Informology
Online ISSN: 2999 - 9790
Homepage: www.informology.org

The Use of Artificial Intelligence in Knowledge Organization and Subject Indexing

Alireza Noruzi

Editor-in-Chief, Associate Professor, Informology Center, Marseille, France. E-mail: anoruzi@gmail.com

Article Info

Article type:
Editorial note

Keywords:
artificial intelligence, knowledge organization, cataloging, classification, indexing.

ABSTRACT

Objective: The library, once a silent repository of knowledge, is undergoing a digital metamorphosis. At the heart of this transformation is the integration of artificial intelligence (AI). Traditionally, cataloging and classifying books was a meticulous, labor-intensive task requiring deep subject knowledge and adherence to complex classification systems. However, the advent of AI is revolutionizing this process, promising greater efficiency, accuracy, and accessibility. This research presents a pilot study on the potential use of AI for semi-automatic subject indexing, cataloging, and classification of books.

Materials and Methods: To collect the data, first we searched for open access books in the Directory of Open Access Books and OAPEN (Online library of open access books). Four books in English were chosen for the current study. The book titles were searched in the Library of Congress Online Catalog to discover the subject headings, the Library of Congress Classification number and the Dewey Decimal Classification number assigned to each book, collecting the necessary information. Third, we searched for each book in ChatGPT, Copilot, and Gemini, using necessary prompts to collect the data.

Results: The results indicate that the degree of cataloging and classification consistency is low. The cataloging and classification consistency are seen as the measure of the similarity of reaction of different machines and human beings processing the same book.

Conclusion: By automating routine tasks, improving classification and cataloging accuracy, and enhancing metadata creation, AI is transforming the way libraries organize and share knowledge. As technology continues to advance, we can expect even more exciting developments in this field.

Cite this article: Noruzi, A. (2024). The use of artificial intelligence in knowledge organization and subject indexing. *Informology*, 3(1), 1-8.

Publisher: Informology Center.

Introduction

The field of library and science (LIS), once predominantly reliant on human expertise and experience, is undergoing a transformative shift with the integration of artificial intelligence (AI). Traditionally, knowledge organization (KO), subject indexing, cataloging and classification of books was a repetitive, labor-intensive and expensive process (University of Toronto Libraries, n.d.), requiring extensive knowledge of subject matter, classification systems (e.g., the Dewey Decimal Classification and the Library of Congress Classification (LCC) systems), the Anglo-American Cataloguing Rules, and meticulous attention to detail of each book. However, the advent of AI is revolutionizing this field of LIS by automating tasks, enhancing accuracy, and unlocking new possibilities for knowledge organization and subject indexing.

One of the most significant applications of AI in knowledge organization and cataloging is the automation of routine tasks. Optical Character Recognition (OCR) technology, powered by AI and Natural Language Processing (NLP), can accurately extract text from book covers, title pages, table of contents, and the contents of documents. The data can then be used to generate preliminary bibliographic records, significantly reducing the time-consuming and repetitive manual data entry process. Additionally, AI can automatically assign and construct subject headings, keywords, notations, class codes, and summaries based on text analysis, saving catalogers and indexers valuable time and effort. By analyzing the content of a book or document, AI algorithms can suggest appropriate terms, descriptors, or classification numbers from controlled vocabularies, thesauri, and classification systems, such as the Library of Congress Subject Headings (LCSH), Medical Subject Headings (MeSH), the Library of Congress Classification (LCC), or the Dewey Decimal Classification (DDC). This not only saves catalogers time but also ensures consistency and accuracy in subject assignment.

