Eurographics Conference on Visualization (EuroVis), Posters Track (2017)
A. Puig Puig and T. Isenberg (Guest Editors)

Diggersdiaries: Using text analysis to support exploration and reading in a large document collection

J. Nualart Vilaplana[†1] M. Pérez-Montoro[1]

[1]Department of Information Science and Media Studies. University of Barcelona. 08014-Barcelona.Catalonia. Spain

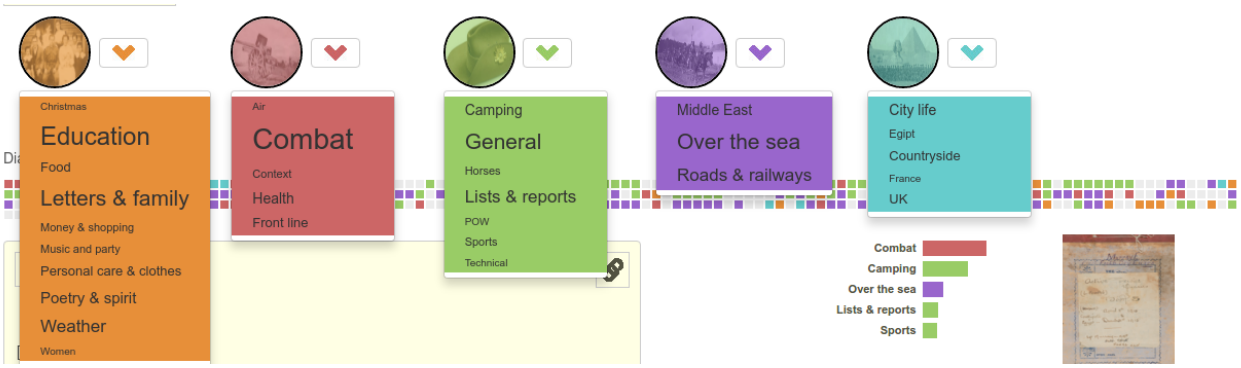

Figure 1: *Detail of Diggersdiaries interface: the two-level menu of topics support reading and exploration of a large collection of World War I ANZAC soldiers' diaries.*

Abstract

This work introduces Diggersdiaries, an web interface to a historical textual document collection. Digital collections are rich in content, but traditional search-based and faceted-based interfaces cannot represent their richness efficiently - e.g. for time-poor and casual browsing. This project addresses this challenge using data analysis (topic models) transparently integrated into reading-centric interface as a two-level browsing menu of semantic topics. The interface offers multiple exploration visualization tools. Its main contribution it is that the interface is fully reading-oriented. The tool is available at `http://diggersdiaries.org`

Categories and Subject Descriptors (according to ACM CCS): H.3.1 [Information Systems]: Content Analysis and Indexing—H.3.7Digital Libraries

1. Introduction

The digitization of cultural heritage by public institutions has led to a growing amount of digital collections available online. Most often, these collections are presented in classic text-based websites accompagned by faceted paginated lists of items with basic interaction, such as filters, sorting, and full-text search. These search-and-list interfaces cannot represent the contents of collections at large scale, at least for two reasons: sparse metadata (which does not describe content), and limited focus and scope of representation (cannot show all, only show a subset after a query).

This work proposes strategies to build rich interfaces for large collections of digitized textual documents. It is presented Diggersdiaries, a project that allows exploration of a transcribed collection of World War I diaries, letters, and reports by Australian soldiers and their families (WWI-Diaries), The collection is hosted at the State Library of New South Wales (SLNSW), Australia [SLN10]. The current SLNSW institutional interface is limited in the typical ways - based on sparse metadata, too big to browse and to read, long faceted-list and search are the only access to the collection.

† Corresponding author:jaume@nualart.cat

The proposed interface is designed in order to help solving the mentioned problems, often found in historcal digitalized collections. As its main feature, the interface allows to read texts that are too long to be fully read. The final result is a combination of interface design, text analysis and the crossread technique.

2. Related Work

Many cultural public institutions use classic interfaces to present their online collections. Data analysis combined with data visualization can be of help in this regard. Different general approaches have been developed for processing large collections of text, including text corpora interfaces, breaking text linearity, and distant reading.

Text corpora interfaces. When multiple visualization options are avalilable, some works propose to offer all of them straight on the home page [EW14], [Col16]. The proposed visualization follow the "Show everything" paradigm, introduced in 2009 by Stamen Studio [sta]. Every diary and every page is visualized in its context (e.g. pages of a diary, all pages, all diaries, etc). Each item is color-encoded according to semantic topics. The exploration paradigm follows the idea of a derive, this is a mixture of random and intuition browsing decisions, to approach the contents. This is studied, among others, by DÃűrk in "The Information Faneur" work [DSW11].

Breaking linearity. One techiniques to explore large texts is to break the linearity of the text. In the 60s the project Xanadu defined the principles of hypertext in the digital age [Ted60]. Xanadu was a visionary definition of standards for the WWW, where jumping from text to text was already well defined. Another relevant work is the concept of rhizome [DG87] where a book is a lot of books, so a narrative can be multiple. More examples can be found in literature [Cor66].

Distant reading. Moretti's work [Mor05] has become a manifesto for literary scholars, suggesting that "literature scholars should stop reading books and start counting, graphing, and mapping them instead". This philosophy is called distant reading. Other authors have analysed text analysis techniques in the digital humanities field, and specifically topic models [Ble12].

