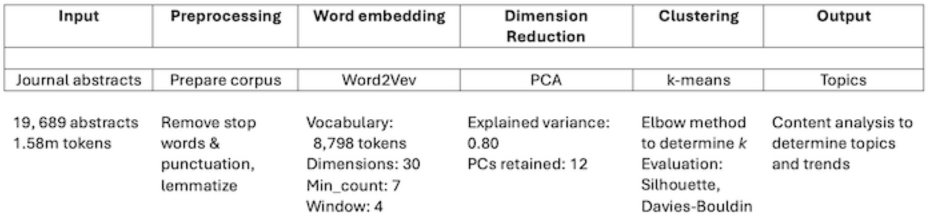
2.4 Research Questions

The purpose of this bibliometric analysis is to discover the evolution of key topics and themes in digital library research from 1996-2023, as represented in abstracts in the Library and Information Science Source database. Key topics are defined as an automated content analysis of article abstracts using machine learning techniques. Key themes are applied using content analysis of changes in the key topics over time.

This study examines the question, how have research topics in digital libraries evolved from 1996-2023, as represented in abstracts in the Library and Information Science Source database? Additional research sub-questions to be addressed include: What are the identified trends in the research topics within digital libraries between 1996-2023? Do the research topics for 2020-2023 continue trends from previous years?

3 Methodology

The framework for this study is presented in Figure 1, which follow a conventional machine learning model using natural language processing techniques. The abstracts from the database search were preprocessed and then used to generate a word embedding model. Dimension reduction was applied to the model to improve performance, and the resulting vector representations were divided into seven sub-corpora of abstracts from four-year time increments (i.e. 1996-1999, 2000-2003, etc). A clustering algorithm was then applied to each of the sub-corpora to extract topics for the four-year time increments. The topics were then manually labelled and analyzed to identify patterns and changes over time.



Input	Preprocessing	Word embedding	Dimension Reduction	Clustering	Output
Journal abstracts	Prepare corpus	Word2Vev	PCA	k-means	Topics
19, 689 abstracts 1.58m tokens	Remove stop words & punctuation, lemmatize	Vocabulary: 8,798 tokens Dimensions: 30 Min_count: 7 Window: 4	Explained variance: 0.80 PCs retained: 12	Elbow method to determine *k* Evaluation: Silhouette, Davies-Bouldin	Content analysis to determine topics and trends

Figure 1: Natural language processing and clustering framework

3.1 Preprocessing

Techniques common in natural language processing (NLP) for preparing textual corpora for analysis were applied during preprocessing, including lemmatization and the removal of stop words and punctuation. All words were converted to lowercase, which resulted in a trade-off between losing information on proper nouns in favor of gaining a deduplicated word count (i.e. no duplicates for a word when it appears first in the sentence as uppercase versus later in the sentence as lowercase). The original corpus contained 1.58 million words, and this was reduced to 19,689 unique tokens after preprocessing[1].

3.2 Word Embedding

The preprocessed dataset was used as the input to generate word embeddings using the Word2Vec model[16]. Word embedding models represent the textual corpus numerically by generating vectors

[1] The number of unique tokens after preprocessing, 19,689, is equal to the number of documents in the corpora. This is purely coincidental.

representing each document (abstract) in the corpus. These numerical representations contain the semantic and syntactic features of the corpus[17].

To represent the corpus numerically, the Word2Vev model was used to create word embeddings. The Gensim Word2Vec model was used with the default parameters, with the following exceptions: the window was decreased to 4 to improve computation since abstracts are relatively short, and the vector size was reduced to 30, which resulted in higher quality clusters. The minimum count was increased to 7 to reduce noise during topic extraction, which reduced the number of analyzed tokens to 8,798. The continuous bag of words algorithm was used to generate the embeddings, which Zhang et al. found performed slightly better than skip-grams for embeddings of article abstracts[22].

3.3 Dimension Reduction

Once represented as vectors, the resulted word embedding model contains thirty dimensions (corresponding to the vector size parameter). While this provides a rich understanding of the corpus, the high dimensionality makes it more difficult to cluster the data. This is commonly called the curse of dimensionality and can be addressed by reducing the number of dimensions in the data[23]. This is necessary as clustering algorithms in general perform better on lower dimensional data. The principal component analysis technique was applied to the embeddings in order to generate better clusters.