Thesauri (e.g., the AGROVOC Multilingual Thesaurus, the UNESCO Thesaurus, the ERIC Thesaurus, etc.), subject headings (such as LCSH, MeSH), and library classification systems (like DDC, and LCC) are complex and require expert knowledge and relevant professional experience. AI helps to improve indexing, cataloging and classification accuracy by analyzing the content of documents and assigning appropriate subject headings, descriptors and class numbers. Machine learning algorithms can identify patterns and relationships between documents, leading to more precise and consistent classification and clustering. This not only benefits library users by improving search results but also aids in collection management and analysis. NLP techniques allow AI systems to understand the nuances of human language, enabling them to extract relevant concepts and terms from text. This capability is particularly valuable for complex or abstract subjects, where human indexing can be challenging. Furthermore, AI can analyze large datasets of bibliographic records to identify patterns and bibliographic relationships between subjects to implement the Functional Requirements for Bibliographic Records (FRBR) model. This

information can be used to refine subject indexing schemes and create more effective subject hierarchies. For instance, AI can suggest new subject headings, descriptors, classification number, and clusters, or identify gaps in the existing vocabulary. Furthermore, metadata (the data about data) is crucial for discoverability of books and documents. AI can assist in creating rich and informative metadata by extracting relevant details from the content. For example, AI can identify authors, publication dates, and even summarize the content. This enhanced metadata can improve search engine optimization (SEO) and make library collections more visible online.

A Generative Pre-trained (GPT) artificial intelligence (AI) model for KO (so-called KOGPT) can be developed and used in information representation and knowledge organization systems such as cataloging, classification and indexing. It can be designed to engage in dialogue with catalogers, classifiers, and indexers in a manner that feels natural and human-like. Like other GPT models, KOGPT can be trained on a vast amount of text data of books and articles to understand language patterns and generate coherent and consistent responses (subject headings, descriptors, index terms). A Generative Pre-trained (GPT) information representation system can read a book, document, or article, analyzing its subject (doing subject analysis), and assign relevant keywords, index terms, subject headings, and class numbers to it (indexing, cataloging and classification). It is worth noting that Generative Artificial Intelligence (GenAI) features prominently for those in the book cataloging and classification of library materials as they add new tools to their knowledge organization kits. However, given GenAI's novelty, much of the conversation about its use in book cataloging, classification, and subject indexing is speculative. GenAI refers to an artificial intelligence system capable of understanding, reasoning, learning, and interacting with extremely large datasets, including the Large Language Models (LLMs). It uses statistical inferences to analyze the relationship between words in a body of text or pixels from an image. From these inferences, GenAI systems can generate human-like content (e.g., book cataloging, classification, and subject indexing) quickly in response to human-provided prompts.

To better understand major trends and common concerns – such as generative AI's role in the cataloging, classification and indexing, this note draws on experimental observations. This exploration delves into the burgeoning intersection of AI and library science, examining how intelligent systems are being harnessed to streamline cataloging processes, improve classification precision, and ultimately enhance the discoverability of library collections.

Materials and Methods

This research presents a pilot study on the potential use of AI for semi-automatic subject indexing, cataloging, and classification of books. The accuracy of the assigned subject headings and classification numbers is compared with the Library of Congress online catalog.

To collect the data, first we searched for open access books in the Directory of Open Access Books (https://www.doabooks.org) and OAPEN (Online library of open access books: https://www.oapen.org). The following four books in English were chosen for the current study: 1. *Entangled Brain: How Perception, Cognition, and Emotion Are Woven Together* by Luiz Pessoa; 2. *Australian Indigenous Knowledge and Libraries* edited by Martin Nakata and Marcia Langton; 3. *Authority Control* edited by Mauro Guerrini and Barbara B. Tillet; and 4. *Digital Humanities and Libraries and Archives in Religious Studies* edited by Clifford B. Anderson. Second, the book titles were searched in the Library of Congress Online Catalog (https://catalog.loc.gov/) to discover the subject headings, the Library of Congress Classification number and the Dewey Decimal Classification number assigned to each book, collecting the necessary information. Third, using three prompts like the following search queries, we searched for each book in ChatGPT (https://chatgpt.com), Copilot (https://copilot.bing.com), and Gemini (https://gemini.google.com):

1. Read the book entitled "ABCD" via the following URLs https://abcd and https://abcd and assign four keywords to it based on the Library of Congress Subject Headings available at https://www.abcd.
2. Assign a class number based on the Library of Congress Classification system accessible via https://www.abcd to the book entitled "ABCD" available at https://www.abcd.
3. Assign a class number based on the Dewey Decimal Classification system accessible via https://www.abcd to the book entitled "ABCD" available at https://www.abcd.

