DOI: 10.7763/IPEDR. 2012. V51. 1

CULTURA: Supporting Enhanced Exploration of Cultural Archives through Personalisation

Eoin Bailey [1] [+], Seamus Lawless [1], Cormac Hampson [1], Alexander O'Connor [1], Mark Sweetnam [2], Owen Conlan [1] and Vincent Wade [1]

[1]Knowledge and Data Engineering Group, School of Computer Science and Statistics,
Trinity College Dublin, Ireland
[2]School of History and Humanities, Trinity College Dublin, Ireland

Abstract. A key challenge facing curators and providers of digital cultural heritage across Europe and Worldwide is to instigate, increase and enhance engagement with cultural archives. To achieve this, a fundamental change in the way cultural resources are experienced and contributed to by communities is required. Furthermore, it is of central importance to create systems which can support the full spectrum of the user community - from professional researchers engaged in answering complex queries, to supporting novice students or members of the public in finding their way through the vast collection of resources. This paper presents CULTURA, a personalisation environment for navigating digitised cultural heritage archives. An initial description of a case study surrounding the 1641 Depositions Archive is followed by an analysis of the results of trials with students in the humanities.

Keywords: Personalisation, Digital Cultural Heritage, CULTURA, User Modelling.

1. Introduction

Substantial effort has been expended in the digitisation and preservation of cultural heritage collections. Traditionally, this effort has been focused on the creation of digital representations of cultural artefacts, and the creation of metadata and documentation associated with these artefacts. As a result, a continually expanding volume of content is now available to humanities scholars, in a variety of formats. After digitisation, these collections are typically monolithic and can often be difficult to search and navigate. Historic manuscript collections often contain text which is highly inconsistent in terms of language, spelling, punctuation and terminology as language and spelling were rarely standardised. Existing platforms [1-3] for digital archives tend either to provide very basic tools for dealing with a range of collections, or more complex tools whose utility is tied to a very specific type of collection, or even to a single archive [4]. By contrast, CULTURA is designed to be a corpus-agnostic platform that provides tools for the detailed exploration and interrogation of a range of digital collections. For this reason, its development to date has involved not close cooperation between humanities scholars and computer scientists. CULTURA is being developed around two very different content collections. Trinity College Dublin is providing the 1641 Depositions [5], a textual source, and the University of Padua is providing the contents of the Imaginum Patavinae Scientiae Archivum (IPSA) [6], an image-based collection.

1.1. Motivation

Understanding a large collection of documents presents a significant challenge, a global and micro view of the documents is required and views will also vary depending on the experience and knowledge of the user. Additionally the language in the documents themselves poses a challenge, for example variations in spelling of words, or the change of language over time. The digitisation process opens these documents up to an array of techniques and processes to aid in these issues, while also making the fragile documents available to a much wider audience. Additionally, different tiers of researchers and users have different tasks, goals, and scaffolding around their research. An environment that supports professional discoveries and new

[+] Corresponding author. Tel.: + 353 1 896 8431; fax: + 353 1 677 2204.
E-mail address: eoin.bailey@scss.tcd.ie.

knowledge is key for these communities, and several communities whom interact with cultural archives have been identified [7]:

- Professional researcher: *scholars, academics, tutors and historical curators*
- Student community*: post-doctoral, postgraduate and undergraduate students*
- Non-professional researcher: *member of a historical society, authors, publishers, members of the public with a sustained interest in history*
- Interested members of the general public: *these can include both adults and school children*

1.2. CULTURA

CULTURA is delivering innovative adaptive services and an interactive user environment which dynamically tailors the investigation, comprehension and enrichment of cultural archives. Through the provision of such functionality, CULTURA empowers all users to investigate, comprehend and contribute to digital cultural collections. CULTURA is a next generation adaptive system providing multi-dimensional adaptivity along several axes, these include a personalised information retrieval and presentation system which responds to models of user and contextual intent; community-aware adaptivity which responds to wider community activity, interest, contribution and experience; content-aware adaptivity which responds to the entities and relationships automatically identified within the artefacts and across collections; and personalised dynamic storylines which are generated across individual as well as entire collections of artefacts. In order for CULTURA to deliver these systems the research is focussed on advancing and integrating key technologies including:

