Evolution of topics in digital library research: Word embedding and clustering to reveal research trends

AUTHORS SECTION

Coble, Zach University of Missouri, USA| coblezc@missouri.edu

Cho, Hyerim University of Missouri, USA | hyerimcho@missouri.edu

ABSTRACT

The history of information science and technology in recent decades is not complete without understanding the role of digital libraries. To understand the evolution of research topics in the digital library literature, this study employs unsupervised machine learning methods to analyze 16,689 abstracts from articles published between 1996-2023 and indexed in the Library and Information Science Source database. Word embedding was performed on the abstracts, and the resulting vectors were divided into 7-year sub-corpora. After dimension reduction, the sub-corpora was clustered, and topics were extracted. A content analysis approach was used to identify the topics and to discuss their evolution over time. Clearly recognizable clusters emerged, along with important topical changes, most notably in the clusters corresponding to electronic resources, metadata, management, and digitization.

KEYWORDS

digital libraries, machine learning, topic analysis, word embedding, bibliometrics

INTRODUCTION

Since emerging in the 1990s as a sub-field of LIS, research on digital libraries has advanced rapidly. Arms (2001) 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" (p. 2). 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 (Cleveland, 1998). Research in digital libraries has brought together a unique combination of related disciplines, including librarianship, computer science and information retrieval, and legal and economic aspects of publishing (Tang et al., 2012; Wolfram, 2016). Simultaneously, bibliometric researchers have leveraged recent technical advances in natural language processing to create opportunities for computational model-based bibliometric analysis, such as topic modeling and clustering (Sugimoto et al., 2010; Zhang et al., 2018). To understand the scientific trend and development in digital libraries, the current study examines how research topics in the digital library literature have evolved from 1996 to 2023, as represented in article abstracts indexed in the Library and Information Science Source database. Particularly, the following research question will be answered: "What are the identified research topics and keywords in digital libraries between 1996 and 2023? What have been the noticeable changes within the research topics in digital libraries?"

LITERATURE REVIEW

Latent Dirichlet allocation (LDA) is a model-based method for topic modeling that has been widely applied to the academic and LIS literature to understand the evolution of research themes (Blei et al., 2003; Sugimoto et al., 2010; Figuerola et al., 2017). 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 (Mikolov et al., 2013; Gastaldi, 2021). Nieminen et al. (2013) established that when applied to abstracts from the academic literature, output from the combination of word embedding and clustering corresponds to topics, and Zhang et al. (2018) further found that k-means clustering was the most effective clustering technique for word embeddings of article abstracts. Gao and colleagues (2022) then combined topic modeling and word embedding techniques to achieve a more nuanced analysis of topics in the LIS literature. There have been several studies that conducted bibliometric analyses within the digital library literature. For example, Ahmad et al. (2018) examined productivity metrics from 4,206 documents from Web of Science published between 2002-2016, and Singh et al. (2017) analyzed 1,000 digital library research articles from 1998-2004 to investigate the publication trend (*e.g.,* single authors VS multiple authors, popular publication venues) and the authors' productivity. Although these studies provide important metrics about the research output, there has been a lack of research regarding identifying the major topics in digital library research and how they have evolved.

METHODS

The current study follows a conventional machine learning model (Goodfellow, Bengio, & Courville, 2016) using natural language processing techniques in Python. The abstracts from the database search were preprocessed using a spaCy pipeline and then used to generate a word embedding model using Gensim Word2vec. Dimension reduction was applied to the model to improve performance, and the resulting vector representations were divided into seven sub-corpora of abstracts from seven-year time increments (*e.g.*, 1996-2003, 2003-2010). The scikit-learn k-means clustering algorithm (scikit-learn developers, 2024) was then applied to each of the sub-corpora to extract topics for

the seven-year time increments. Topics for each cluster were manually labeled based on the forty most representative keywords and the ten most representative abstracts, where representative was defined as nearest to the cluster centroid (Řehůřek, 2022). In this step, a content analysis approach was utilized (Chuang et al., 2015).

RESULTS

In total, 41 clusters were created, with an average of ten clusters per seven-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 (Shahapure & Nicholas, 2020), which averages each document's individual Silhouette score. The Silhouette coefficients ranged between 0.08-0.10, suggesting weak clusters.

Table 1: Silhouette Coefficient Scores for Date Ranges

Years	*Size*	*Total clusters,* k	*Silhouette coefficient*
1996-2003	2,631	8	0.08
2003-2010	7,577	11	0.08
2010-2017	7,089	12	0.10
2017-2023	4,925	10	0.10

While the Silhouette scores were relatively low, clear topics emerged from the clusters. *Table 2* provides a summary of the most prominent topics. The labeled topics were created by the authors, based on the keywords generated during clustering. The notable changes are based on changes over time in keywords associated with labeled topics.