Results

Table 1 shows the subject headings and keywords assigned to the selected books in the Library of Congress online catalog, ChatGPT, Copilot, and Gemini.

Table 1. Subject headings assigned to the books

Books	LC Subjects assigned in the LC Catalog	Copilot	ChatGPT	Gemini
Entangled Brain: How Perception, Cognition, and Emotion Are Woven Together	1. Perception. 2. Emotions and cognition. 3. Brain. 4. Neuropsychology.	1. Brain Networks 2. Neuroscience 3. Cognitive Processing 4. Emotional Integration	1. Neurosciences 2. Brain 3. Cognition 4. Emotions	Unable.
Australian Indigenous Knowledge and Libraries	1. Libraries and Aboriginal Australians--Congresses.	1. Indigenous Knowledge Documentation 2. Cultural Protocols 3. Access and Control 4. Libraries and Aboriginal Australians	1. Indigenous peoples 2. Knowledge management 3. Libraries and Indigenous people 4. Australia	1. Libraries and Aboriginal Australians 2. Indigenous knowledge - Australia 3. Aboriginal Australians - Archives 4. Information services – Australia

Authority Control	1. Authority files (Information retrieval) -- Congresses.	1. Authority Control 2. Subject Authority File 3. Controlled Vocabulary 4. Thesaurus and Controlled Terms	1. Authority files (Information retrieval) 2. Cataloging 3. Names, Personal -- Authority files 4. Library catalogues	Unable.
Digital Humanities and Libraries and Archives in Religious Studies	1. Academic libraries-- Effect of technological innovations on. 2. Digital humanities-- Religious aspects. 3. Theological libraries--Administration. 4. Theology--Research. 5. Theology--Study and teaching.	1. Digital Humanities 2. Libraries and Archives 3. Subject Access 4. Religious Studies 5. Innovative Methods	1. Digital humanities 2. Religious studies 3. Libraries and archives 4. Theology 5. Information technology	1. Digital humanities 2. Religious studies - Digital libraries 3. Religious studies - Archives - Digital formats 4. Information specialists - Collaboration with religious scholars 5. Religious studies - Research - Technological innovations

Table 2 indicates the Library of Congress (LC) class number assigned to the selected books in the Library of Congress online catalog, ChatGPT, Copilot, and Gemini.

Table 2. LC class numbers assigned to the books

Books	**LC call number assigned in the LC Catalog**	**Copilot**	**ChatGPT**	**Gemini**
Entangled Brain: How Perception, Cognition, and Emotion Are Woven Together	BF311	Q – Science	QP385	Unable.
Australian Indigenous Knowledge and Libraries	Z711.92.A24	Z – Bibliography. Library Science. Information Resources (General)	Z711.92.A8	Z 675 .A8 A87 2005
Authority Control	Z693.3.A88	Z – Bibliography. Library Science. Information Resources (General)	Z695.24	Unable.
Digital Humanities and Libraries and Archives in Religious Studies	AZ195	Z – Bibliography. Library Science. Information Resources (General)	BL 256 BL – Religion 256 – Digital Humanities	Unable.

Table 3 addresses the Dewey Decimal Classification (DDC) class number assigned to the selected books in the Library of Congress online catalog, ChatGPT, Copilot, and Gemini.

Table 3. DDC class numbers assigned to the books

Books	DDC call number assigned in the LC Catalog	Gemini	Copilot	ChatGPT
Entangled Brain: How Perception, Cognition, and Emotion Are Woven Together	153	Based on my search the DDC for "The Entangled Brain: How Perception, Cognition, and Emotion Are Woven Together" is likely to be in the 150 range (Psychology) or 612.8 (Human neuroanatomy)	153	612.82
Australian Indigenous Knowledge and Libraries	N.A.	Couldn't find the DDC number for this book.	305.89915 – Indigenous Peoples and Cultures in Australia	027.70899915
Authority Control	025.3/222	025.33	025.33	025.3
Digital Humanities and Libraries and Archives in Religious Studies	001.30285	025.04.	Recommend assigning it to: Library and Information Sciences / Museology / Museology and Heritage Studies / Religion: General	006.3