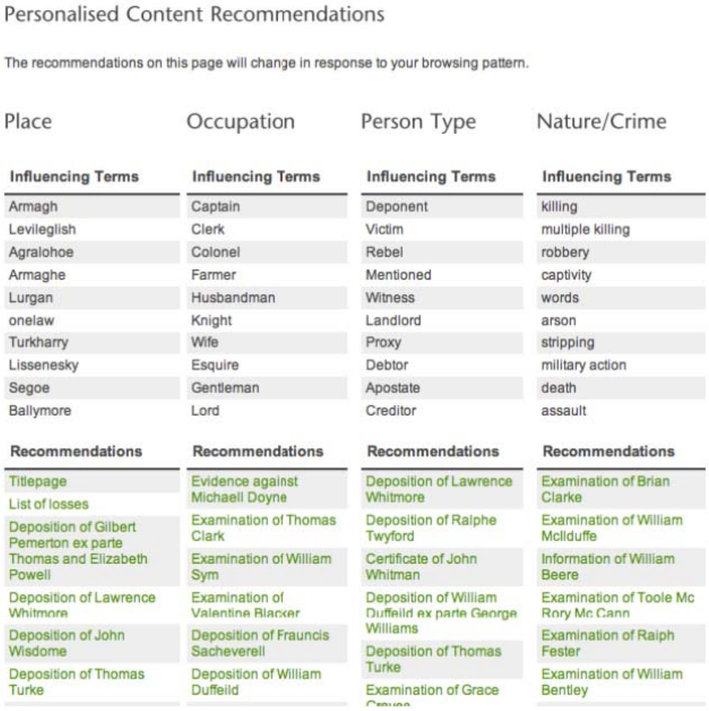
- cutting edge natural language processing, which normalises ambiguities in noisy historical texts
- entity and relationship extraction, which highlights the key individuals, events, dates and other entities and relationships within unstructured text
- social network analysis of the entities and relationships within the content, and also of the individuals and broader community of users engaging with the content
- multi-model adaptivity to support dynamic reconciliation of multiple dimensions of personalisation

2. The 1641 Depositions

The 1641 Depositions are a collection of noisy text documents, mainly of a legal nature, dating from the 17th Century. They primarily contain witness testimonies from Protestants, but also some Catholics, from all social backgrounds. The collection, which has been digitised and transcribed, contains over 8,500 depositions or 20,000 pages, in which men and women of all classes and from all over Ireland told of their experiences following the outbreak of rebellion by the Catholic Irish in October 1641. This body of material provides a unique source of information for the causes and events surrounding the 1641 rebellion and for the social, economic, cultural, religious, and political history of seventeenth-century Ireland, England and Scotland. This is typical of the category of digital resource which will benefit most from CULTURA as it is inconsistent in spelling, punctuation, nomenclature and word forms, and reflects a cultural outlook quite different to the modern one. From a technological perspective, the 1641 Depositions represent a textually rich digital humanities collection, which is characterised by noisy text, inconsistent sentence structure, grammar and spelling. These artefacts have active communities of interest because of their wider social and historical implications that transcend geographical and chronological boundaries and continue to shape opinions and values to this day. The depositions display important similarities to much of the user-generated content found on the world-wide web today. They are inconsistent in almost every aspect, including spelling, punctuation, case and language. These similarities allow CULTURA to draw upon state of the art approaches in Adaptive Hypermedia and Adaptive Web systems research. Such research is concerned with improving the retrieval and composition of information based on an individual’s needs and interests [8,9].

