Personal Research Assistant for Online Exploration of Historical News

Lidia Pivovarova[1], Axel Jean-Caurant[2], Jari Avikainen[1], Khalid Alnajjar[1], Mark Granroth-Wilding[1], Leo Leppänen[1], Elaine Zosa[1], and Hannu Toivonen[1]

[1] University of Helsinki
firstname.familyname@helsinki.fi
[2] University of La Rochelle
axel.jean-caurant@univ-lr.fr

Abstract. We present a novel environment for exploratory search in large collections of historical newspapers developed as a part of the NewsEye project. In this paper we focus on the intelligent Personal Research Assistant (PRA) component in the environment and the web interface. The PRA is an interactive exploratory engine that combines results of various text analysis tools in an unsupervised fashion to conduct autonomous investigations on the data according to users' needs. The PRA is freely available online together with some datasets of European historical newspapers. The methods used by the assistant are of potential benefit to other exploratory search applications.

Keywords: Exploratory Search · Intelligent Personal Assistant

1 Introduction

We present the NewsEye Personal Research Assistant (PRA)[1], able to analyse large collections of historical news using an extensible inventory of text-processing tools. These include query-based document search, finding related documents, named entity recognition, stance detection and describing the topics in a collection. The core component – the *Investigator* – performs exploratory corpus analysis on behalf of the user to discover potentially interesting phenomena in the data. The Investigator acts within the modern exploratory search paradigm [2, 10], though it uses a broad inventory of text processing tools that can be applied to various document sets depending on the query.

Intelligent personal assistants have been employed in various applications, due to their ability to provide context-based support to users efficiently, saving time and allowing them to focus on important tasks: e.g. navigation [5], time management [4], e-mail organization [7] or patient healthcare [6].

It has been noted that scholars have special information needs and require support for corpus management [8]. Historians are typically interested in analyzing historical data on a level of abstraction that computational models cannot

[1] This work has been supported by the European Union's Horizon 2020 research and innovation programme under grant 770299 (NewsEye).

fully learn on their own. Applying potentially informative computational analyses on multiple sub-collections is not only tedious and time-consuming, but sometimes ruled out by the lack of easy-to-use tools and specialist skills (e.g. programming). As a result, a tool is required that is capable of automatically analyzing historical data while giving historians the freedom to dynamically adjust the parameters and context of the analysis.

The Personal Research Assistant is implemented as a part of the NewsEye Project, which aims to develop novel methods facilitating access to digitized historical newspapers for a broad range of users, including professional historians as well as the general public. Computer scientists, historians and librarians are involved in the project, which allows developing and testing computational solutions that meet the needs of digital humanities research studying historical newspapers[2].

A platform has been built for the NewsEye project that incorporates a broad range of features such as text recognition [3], semantic annotation [9], advanced textual analytics [11] and an intelligent personal assistant. It includes a web interface that permits users to find relevant documents based on queries[3].

Users interact with the PRA through a web-interface, where the PRA returns requested information and analysis, as well as the results of the Investigator's autonomous search, along with automatically generated natural language reports, when applicable. Though the NewsEye Investigator is developed specifically for historical research, we believe the same design principles are applicable in other humanities disciplines, where objectivity is a crucial issue.

Though it is still under development, the PRA already performs independent analysis and produces meaningful results.

2 NewsEye Data Analysis Platform

The NewsEye platform provides access to a number of Austrian, French and Finnish newspapers from 19th and early 20th centuries and provides a number of analytical tools to facilitate historical research. These come in various levels of complexity, from straightforward word counts to more sophisticated probabilistic models. The data set and the tool inventory are easily extensible.

The general information flow within the infrastructure is presented in Figure 1. Images of scanned newspapers are provided by National Libraries of Austria, France and Finland. The images are processed to extract text and separate pages into articles. Articles are then semantically annotated by a number of NLP methods including named entity recognition, sentiment analysis, and novelty and event detection. All these operations are performed offline and the results are stored and made accessible through a Solr index. Dynamic text analysis is run on demand and performs query-specific analysis of sets of documents, document linking and comparative analysis of multiple document sets.

[2] For additional information on the project, its datasets, tools and publications visit https://www.newseye.eu/.