Principal component analysis is a linear technique for dimension reduction that is used to explain the maximal amount of variance in the data. An explained variance ratio of 0.80 was used, meaning that 80% of the information in the embedding vectors was kept and 20% was lost. This resulted in twelve principal components, meaning that the data was reduced from thirty dimensions to twelve.

3.4 Clustering and Topic Extraction

The k-means algorithm was used for clustering, which Zhang et al. found was the most effective clustering technique for word embeddings of article abstracts[22]. The elbow method was used to determine the most appropriate number of clusters, (k), for each of the seven sub-corpora.

Once the clusters were generated, internal analysis metrics were calculated to assess the quality and accuracy of the clustering process. Ideal clusters are compact and well-separated, meaning that similar documents are closely positioned to each other and clearly distinct from other groups of documents. Conversely, poor clustering results in sparse and overlapping clusters, which indicates that the topics are porous and not clearly defined. Analysis metrics included the Silhouette score and the Davies–Bouldin index, which use the within-cluster sum of squares to measure overall cluster quality.

3.5 Topic Extraction

The resulting clusters were then manually analyzed using content analysis to provide labels for the extracted topics and to assess any changes in the topics over time. For each cluster, the 40 most representative words were extracted, which correspond to the topics for the that cluster. Additionally, the ten most representative documents abstracted were extracted and analyzed. Labels were assigned using the most representative words and the most representative documents, along with the author's experience as a researcher and practitioner in the field of digital librarianship.

4 Results

In total, 77 clusters were created, with an average of eleven clusters per four-year increment. Table 1 provides an overview of these groups, including the number of clusters and the number of abstracts contained in each cluster. The table also reports the Silhouette coefficient, which is an average of each document's individual Silhouette scores. The Silhouette coefficients range between 0.08 - 0.12, which suggests lower quality clusters. Typically, Silhouette scores above 0.5 are considered good.

While the Silhouette scores are relatively low, clear topics did emerge from the clusters. Appendix 1 provides the labels applied to each of the 77 clusters. Figure 2 provides a summary of the cluster labels by showing the top three clusters (in terms of Silhouette scores) for each time frame along

Table 1: Silhouette Coefficient Scores for Date Ranges

YEARS	SIZE	TOTAL CLUSTERS, K	SILHOUETTE COEFFICIENT
1996-1999	690	13	0.10
2000-2003	1941	12	0.08
2004-2007	3821	10	0.09
2008-2011	4012	13	0.09
2012-2015	3499	11	0.10
2016-2019	3035	10	0.10
2020-2023	2691	8	0.12

with the label. For further details, Figure 3 shows an example of the report that was analyzed for each individual cluster, which contains the Silhouette score, the top 40 most representative words, and ten most representative documents.

5 Discussion

Although the analysis is not yet complete, several topics have appeared consistently over time. The European and international topics appears consistently across all time spans analyzed. The most representative words primarily contain geographic entities (e.g. Bavarian, Europeana, India), and may be explained by international researchers having interests that are distinct from U.S. researchers[20]. This cluster periodically includes terms with British spelling (e.g. centre, digitised) and could be something to normalize in future studies.

The topic for software, systems, development shows the most distinct changes over time. It first appears clearly in 2004-2007, with terms specific to software development (operational, develop, standardize) along with terms specific to digital library software (DSpace, Sword, ETD). In particular, the presence of the terms ”setup” and ”establish” indicate that authors are writing about early efforts to establish digital library infrastructure. In 2008-2011, most of the same terms appear (develop, standardize, DSpace, ETD) along with news terms, such as ”WCAG,” ”dashboard,” and ”compliant” that suggest that software development is reaching a mature phase. Authors are less concerned with establishing operational software and can now focus on digital accessibility (WCAG compliance) and customizing dashboard.

6 Conclusion

tbd

Acknowledgments

Thank you to Dr. Yunxin Zhao for excellent teaching of the methods used here, and to Dr. Hyerim Cho for expert advice at every phase of the project.

