Integrating LLMs and RAG for Improved Usability in Multimedia Archives

Sean Carroll
De Montfort University
Leicester, England
P2546407@dmu.ac.uk

Abstract

This paper explores the roles of generative AI technologies, such as Large Language Models (LLMs) and Retrieval Augmentation Generation (RAG) systems to enhance multimedia archives. In it, archives are redefined as dynamic entities for knowledge construction and validation, focusing on improving cultural accessibility and interaction. The paper addresses the challenges in archiving new media art, emphasising how semantic embeddings and LLMs make user interactions with archives more intuitive and accessible. The practical implementation of these technologies is explored in the Computer Art Archive's pilot program, which incorporates this technology into systems akin to current knowledge retrieval frameworks. The paper concludes by discussing how these advancements democratise knowledge access and contribute to the enrichment of digital cultural heritage.

Keywords

Multimedia, Archives, AI, LLMs, RAG, Semantics, Accessibility, Archiving, Chatbot, Heritage

Introduction

This paper delves into the transformative potential of advanced AI technologies, particularly Large Language Models (LLMs) and Retrieval Augmentation Generation (RAG) systems, in enhancing the accessibility and interaction within multimedia archives. The study begins by examining how these technologies can shift the paradigm from traditional search-centric methods to a more intuitive, engagement-driven approach in archiving. It then explores the implementation of semantic embedding and LLMs to navigate the complexities of new media art archiving, addressing challenges in metadata categorization and user accessibility.

The New Media Archive

Featherstone [1] characterises the archive as a 'place for the storage of documents and records… the storehouse for the material from which memories [are] constructed,' while Osborne [2] views it as a 'means of generating ethical and epistemological credibility.' Combined, these perspectives position the archive as more than a mere repository; it emerges as a crucial instrument in the construction and validation of knowledge for its stakeholders. This role involves determining what knowledge is vital for the archive, effectively retaining it, and supporting the utilisation of this stored knowledge.

Bourdieu's [3] observations add a critical dimension to our understanding of archives. He notes, '...the inheritance of cultural wealth that has been accumulated and bequeathed by previous generations only really belongs (although it is theoretically offered to everyone) to those with the means of appropriating it for themselves.' This statement underscores the need for archives to go beyond simple search functionalities, advocating for a paradigm shift towards deeper engagement and broader accessibility. It highlights the importance of making archives not just as repositories of knowledge but also accessible platforms where cultural heritage can be meaningfully appropriated by people from diverse backgrounds.

At ISEA 2022, Mitchell et al. [4] presented their findings on the challenges of archive connectivity, focusing on the need for standardised taxonomies in organising archival data. Their discussion emphasised that archival data, often extensive and varied in structure, faces significant categorisation challenges due to the lack of uniformity in metadata and taxonomies. This standardisation is crucial for organising and effectively utilising vast amounts of information.

However, as Ippolito [5] points out in his essay 'Death By Wall Label,' this task is further complicated when considering the diverse nature of new media artworks. He critically examines the traditional archival methodologies applied to new media art, focusing on the static nature of cataloguing systems. He contends that these conventional practices fail to capture the inherently dynamic and collaborative essence of new media artworks. Building on this argument, Cepeda [6] delves into the multifaceted nature of new media art archiving. Cepeda emphasises the necessity of archiving not just the complete artworks but also their individual components, such as textures, videos, 3D meshes, and audio files, to facilitate future restorations or reinterpretations. These perspectives provide a deeper understanding of the categorisation challenges that Mitchell et al. identified at ISEA 2022.

Integration of Large Language Models

Reporting on research that aims to build on these discussions, preliminary findings identify how emerging AI technologies can enhance the accessibility and capacity of New Media Archives, specifically aiming to make knowledge available not only to academics but also to curators and artists. To further understand this context, Whitelaw's [7] critical analysis of digital cultural collections in his paper on 'Generous interfaces' challenges the constrained nature of traditional search interfaces, which typically prompt users for precise queries and often yield limited outputs. Advocating for a radical transformation, Whitelaw proposes the design of interfaces that are richly informative and inherently navigable, inviting users to engage with the full spectrum of an archive's content. Enhanced by advanced Large Language Models (LLMs), conversational interfaces have potential to revolutionise archive accessibility, accommodating a range of linguistic styles from specialised academic language to general and beginner terms. These interfaces allow users to intuitively describe their search needs without relying on precise terminology, democratising cultural access in line with Bourdieu's [3] ideals and resonating with Whitelaw's [7] vision of informative and navigable interfaces. This fosters a broader, more meaningful engagement with archival content.

Complementing this, the technical backbone of these interfaces is often a Retrieval Augmentation Generation (RAG) system with semantic embedding which underpins the ability to handle complex queries [8]. A RAG system translates user queries into a vector, capturing the nuances of language as points in high-dimensional space. It then matches these vectors against a pre-indexed map of the archive's content, represented similarly, ensuring that responses are not only relevant but also semantically aligned with the user's intent. By integrating these technologies, the archive can effectively draw upon its vast knowledge base, providing responses that deeply resonate with and accurately reflect users' inquiries, and simultaneously avoid issues associated with LLMs such as hallucination.

The RAG system's nuanced approach directly addresses Whitelaw's [7] and Ippolito's [5] critiques concerning traditional archival methods. While Whitelaw highlights the constraints of conventional search interfaces, the RAG system's semantic matching capabilities offer a more contextual and intent-driven search experience. This not only aligns with Whitelaw's call for richer interfaces but also mitigates the issues of metadata overload and lack of detail, as outlined by Mitchell et al. [4] and Ippolito [5]. By proficiently handling queries even in the absence of direct matches, a RAG system presents a more dynamic and inclusive approach to archival interaction.

