CHALLENGES IN IMPLEMENTING A RECOMMENDER SYSTEM FOR HISTORICAL RESEARCH IN THE HUMANITIES

Florian Atzenhofer-Baumgartner[1,2]   Bernhard C. Geiger[3,4]   Christoph Trattner[5]   Georg Vogeler[2]
Dominik Kowald[1,3]

[1]Institute of Interactive Systems and Data Science, Graz University of Technology, Graz, Austria
[2]Department of Digital Humanities, University of Graz, Graz, Austria
[3]Know Center Research GmbH, Graz, Austria
[4]Signal Processing and Speech Communication Laboratory, Graz University of Technology, Graz, Austria
[5]University of Bergen & Media Futures, Bergen, Norway

ABSTRACT

This extended abstract describes the challenges in implementing recommender systems for digital archives in the humanities, focusing on Monasterium.net, a platform for historical legal documents. We discuss three key aspects: (i) the unique characteristics of so-called charters as items for recommendation, (ii) the complex multi-stakeholder environment, and (iii) the distinct information-seeking behavior of scholars in the humanities. By examining these factors, we aim to contribute to the development of more effective and tailored recommender systems for (digital) humanities research.

Keywords: Multistakeholder Recommender Systems, Digital Humanities, Historical Research, Information-Seeking Behavior

1   Introduction

Recommender systems (RecSys) are ubiquitous in various domains (e.g., entertainment and multimedia [1]), enhancing user experience and facilitating content discovery. However, their application in digital humanities venues is underutilized and presents unique challenges. This work focuses on Monasterium.net, a digital archive for charters, to explore three main challenges (see Sections 2 – 4) and propose potential directions for future research (see Section 5).

2   Charters as Unique Items for Recommender Systems

Charters, the primary content of Monasterium.net, have distinct characteristics that set them apart from typical items in RecSys. These historical documents record legal actions and provide rich insights into past societies [2]. Unlike commercial products or entertainment content, charters are complex, multifaceted artifacts with varying levels of completeness and authenticity. In addition, they exhibit a specific type of public value as cultural heritage objects.

Charters in Monasterium.net are represented as digitized scans, scholarly editions, and semi-structured data using specialized schemas like the Charter Encoding Initiative (CEI) [3]. This representation includes diverse metadata such as material properties, production details, authentication methods, and scholarly annotations. The high dimensionality and often sparse population of this metadata pose significant challenges for traditional recommendation algorithms.

Furthermore, the historical nature of charters introduces additional complexities. Unlike continuously growing datasets in modern RecSys, the set of historical charters is finite, though new discoveries or digitizations may occasionally expand the corpus. Charters span different historical periods, requiring consideration of temporal relevance in recommendations.

Figure 1: Recto side of a charter in which five Cardinal priests grant a 100-day indulgence for St. John's Chapel in Mautern on the Danube (1475). The document, requested by local petitioners, details specific visitation requirements and contributions; it demonstrates typical medieval ecclesiastical documentation practices and authentication methods. Adapted from https://www.monasterium.net/mom/AT-StiAG/GoettweigOSB/1475_III_20/charter.

Some charters may be identified as forgeries long after their creation, affecting their relevance and interpretation. Many charters contain non-resolvable, missing references of entities, or damaged sections, which requires special handling in data representation and recommendation algorithms. Additionally, charters often contain text in various languages and scripts, which furthers complicates text-based recommendation approaches.

3   Multi-Stakeholder Environment in Digital Archives

The implementation of RecSys for Monasterium.net must consider a complex ecosystem of stakeholders in a mostly non-commercial setting, each with distinct values and objectives. Key stakeholders include researchers, primarily historians and other humanities scholars, who seek relevant documents for their studies. They value accuracy, comprehensiveness, and the ability to discover unexpected connections [4]. Content creators, including archivists and editors who contribute to the platform, aim for visibility and recognition of their work [5]. Platform owners, such as ICARus for Monasterium.net, focus on user satisfaction, engagement, and strategic growth. Funding agencies, who support the digitization and preservation of charters, value the extent of use and impact of the collections.

Balancing these diverse and sometimes conflicting interests presents a significant challenge. For example, recommendations that prioritize popular or well-documented charters might satisfy casual users but could limit the exposure of lesser-known documents that might be crucial for specialized research. Similarly, the needs of content creators for visibility might conflict with researchers' requirements for the most relevant documents, regardless of their popularity or recency of addition to the platform.