3. User Modeling in CULTURA

The CULTURA adaptive environment empowers a user's investigation, comprehension and experience with cultural archives by adapting to each individual user, adjusting the selection, the sequencing and the presentation of the components of the archive in response to user choices and activities. Users are modelled as they navigate the cultural archive, and information on what artefacts they have viewed, bookmarked, and annotated are used in the initial deployment of the CULTURA environment. In the trial these actions each influence the terms found under four headings contained in the 1641 Collection (Place, Occupation, Person Type, and Nature/Crime). Terms from each heading associated with the artefact being acted upon by the user are given an increased weight, thereby increasing their significance for that user, while all other terms are given a lowered weight. User actions also impact the level of confidence associated with an action e.g. the act of viewing an artefact has less significance than the act of bookmarking or annotating the same artefact. The result of this process is an ordered list of terms for each user of the system; these terms model a user's interests within the cultural archive.



Personalised Content Recommendations

The recommendations on this page will change in response to your browsing pattern.

Place	Occupation	Person Type	Nature/Crime
Influencing Terms	**Influencing Terms**	**Influencing Terms**	**Influencing Terms**
Armagh	Captain	Deponent	killing
Levileglish	Clerk	Victim	multiple killing
Agralohoe	Colonel	Rebel	robbery
Armaghe	Farmer	Mentioned	captivity
Lurgan	Husbandman	Witness	words
onelaw	Knight	Landlord	arson
Turkharry	Wife	Proxy	stripping
Lissenesky	Esquire	Debtor	military action
Segoe	Gentleman	Apostate	death
Ballymore	Lord	Creditor	assault
Recommendations	**Recommendations**	**Recommendations**	**Recommendations**
Titlepage List of losses	Evidence against Michaell Doyne	Deposition of Lawrence Whitmore	Examination of Brian Clarke
Deposition of Gilbert Pemerton ex parte Thomas and Elizabeth Powell	Examination of Thomas Clark	Deposition of Ralphe Twyford	Examination of William Mcllduffe
	Examination of William Sym	Certificate of John Whitman	Information of William Beere
Deposition of Lawrence Whitmore	Examination of Valentine Blacker	Deposition of William Duffeild ex parte George Williams	Examination of Toole Mc Rory Mc Cann
Deposition of John Wisdome	Deposition of Frauncis Sacheverell	Deposition of Thomas Turke	Examination of Ralph Fester
Deposition of Thomas Turke	Deposition of William Duffeild	Examination of Grace	Examination of William Bentley

Fig. 1: An example of the personalised content which is displayed to users within the CULTURA portal.

4. Evaluation and Analysis

A trial of the initial CULTURA environment was run in the School of History in Trinity College Dublin with both professional researchers and post-graduate students. The trial utilised the 1641 Depositions, limited to county Armagh. The professional researchers used CULTURA as an aid for writing a research paper, while the post-graduate students used CULTURA while completing assignments as part of their studies.

4.1. Experimental Approach

The post-graduate students were split into three groups. Each group had a different topic to research and were required to write an essay based on what they understood from the depositions. The professional researchers used the system while writing a research paper related to the 1641 Depositions. The trial is intended to provide a baseline from which future iterations of CULTURA can be compared and contrasted.

4.2. Analysis

Figure 2 outlines the different points from which a user of the CULTURA environment visited a deposition. The personalised listings account for 7.88% of the total number of depositions viewed when looking at all users. However, when only accounting for student users of CULTURA, the personalised listings account for 8.51% of all depositions viewed. This compares with only 1.26% when looking at the professional researcher group. This immediately provides a view that the personalised recommendations are more useful to users with less knowledge of the artefacts. Also of note is the use of bookmarks, the student groups used bookmarks to access depositions only 0.77% of the time, whereas the professional researchers used bookmarks to access the depositions 29.56% of the time. This clearly shows the importance of delineating groups when performing adaptations, as even these groups, both of whom have research backgrounds, approach and utilise different aspects of the CULTURA environment differently.