References

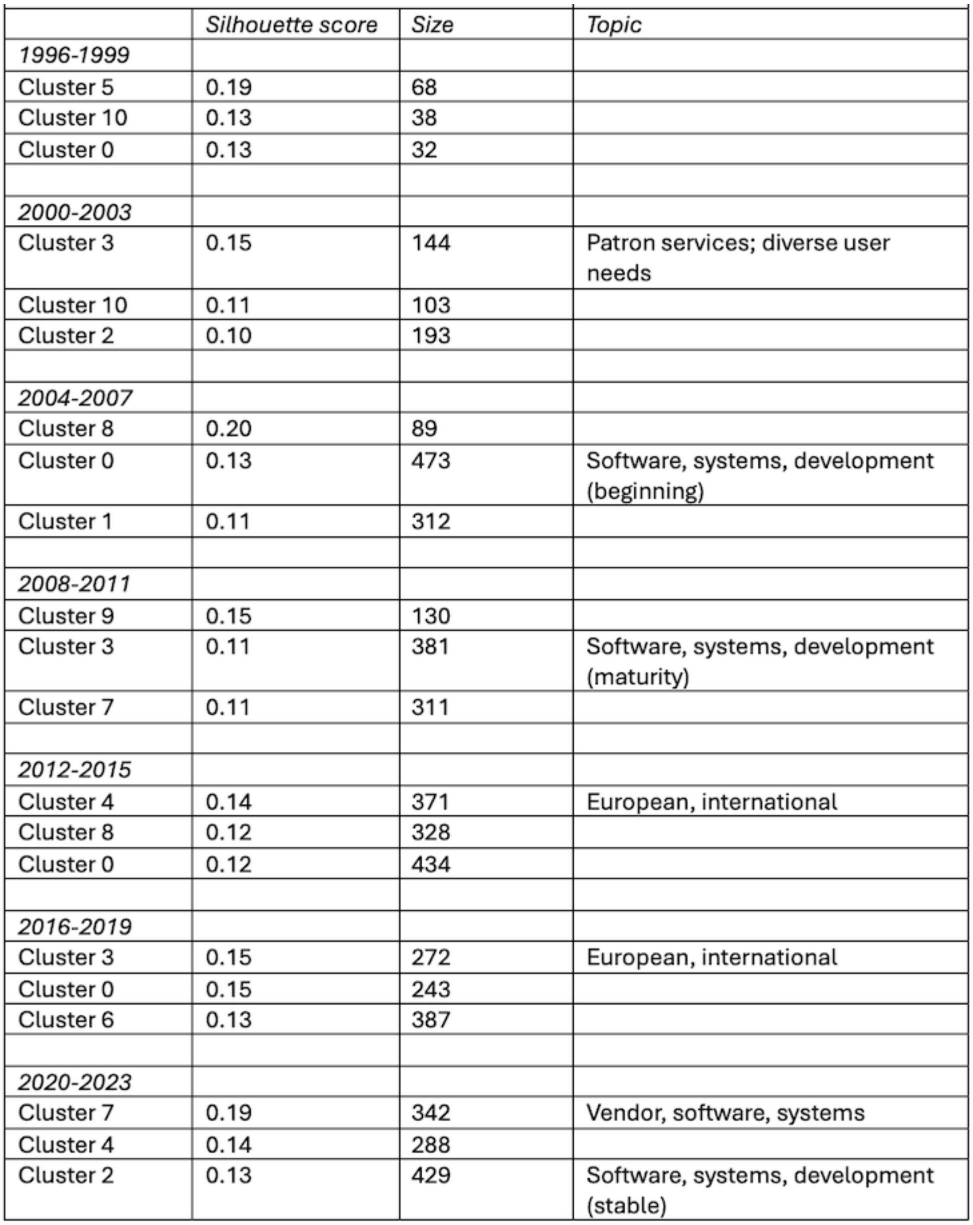
[1] Arms, W.Y. (2001) *Digital libraries.* Cambridge, MA: MIT Press.

[2] Cleveland, G. (1998) *Digital libraries: Definition, issues, and challenges (Occasional Paper #8).* International Federation of Library Associations and Institutions. https://archive.ifla.org/VI/5/op/udtop8/udt-op8.pdf

[3] Han, X. (2020) Evolution of research topics in LIS between 1996 and 2019: an analysis based on latent Dirichlet allocation topic model. *Scientometrics* **125**, 2561–2595. DOI: 10.1007/s11192-020-03721-0

[4] Blei, D.M., Ng, A.Y., & Jordan, M.I. (2003) Latent dirichlet allocation. *Journal of Machine Learning Research*, **3**(2003), 993-1022. https://www.jmlr.org/papers/volume3/blei03a/blei03a.pdf

[5] Yau, C.K., Porter, A., Newman, N., & Suominen, A. (2014) Clustering scientific documents with topic modeling. *Scientometrics*, **100**(3), 767–786. DOI: 10.1007/s11192-014-1321-8