The topic of management was consistently present, and the corresponding abstracts clearly reflect the evolution of the wider field. For instance, the most representative abstract from 1993-2003 (Moore et al., 2003) begins by asking, "What is the organizational impact of becoming a digital library, as well as a physical entity with facilities and collections? Is the digital library an add-on or an integrated component of the overall library package?" (p. 14). This contemplation of integrating physical and digital services stands in contrast to the most representative abstract from 2017-2023 (Cox, 2017): "National University of Ireland Galway digitized the archive of the Abbey Theatre between 2012 and 2015...It has had a transformative effect on the Library as leader of the project. The role of the archivist has changed and partnerships with the academic community have strengthened" (p. 20). These two abstracts encapsulate the growth and maturity of digital library research during this period, from pondering potential transformations to case studies summarizing exactly how they occurred.

Electronic resources were the most prevalent topic. The topic was initially almost exclusively about journals and had minimal mention of ebooks, although by 2017-2023 it had switched to being predominately about ebooks. The conference summaries topic, as well as the announcements and project reports topic, appeared to cluster around similarities in genre rather than subject. For instance, most abstracts for announcements and project reports were shorter, typically around 50 words, and had a similar formulaic structure, regardless of the publication venue.

Table 2: Evolution of Most Prominent Topics

Labeled topic	*Keywords*	*Notable changes*
Electronic resources	serial, journal, ebook, acquire, subscription, demand, electronic	Shift from journals to ebooks.
Metadata	metadata, data, interoperability, discovery, data science, catalog	Always present but never dominant. Combined with archives (2003-2010), data science (2010-2017), and interoperability (2017-2023).
Management	manage, future, development, policy, organization, offering, sustain	Shift from early keywords reflecting new initiatives (offering, expand) to later ones reflecting maturity (future, sustain)
Digitization	digital, collection, digitization, local, preservation, repository, content	Not clearly present until 2003-2010. Initially with access/interfaces, then metadata/discovery, then archives/processing.
Conference summaries	british, norwegian, scotland, national, seminar, conference, congress	More genre than topic. Most consistent cluster over time. Primarily city and country names.
Announcements & project reports	library, development, service, staff, provision, increasingly, consortia, offering	More genre than topic. Miscellaneous updates, website reviews, and scene reports. Consistent due to formulaic abstract structure

CONCLUSION

The study contributes a macro-level analysis of digital library research topics and their evolution over time. Areas for future study include comparison with topics extracted using the current state of the art attention-based large language models (Grootendors, 2022). The study support research in information science by providing insight into the major research themes in digital libraries, which is essential to understanding the history and development of information science and technology during a transformative era.

GENERATIVE AI USE

We confirm that we did not use generative AI tools/services to author this submission.

REFERENCES

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. https://doi.org/10.1108/EL-02-2017-0036

Arms, W.Y. (2001). *Digital Libraries*. Cambridge, MA: MIT Press.

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

Chuang, J., Roberts, M.E., Stewart, M.E., et al. (2015). TopicCheck: Interactive alignment for assessing topic model stability. Human Language Technologies: The 2015 Annual Conference of the North American Chapter of the ACL. https://aclanthology.org/N15-1018.pdf

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

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. https://doi.org/10.1007/s11192-017-2432-9

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

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

Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep Learning*. Cambridge, MA: MIT Press.

Grootendors, M. (2022). BERTopic: Neural topic modeling with a class-based TF-IDF procedure. *arXiv*. https://doi.org/10.48550/arXiv.2203.05794

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

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

Moore, M.E., Garrison, S., Hayes, B., & McLendon, W. (2003). Reinventing a health sciences digital library--organizational impact. Medical Reference Services Quarterly, 22(4), 75–82. https://doi.org/10.1300/J115v22n04_08

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

Sugimoto, C. R., Li, D., Russell, T. G., Finlay, S. C., & Ding, Y. (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*, 62(1), 185–204. DOI: 10.1002/asi.21435

scikit-learn developers (2024). *2.3. Clustering*. https://scikit-learn.org/stable/modules/clustering.html

Shahapure, K. R., & Nicholas, C. (2020). Cluster Quality Analysis Using Silhouette Score. *2020 IEEE 7th International Conference on Data Science and Advanced Analytics*, 747-748. https://doi.org/10.1109/DSAA49011.2020.00096

Tang, J., Fong, A.C.M., Wang, B. & Zhang, J. (2012). A unified probabilistic framework for name disambiguation in digital library. *IEEE Transactions on Knowledge and Data Engineering*, 24(6), 975-987. https://doi.org/10.1109/TKDE.2011.13

Wolfram, D. (2016). Bibliometrics, information retrieval and natural language processing: Natural synergies to support digital library research. *BIRNDL 2016 Joint Workshop on Bibliometric-enhanced Information Retrieval and NLP for Digital Libraries*. https://aclanthology.org/W16-1501.pdf

Zhang, Y., Lu, J., Feng, L., Liu, Q., Porter, A., Chen, H, Zhang, G. (2018). Does deep learning help topic extraction? A kernel k-means clustering method with word embedding. J*ournal of Informetrics*, 12(4), 1099-1117. https://doi.org/10.1016/j.joi.2018.09.004