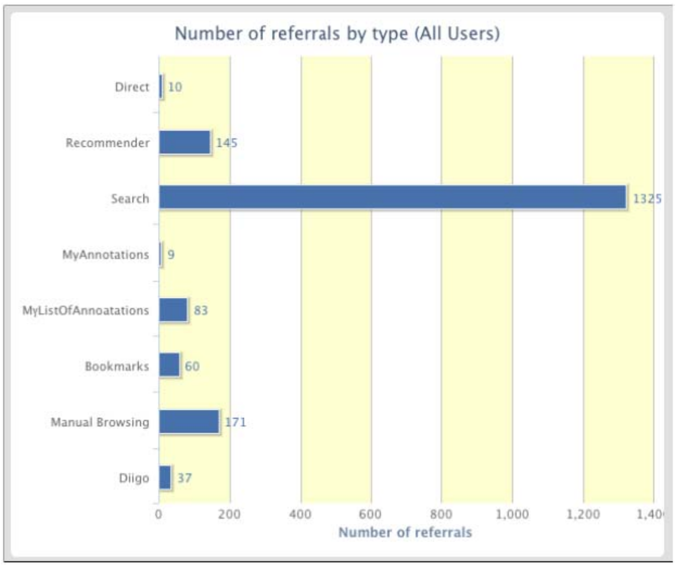


Fig. 2: Referrer location for each view of a 1641 deposition, all users.

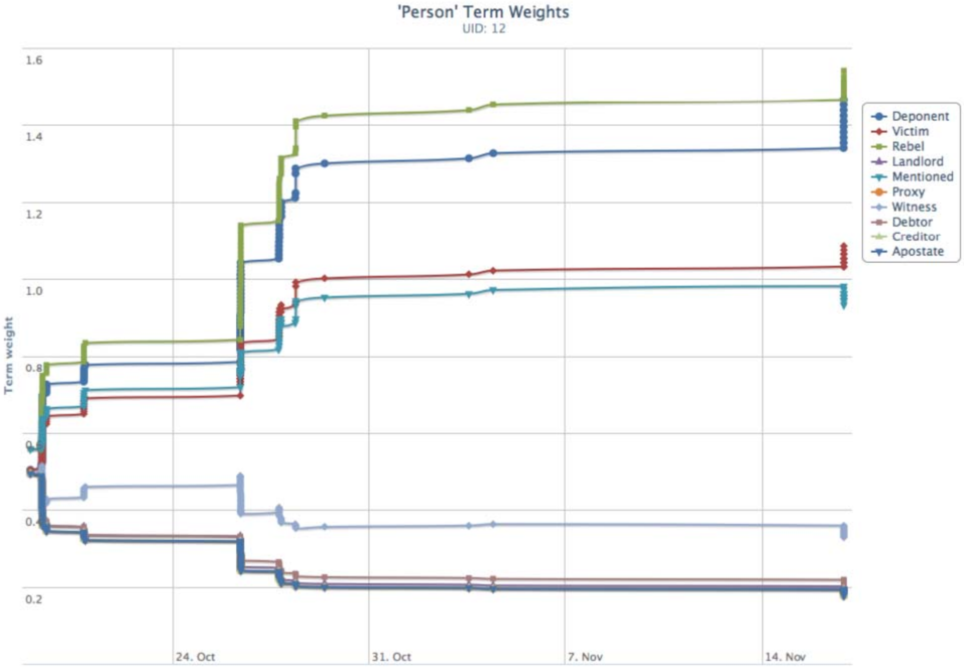


Fig. 3: Changing term weights for a single user.

Figure 3 shows how the term weights in the person category changed for a single user during the course of that user navigating the CULTURA environment. The terms rise and fall as the user's interests change. The task this user was set involved rebels, this is the highest ranked term for this user in this category, indicating that the CULTURA environment, working with the user, accurately represented the user's interests. Analysis of the changing term weights for the post-graduate researchers by professional researchers expert in the 1641 depositions indicate that the terms weights for each of the three groups are indicative of the essays assigned to each group. This indicates that the system captured the intent of this group of users.