It should be noted that we did not use Robert Hooper's measure of indexing consistency to calculate the consistency percentages, because a quick look at the tables indicate that the degree of cataloging (subject headings) and classification (class numbers) consistency is low. The cataloging and classification consistency are seen as the measure of the similarity of reaction of different machines and human beings processing the same book. Hooper's measure indicates the percentage consistency of indexing (Hooper, 1965). It is calculated as follows:

$$\text{Consistency (A, B)} = \frac{i}{i + j + k}$$

where:

(i) represents the number of consistent index terms.
(j) represents the number of inconsistent index terms.
(k) represents the number of missing index terms.

Discussion and Conclusion

The traditional methods of cataloging, classification and subject indexing, once the cornerstone of library organization, are undergoing a significant transformation due to the advent of artificial intelligence (AI). This powerful technology is poised to revolutionize how libraries organize and access their collections. While AI offers immense potential, challenges remain. Ensuring the cataloging, classification, and indexing consistency is paramount. The cataloging and indexing

consistency are seen as the measure of the similarity of reaction of different machines and human beings processing the same document. The catalogers, indexers and AI machine algorithms act considerably different (among themselves and with others) in their '*aboutness*' judgment as to which subject headings, descriptors, or terms reflect the contents of the document most adequately (Zunde & Dexter, 1969). Accurate information representation is the most important input function of the information retrieval system. Additionally, AI systems must be trained on high-quality data to produce accurate results. Human oversight and intervention will continue to be necessary to address complex cases and ensure the quality of subject cataloging and indexing. Looking ahead, AI is poised to play an even more significant role in indexing, cataloging and classification. Natural language processing (NLP) will enable more sophisticated text analysis, leading to better subject assignment and improved search capabilities. AI-powered recommendation systems can suggest books and resources to users based on their reading preferences, enhancing the overall library experience.

In addition, the generative AI tools for knowledge organization and subject indexing are developing and might provide much better results in the future. However, their results (suggested index terms, descriptors, subject headings, and class numbers) are not consistent enough and need to be constantly trained to improve. The output of AI chatbots like ChatGPT, Copilot, and Gemini can be improved by giving (human) feedback "to produce more consistently useful or correct results." (Thelwall, 2024).

Furthermore, the complexity of human language and the evolving nature of knowledge can pose difficulties for AI systems. Despite these challenges, the future of cataloging and subject indexing is bright. AI has the potential to transform libraries into more efficient, user-centric organizations. By automating routine tasks and improving the accuracy of subject indexing and cataloging, AI can enhance the discoverability of library collections and facilitate research. In conclusion, the integration of AI in cataloging, classification, and subject indexing is a significant development that promises to reshape the library landscape. As technology continues to advance, we can expect to see even more innovative applications of AI in this field.

Data Availability Statement

Not applicable.

Ethical considerations

The author avoided from data fabrication and falsification.

Funding

Not applicable.

References

Golub, K., Wang, J., & Widegren, J. (2024). Using ChatGPT for (semi-) automatic subject indexing of different document types. *Digital Humanities in the Nordic and Baltic Countries: DHNB*. https://www.diva-portal.org/smash/record.jsf?pid=diva2:1859349

Hooper, R. (1965). *Indexer consistency tests — Origin, measurement, results, and utilization*. Bethesda, Md.: IBM Corporation; 1965.

Thelwall, M. (2024). *Can ChatGPT evaluate research quality*? Retrieved 3 Apr 2024, from https://arxiv.org/abs/2402.05519

University of Toronto Libraries. (n.d.). Organization of the Collection & Card Catalogues. University of Toronto Libraries. https://exhibits.library.utoronto.ca/exhibits/show/utl125/access-and-discovery/organization-card-catalogues

Zunde, P., & Dexter, M. E. (1969). Indexing consistency and quality. *American Documentation*, 20(3), 259-267.