	Silhouette score	*Size*	*Topic*
1996-1999			
Cluster 5	0.19	68	
Cluster 10	0.13	38	
Cluster 0	0.13	32	
2000-2003			
Cluster 3	0.15	144	Patron services; diverse user needs
Cluster 10	0.11	103	
Cluster 2	0.10	193	
2004-2007			
Cluster 8	0.20	89	
Cluster 0	0.13	473	Software, systems, development (beginning)
Cluster 1	0.11	312	
2008-2011			
Cluster 9	0.15	130	
Cluster 3	0.11	381	Software, systems, development (maturity)
Cluster 7	0.11	311	
2012-2015			
Cluster 4	0.14	371	European, international
Cluster 8	0.12	328	
Cluster 0	0.12	434	
2016-2019			
Cluster 3	0.15	272	European, international
Cluster 0	0.15	243	
Cluster 6	0.13	387	
2020-2023			
Cluster 7	0.19	342	Vendor, software, systems
Cluster 4	0.14	288	
Cluster 2	0.13	429	Software, systems, development (stable)

Figure 2: Top Three Clusters per Date Range

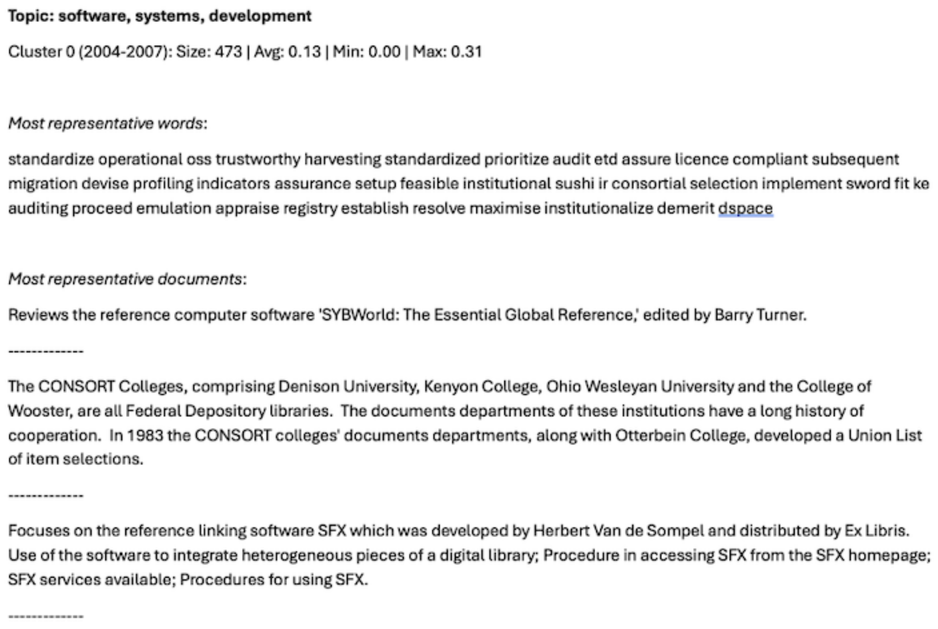
[6] Blei, D. M., & Lafferty, J. D. (2007). A correlated topic model of science. *The Annals of Applied Statistics*, bf 1(1), 17–35. DOI: 10.1214/07-aoas114

[7] Suominen, A., & Toivanen, H. (2015) Map of science with topic modeling: Comparison of unsupervised learning and human-assigned subject classification. *Journal of the Association for Information Science and Technology*, **67**(10), 2464–2476. DOI: 10.1002/asi.23596

[8] Nieminen, P., Polonen, I., & Sipola, T. (2013). Research literature clustering using diffusion maps. Journal of Informetrics, 7(4): 874-886. https://doi.org/10.1016/j.joi.2013.08.004

[9] Griffiths, T. L., & Steyvers, M. (2004) Finding scientific topics. *Proceedings of the National Academy of Sciences*, **101**(suppl 1), 5228–5235. DOI: 10.1073/pnas.0307752101

[10] Jeong, D.H., & Song, M. (2014) Time gap analysis by the topic model-based temporal technique. *Journal of Informetrics*, **8**(3), 776–790. DOI: 10.1016/j.joi.2014.07.005



Topic: software, systems, development

Cluster 0 (2004-2007): Size: 473 | Avg: 0.13 | Min: 0.00 | Max: 0.31

Most representative words:

standardize operational oss trustworthy harvesting standardized prioritize audit etd assure licence compliant subsequent migration devise profiling indicators assurance setup feasible institutional sushi ir consortial selection implement sword fit ke auditing proceed emulation appraise registry establish resolve maximise institutionalize demerit dspace

Most representative documents:

Reviews the reference computer software 'SYBWorld: The Essential Global Reference,' edited by Barry Turner.