5. Summary and Future Work

The outcome of the user trial has shown a number of interesting results; professional researchers whom are well versed in the cultural collection are less likely to utilise recommendations than researchers learning about the collections. However the same professional researchers found a greater use in bookmarking and annotating parts of the cultural archive. Focusing on the post-graduate students, the models of the users and the recommendations that came from the combination of the user model and the adaptive service were useful and of interest to the researchers performing tasks with the environment. More work needs to be completed on the definition of the distinct user groups from the user modeling perspective; this will enable the next iteration of CULTURA to provide more accurate personalisation for each user.

Integration of a second collection, Imaginum Patavinae Scientiae Archivum (IPSA) [6], is also in progress. From a technical perspective, IPSA represents a very different kind of cultural archive to the 1641 Collection. The IPSA collection is primarily image based, with substantive metadata available. This metadata not only provides descriptive passages, but is also historically valuable as it captures the scientific processes which were prevalent during the creation of the original collection. This collection differs significantly from the 1641 collection as it is largely composed of images and not text. The metadata includes information on the content and the provenance of the digital images. These contrasting, yet complimentary collections are used to validate the spectrum of techniques offered by CULTURA.

The next release of CULTURA will normalise collections (such as the 1641 depositions) in order to reduce the variation in terms and references to entities. Information extraction can then be applied to the normalised text to annotate the content with discovered entities and relationships. Social Network Analysis (SNA) and Influencer Network Analysis (INA) will then be conducted to identify the communities and influential individuals described in the content collections. SNA and INA will also be conducted on the community of researchers using the collections in order to foster collaboration and to identify influential members of the community.

6. Acknowledgements

CULTURA is funded by the European Commission's Seventh Framework Programme (FP7/2007-2013), activity ICT-6-4.1 Digital Libraries and Preservation.

7. References

[1] Tuukka Ruotsalo et al. Smartmuseum: Personalized Context-aware Access to Digital Cultural Heritage. Proceedings of the International Conferences on Digital Libraries and the Semantic Web 2009, Trento, Italy.

[2] The Cultural Heritage Information Presentation project. http://www.chipproject.org (accessed 19th May 2012).

[3] MultimediaN N9C Eculture project. http://e-culture.multimedian.nl (accessed 19th December 2012).

[4] Mark S. Sweetnam and Barbara A. Fennell Natural language processing and early modern dirty data: applying IBM Languageware to the 1641 depositions, Lit Linguist Computing, first published online December 15, 2011

[5] 1641 Depositions http://1641.tcd.ie (accessed 3rd February 2012).

[6] IPSA Imaginum Patavinae Scientiae Archivum http://www.ipsa-project.org (accessed 19th May 2012).

[7] Sweetnam, M., Agosti, M., Orio, N., Ponchia, C., Steiner, C., Hillemann, E., Ó Siochrú, M and Lawless, S. "User Needs for Enhanced Engagement with Cultural Heritage Collections". In the Proceedings of the 16th International Conference on Theory and Practice of Digital Libraries, TPDL 2012, Pafos, Cyprus. In Press. 2012.

[8] Brusilovsky, P. "Adaptive Navigation Support", In The Adaptive Web, LNCS, vol. 4321, P. Brusilovsky, A. Kobsa, W. Nejdl (eds.), Berlin Heidelberg New York: Springer-Verlag. 2007.

[9] Brusilovsky, P., Kobsa, A. and Nejdl, W. (eds.). "The Adaptive Web: Methods and Strategies of Web Personalisation". In The Adaptive Web, Lecture Notes in Computer Science, LNCS, vol. 4321, Berlin Heidelberg New York: Springer-Verlag. 2007.