[3] Free accessible through https://platform.newseye.eu/

Thus the PRA deals with a heterogeneous data and a variety of analytical tools. The goal of the PRA is to make effective use of theese tools to find peculiarities of potential interest for historical research. The PRA produces a set of natural language reports detailing its findings. These are produced by an automatic natural language generation system [1] and can be generated in English, French, Finnish or German.

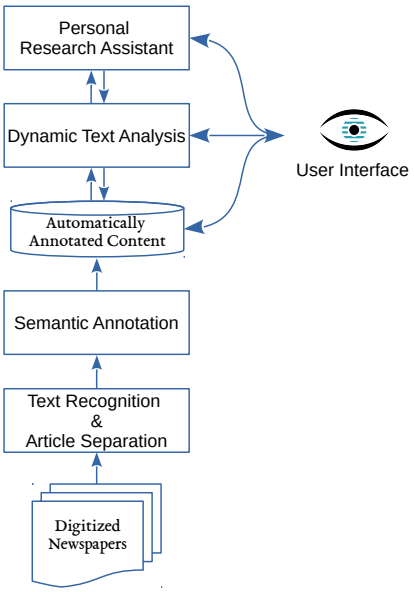


Fig. 1: Information flow within the NewsEye infrastructure.

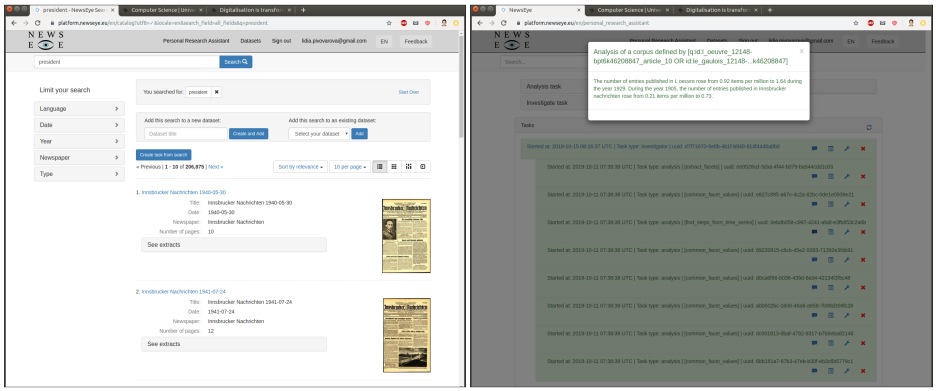
The user interface allows users to query data on various levels. First, it is possible to directly query the database index for simple data collection. The search outputs can be saved and combined to build users' own sub-corpora. Then the Investigator starts autonomous exploratory analysis based on a sub-corpus. The requirement of autonomy comes from the needs of humanities studies, where the option to approach history without predefined questions is seen as a key advantage of modern data-driven methods. In adition, the user can directly call a specific analysis tool on the sub-corpus.

This functionality is exemplified in Figure 2. Figure 2(a) presents the search interface that allows the user to browse the collection and create sub-corpora. Figure 2(b) shows analysis output, organized as a set of analysis tasks. Using icons to the right of each task the user can request a natural-language report, raw results or task parameters. A report for one task is shown.

3 Current Status and Further Work

Main parts of the data processing pipeline are implemented, at least at a prototype level. Future work will include development and integration of more sophisticated methods for text analysis. We also plan to make more newspapers available through the NewsEye platform. Thus, the PRA data and tool inventory will be expanded. This expansion does not theoretically require any changes in the interface, since most of the user forms in the interface are produced automatically based on the tool specification provided by the PRA API.

Nevertheless, some data analysis instruments can be more efficiently exploited via a more specific interface. For example, the NewsEye interface has a special section to represent *topic models*, where the user may request word clouds for each topic (Figure 3). This and other analysis tools, e.g. time series analysis, require specialized visual support.



(a) Search interface (b) Analysis interface

Fig. 2: Example screenshots from the NewsEye user interface.