The CONSORT Colleges, comprising Denison University, Kenyon College, Ohio Wesleyan University and the College of Wooster, are all Federal Depository libraries. The documents departments of these institutions have a long history of cooperation. In 1983 the CONSORT colleges' documents departments, along with Otterbein College, developed a Union List of item selections.

Focuses on the reference linking software SFX which was developed by Herbert Van de Sompel and distributed by Ex Libris. Use of the software to integrate heterogeneous pieces of a digital library; Procedure in accessing SFX from the SFX homepage; SFX services available; Procedures for using SFX.

Figure 3: Cluster report summary, with Silhouette score, top 40 most representative keywords, and top 10 most representative documents

[11] Rosen-Zvi, M., T. Griffiths, M. Steyvers and P. Smyth (2012) The author-topic model for authors and documents. *arXiv*. DOI: 10.48550/arXiv.1207.4169

[12] Wang, X., C. Zhai & Roth, D. (2013) Understanding evolution of research themes: a probabilistic generative model for citations. *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining*. DOI: 10.1145/2487575.2487698

[13] Sugimoto, C. R., Li, D., Russell, T. G., Finlay, et a. (2010) The shifting sands of disciplinary development: Analyzing North American Library and Information Science dissertations using latent Dirichlet allocation. *Journal of the American Society for Information Science and Technology*, bf 62(1), 185–204. DOI: 10.1002/asi.21435

[14] Figuerola, C. G., García Marco, F. J., & Pinto, M. (2017) Mapping the evolution of library and information science (1978–2014) using topic modeling on LISA. *Scientometrics*, **112**(3), 1507–1535. DOI: 10.1007/s11192-017-2432-9

[15] Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013) Efficient estimation of word representations in vector space. *arXiv*. https://doi.org/10.48550/arXiv.1301.3781

[16] Gastaldi, J. L. (2021) Why can computers understand natural language?" *Philosophy & Technology*, **34**(1), 149–214. DOI: 10.1007/s13347-020-00393-9

[17] Fu, R.J., Guo, J., Qin, B., et al (2014). Learning semantic hierarchies via word embeddings. *52nd Annual Meeting of the Association for Computational Linguistics (ACL)*, Baltimore, MD, Association of Computational Linguistics. https://aclanthology.org/P14-1113

[18] Hamilton, W. L., Leskovec, J. & Jurafsky, D. (2016) Diachronic word embeddings reveal statistical laws of semantic change. *54th Annual Meeting of the Association for Computational Linguistics (ACL)*, Berlin, Germany, Association of Computational Linguistics. https://aclanthology.org/P16-1141

[19] Gao, Q., Huang, X., Dong, K. et al. Semantic-enhanced topic evolution analysis: a combination of the dynamic topic model and word2vec. *Scientometrics* **127**, 1543–1563 (2022). DOI: 10.1007/s11192-022-04275-z

[20] Ahmad, K., Jian Ming, Z., & Rafi, M. (2018) Assessing the digital library research output: Bibliometric analysis from 2002 to 2016. *The Electronic Library*, **36**(4), 696–704. DOI: 10.1108/EL-02-2017-0036

[21] Singh, G., Mittal, R., & Ahmad, M. (2007) A bibliometric study of literature on digital libraries. *The Electronic Library*, **25**(3), 342–348. DOI: 10.1108/02640470710754841

[22] Zhang, Y. et al. (2018). Does deep learning help topic extraction? A kernel k-means clustering method with word embedding. *Journal of Informetrics*, **12**(4): 1099-1117. https://doi.org/10.1016/j.joi.2018.09.004

[23] Bellman, R. 1957 *Dynamic programming*. Princeton University Press, Princeton, NJ.

[24] Nieminen, P., Polonen, I., & Sipola, T. (2013) Research literature clustering using diffusion maps. *Journal of Informetrics*, **7**(4): 874-886. https://doi.org/10.1016/j.joi.2013.08.004