4   Information-Seeking Behavior in Humanities Research

The information-seeking behavior of humanities scholars differs significantly from the user behavior in commercial RecSys. For instance, historians often engage in exploratory searches, seeking to uncover unexpected connections and generate new research questions [6, 7]. Their research process is typically non-linear, involving iterative refinement

of queries and exploration of diverse paths, and thus, suggesting incorporating context information such as time or semantic context cues [8, 9].

Moreover, the concept of relevance in the humanities is multifaceted and context-dependent. A charter's relevance may not be immediately apparent and could emerge through the process of investigation. This characteristic challenges traditional relevance metrics used in RecSys. Connected to this, the long-term nature of humanities research also impacts user interaction with the system. Unlike in conventional RecSys, where immediate user feedback is common, the value of a recommendation in humanities research might only become apparent after extended periods of study.

5   Conclusion and Future Work

Implementing effective RecSys for digital archives like Monasterium.net requires addressing the unique nature of historical documents, navigating a complex multi-stakeholder environment, and accommodating the distinct information-seeking behavior of humanities scholars. Future work should focus on developing specialized algorithms that can handle the complexity and sparsity of charter metadata, designing evaluation frameworks that reflect the diverse stakeholder values, and creating interfaces that support the exploratory and long-term nature of humanities research. These endeavors meet the recent calls of the RecSys community for more detailed user studies and context-aware RecSys [10]. Also, by addressing these challenges, we can enhance the utility of digital archives, enable new discoveries and deeper understanding in historical research. This work not only benefits Monasterium.net, but also provides insights applicable to other digital humanities platforms and cultural heritage institutions [11].

References

[1] Dominik Kowald and Emanuel Lacic. Popularity bias in collaborative filtering-based multimedia recommender systems. In *International Workshop on Algorithmic Bias in Search and Recommendation*, pages 1–11. Springer, 2022.

[2] Philip Slavin. *The sources for manorial and rural history.*, page 131–148. Routledge, London, 1 edition, 2012.

[3] Georg Vogeler. Towards a standard of encoding medieval charters with XML. *Digital Scholarship in the Humanities*, 20(3):269–280, September 2005.

[4] Adam Nix, Stephanie Decker, David A. Kirsch, and Santhilata Kuppili Venkata. *Archival research in the digital era*, page 156–172. Edward Elgar Publishing, July 2023.

[5] Mathieu Andro and Imad Saleh. *Digital libraries and crowdsourcing: a review*, page 135–161. Wiley, January 2017.

[6] Kim Martin and Anabel Quan-Haase. The role of agency in historians' experiences of serendipity in physical and digital information environments. *Journal of Documentation*, 72(6):1008–1026, October 2016.

[7] Volkmar P Engerer. 'attention, attention, exploring minds acknowledge digital structure!' the shift to digital humanities has happened, so what should information scientists do in response? *Digital Scholarship in the Humanities*, 36(1):43–63, April 2021.

[8] Dominik Kowald, Simone Kopeinik, Paul Seitlinger, Tobias Ley, Dietrich Albert, and Christoph Trattner. Refining frequency-based tag reuse predictions by means of time and semantic context. In *Mining, Modeling, and Recommending'Things' in Social Media: 4th International Workshops*, pages 55–74. Springer, 2015.

[9] Dominik Kowald, Paul Seitlinger, Simone Kopeinik, Tobias Ley, and Christoph Trattner. Forgetting the words but remembering the meaning: Modeling forgetting in a verbal and semantic tag recommender. In *Mining, Modeling, and Recommending'Things' in Social Media: 4th International Workshops*, pages 75–95. Springer, 2015.

[10] Himan Abdollahpouri, Gediminas Adomavicius, Robin Burke, Ido Guy, Dietmar Jannach, Toshihiro Kamishima, Jan Krasnodebski, and Luiz Pizzato. Multistakeholder recommendation: survey and research directions. *User Modeling and User-Adapted Interaction*, 30(1):127–158, 2020.

[11] Giovanni Colavizza, Tobias Blanke, Charles Jeurgens, and Julia Noordegraaf. Archives and AI: an overview of current debates and future perspectives. *Journal on Computing and Cultural Heritage*, 15(1):1–15, February 2022.