The core PRA component, the autonomous Investigator, is due to change. The current investigator uses patterns – predefined sequences of tools that are run in parallel. In the future, it should be able to adjust its exploration plan on the fly. In principle, the output of its work could be presented in the simple list of tasks, as in Figure 2(b), but we plan to develop more friendly interface for the investigator.

In this paper we presented the NewsEye exploratory platform, which facilitates historical newspapers studies. The platform provides access to a number of search and text analysis tools. The current interface allows users to access to a large collection of newspapers from the 19th-20th centuries and to analyse them using the autonomous Investigator, which processes data using a variety of analysis tools. The data collection and the tool inventory will be expanded in the near future.

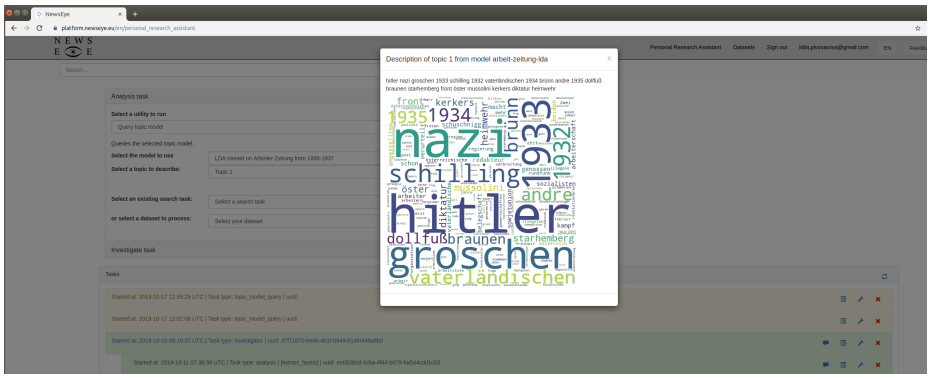


Fig. 3: Topic modelilng representation in the NewsEye user inteface.

References

1. Leppänen, L., Munezero, M., Granroth-Wilding, M., Toivonen, H.: Data-driven news generation for automated journalism. In: Proceedings of the 10th International Conference on Natural Language Generation. pp. 188–197 (2017)
2. Marchionini, G., White, R.: Find what you need, understand what you find. International Journal of Human-Computer Interaction **23**(3), 205–237 (2007)
3. Michael, J., Labahn, R., Gruning, T., Zollner, J.: Evaluating sequence-to-sequence models for handwritten text recognition. In: International Conference on Document Analysis and Recognition (ICDAR) (2019)
4. Myers, K., Berry, P., Blythe, J., Conley, K., Gervasio, M., McGuinness, D.L., Morley, D., Pfeffer, A., Pollack, M., Tambe, M.: An intelligent personal assistant for task and time management. AI Magazine **28**(2), 47 (Jun 2007)
5. Page, L.C., Gehlbach, H.: How an artificially intelligent virtual assistant helps students navigate the road to college. AERA Open **3**(4), 2332858417749220 (2017)
6. Santos, J., Rodrigues, J.J., Silva, B.M., Casal, J., Saleem, K., Denisov, V.: An iot-based mobile gateway for intelligent personal assistants on mobile health environments. Journal of Network and Computer Applications **71**, 194 – 204 (2016)
7. Segal, R.B., Kephart, J.O.: Swiftfile: An intelligent assistant for organizing e-mail. In: In AAAI 2000 Spring Symposium on Adaptive User Interfaces. Stanford, CA (2000)
8. Singh, J., Nejdl, W., Anand, A.: Expedition: a time-aware exploratory search system designed for scholars. In: Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval. pp. 1105–1108. ACM (2016)
9. Sumikawa, Y., Jatowt, A., Doucet, A., Moreux, J.P.: Large scale analysis of semantic and temporal aspects in cultural heritage collection's search. In: 2019 ACM/IEEE Joint Conference on Digital Libraries (JCDL). pp. 77–86. IEEE (2019)
10. White, R.W., Roth, R.A.: Exploratory search: Beyond the query-response paradigm. Synthesis lectures on information concepts, retrieval, and services **1**(1), 1–98 (2009)
11. Zosa, E., Granroth-Wilding, M.: Multilingual dynamic topic model. In: Recent Advances in Natural Language Processing (RANLP) (2019)