The incorporation of semantic vector embeddings in new media art archives is a significant progression of longstanding research in semantic technologies, as exemplified by the Semantic Map project by Fleischmann & Strauss [9]. By leveraging semantic vector embeddings, it is possible to unlock an unprecedented level of accessibility, and perhaps even more critically, adaptability. Adaptability is crucial in facilitating a dynamic and collaborative development process. The utilisation of semantic embeddings provides a robust foundation for sharing access to databases. It simplifies processes like forking—creating branches from the main body of content—that allow for parallel development and innovation. This approach also simplifies version control and backup procedures, laying the groundwork for creative engagements such as interface hackathons where participants can reimagine ways to interact with archival content.

In parallel, there is a concurrent evolution in API development, utilising state-of-the-art tools such as LangChain, Pinecone, Canopy, or even Large Language Models (LLMs) to self-code tools, thereby lowering the barriers to technological entry [10]. This democratisation of development extends the capability to innovate within the archival domain to a broader community, enabling a collective contribution to the enrichment of digital cultural heritage.

Computer Arts Archive

The Computer Arts Archive is a nonprofit dedicated to preserving and promoting computer arts, serving artists, audiences, curators, educators, and researchers. Drawing on literature, and to test user experience, an AI model has been built that leverages the Information Search Process (ISP) model [11] and Dervin's [12] Sense-Making Methodology to shift the archive's focus from search-centric to engagement-centric. Personalised pathways are offered based on user preferences, aligning with the initiation phase of the ISP model [11]. The program utilises a context-aware chatbot, grounded in Dervin's principles, to ask users probing questions, enriching the understanding of their goals and needs. These interactions are crucial, as the questions and user responses are fed back into the RAG system [8]. This feedback has been used to refine the semantic embeddings created from user searches, to date enhancing the system's ability to generate increasingly relevant and contextually appropriate prompts and recommendations.

The next phase of investigation will utilise Large Language Models (LLMs) enhanced with semantically retrieved knowledge to support users during the ideation phase, drawing upon the formulation stage of Kuhlthau's Information Search Process (ISP) model [11]. The LLM comes into play by analysing the themes and content of the items collected by the users. Based on this analysis, the LLM engages in a contextually relevant conversation, thus aiding users in developing their creative concepts. This interactive process is designed to enhance the user's ability to formulate nuanced ideas and engage deeply with the archival content, embodying the essence of cultural capital through the active participation of the user in the creative process.

It is anticipated that a summary of findings from these phases of the investigation will be presented at ISEA 2024.

Summary

The paper acknowledges the significant advancements made in new media art archiving, while also identifying the

need for ongoing development, particularly in deepening engagement with artists and curators. It proposes an approach that leverages AI technologies to enhance the accessibility and interactivity of archives like the Computer Art Archive, aiming to create more dynamic connections between the archive, its contents, and a diverse audience. Furthermore, the paper advocates for the development of knowledge retrieval frameworks and the establishment of a flexible ecosystem of APIs and low-code solutions. This strategy is intended to facilitate experimentation with these technologies in archival settings at a relatively low cost. By doing so, the paper aims to encourage the use of advanced AI technologies in the archival field, promoting more inclusive and innovative interactions with archival content.

References:

[1] Featherstone, M. (2006). Archive. Theory, Culture & Society, 23(2–3), 591–596. https://doi.org/10.1177/0263276406023002106

[2] Osborne, T. (1999). The ordinariness of the archive. History of the Human Sciences, 12(2), 51–64. https://doi.org/10.1177/09526959922120243

[3] Bourdieu, P. (1973). 'Cultural Reproduction and Social Reproduction.' In Knowledge, Education and Social Change: Papers in the Sociology of Education, edited by R. Brown, 71–112. Tavistock, UK: Tavistock Publications.

[4] Mitchell, B., Searleman, J. T., van der Plas, W., & Wong, T. C. W. (2022). Connecting new media art archives worldwide. In ISEA 2022 Proceedings. Accessed on 05/11/23 from: https://isea2022.isea-international.org/wp-content/uploads/2023/02/ISEA2022-BCN-Proceedings.pdf

[5] Ippolito, J. (2008). Death by Wall Label. Accessed on 22/10/23 from: http://thoughtmesh.net/publish/printable.php?id=11

[6] Cepeda, R. G. (2019). Rescuing new media art from technological obsolescence. DAT Journal, 4(3), 37–46. https://doi.org/10.29147/dat.v4i3.145

[7] Whitelaw, M. (2015). Generous Interfaces for Digital Cultural Collections. Accessed 16/10/23 from: https://www.digitalhumanities.org/dhq/vol/9/1/000205/000205.html

[8] Lewis et al. (2020). Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. Accessed on 17/10/23 from: https://arxiv.org/abs/2005.11401

[9] Fleischmann, M., & Strauss, W. (2023). Exploring the Digital Archive as a Thinking Space – AI Aspects on Documentation, Access and Knowledge Discovery. In ISEA 2023 Proceedings. Accessed on 07/11/23 from: https://isea-archives.org/docs/2023/3rd_SNMAA_Provisional_Proceedings.pdf

[10] "Canopy RAG Framework." Pinecone. Accessed from: https://www.pinecone.io/blog/canopy-rag-framework/

[11] Kuhlthau, Carol, Maniotes, Leslie K., and Caspari, Ann K. (2007). Guided Inquiry: Learning in the 21st Century. Westport, Connecticut: Libraries Unlimited.

[12] Dervin, Brenda, Foreman-Wernet, Lois, and Lauterbach, Eric (Eds.). (2003). Sense-making methodology reader: selected writings of Brenda Dervin. Cresskill, N.J.: Hampton Press.