The work Topic Modeling Martha Ballards Diary [Ble10] analyses the diaries of Martha Ballard, a New England midwife who kept a daily diary for over twenty-seven years, starting on 1785. The collection consists of 1400 handwritten pages. The topics generated are manually labelled with descriptive titles, such as: midwifery, church, death, etc. Another related project is Mining the dispatch [Nel10], Nelson applied topic model analysis to a collection of news texts from the U.S. Civil War during the years 1860-65.

3. Methods and Resutls

This research project presents an interface to part of the Word War I Diaries collection from the SLNSW. At the time of collecting the data, the collection contained 688 diaries, letters, and reports, written by 337 authors, with a total of 81763 pages. The collection was acquired by the Library through a combination of donations and acquisitions. Once digitized, the documents were crowd-edited partially by SLNSW employees and volunteers. Diggersdiaries is a new tool for reading and explore this collection. Technically the project is a client side Javascript (AngulaJS) application; the data source is a set of static JSON files that encode the collection contents and analysis data.

3.1. Interface design decisions

Diggersdiaries provides a page-centric approach for exploring and, especially, reading this collection. In each part of the interface, letter and diary pages are represented as small colour-coded squares using the 5 main topic categories mentioned in subsection 3.2. Each page has a score for each of the topics; the page color indicates the group of topics which the highest score. Three data visualisation elements are the tools to explore and overview the collections: by pages, by diaries, and by date.

- **By-pages overview**: is a page-grid visualisation device coloured according to the five categories of topics. The pages can be filtered by category.
- **By-diaries overview**: is a facetted view of all diaries that can be sorted by date and alphabetically by author names and topic names.
- **By-date overview**: shows the start and end dates of each diary. Each diary is epresented as a bar that is related to a timeline axis.

The main element of the interface is the reader. The reader is integrated in all the visualizations. In the proposed method of reading the collection, the user can randomly read pages from the collection, still focusing on a given topic. The reader element introduces an inside-out approach exploration of textual document collections, and the idea that the construction of an overview of a collection of texts can be made upon from samples of the collection

3.2. Data Analysis

The texts of the collection were analyzed by topic models using the open source toolkit MALLET [McC02]. The resulting list of topics were reviewed and labeled manually according to their coherence, and semantic meaning. Second, each topic was classified as either removable, well-defined, or a synonym of a well-defined topic. The final topic set was organized at two levels.

- **Personal (10)**: Christmas, Education, Food, Letters & family, Money & shopping, Music and party, Personal care & clothes, Poetry & spirit, Weather, Women.
- **War (5)**: Air, Combat, Context, Health, Front line.
- **Military life (7)**: Camping, General, Horses, Lists & reports, POW, Sports, Technical.
- **Travelling (3)**: Middle East, Over the sea, Roads & railways.
- **The accidental tourist (5)**: City life, Egypt, Countryside, France, UK.

4. Conclusions

Diggersdiaries provides an innivative interface for exploring and, especially, reading a large historical collection of transcribed manuscripts. Since this project is a work in progress, future work will include a qualitative user evaluation study investigating the usability and usefulness of reading a collection of texts in piece through a clean and easy to use interface.

Acknowledgments

This work is part of the projects "Open access in Spain: an assessment of its impact on scientific communication system" (CS02014-52830-p) and "Interactive content and creation in multimedia information communication: audiences, design, systems and styles" (CSO2015-64955-C4-2-R) (MINECO/FEDER). Spanish Ministry of Economy and Competitiveness.

References

[Ble10] BLEVINS C.: Topic modeling martha ballard s diary. *Pers. Blog* (2010). (Visited on 12/12/2015). 2

[Ble12] BLEI D.: Topic modeling and digital humanities. *Journal of Digital Humanities* 2, 1 (2012), 8–11. 2

[Col16] COLLECTIVE: The eugenics archives. http://eugenicsarchive.ca/, 2016. (Visited on 08/04/2016). 2

[Cor66] CORTÁZAR J.: Hopscotch (rayuela). *New York: Pantheon* (1966). 2

[DG87] DELEUZE G., GUATTARI F.: Introduction: rhizome. *A thousand plateaus: Capitalism and schizophrenia* (1987), 3–25. 2

[DSW11] DOERK M. C., SHEELAGH, WILLIAMSON C.: The information flaneur: a fresh look at information seeking. In *Proceedings of the 2011 annual conference on Human factors in computing systems* (2011), pp. 1215–1224. 2

[EW14] ENNIS B., WHITELAW M.: Australian prints + printmaking. http://printsandprintmaking.gov.au/, 2014. (Visited on 02/01/2016). 2

[McC02] MCCALLUM A. K.: Mallet: A machine learning for language toolkit. http://mallet.cs.umass.edu, 2002. (Visited on 11/28/2015). 2

[Mor05] MORETTI F.: *Graphs, maps, trees: abstract models for a literary history*. Verso, 2005. 2

[Nel10] NELSON R. K.: Mining the dispatch. http://dsl.richmond.edu/dispatch/pages/home, 2010. 2

[SLN10] SLNSW: State library of nsw - transcripts. (Visited on 12/22/2015). 1

[sta] Stamen.com. http://stamen.com/. Accessed: 2013-09-26. (Archived by WebCite at http://www.webcitation.org/6Jv9xzRP8). 2

[Ted60] TED NELSON, ET AL: Project Xanadu. http://xanadu.com/, 1960. [Accessed: 2014-07-26. (Archived by WebCite at http://www.webcitation.org/6RLo0HzFo)]. 2