Clipping the Page – Automatic Article Detection and Marking Software in Production of Newspaper Clippings of a Digitized Historical Journalistic Collection

Kimmo Kettunen[0000-0003-2747-1382], Tuula Pääkkönen [0000-0003-3958-9732] and Erno Liukkonen

University of Helsinki, The National Library of Finland, DH Research
Saimaankatu 6, Mikkeli, FI-50100 Finland
Firstname.lastname@helsinki.fi

Abstract. This paper describes utilization of article detection and extraction on the Finnish Digi[1] newspaper material of the National Library of Finland (NLF) using data of one newspaper, *Uusi Suometar* 1869–1918. We use PIVAJ software [1] for detection and marking of articles in our collection. Out of the separated articles we can produce automatic clippings for the user. The user can collect clippings for own use both as images and as OCRed text. Together these functionalities improve usability of the digitized journalistic collection by providing a structured access to the contents of a page.

Keywords: article extraction, digitized historical newspaper collections, PIVAJ software, content marking and distribution

1 Introduction

It is a common practice that historical newspaper collections are digitized on page level: pages of the physical newspapers are scanned and OCRed and the page images serve as the basic browsing and searching unit of the collection. Searches to the collection are made on page level and results are shown on page level to the user. Page, however, is not any kind of basic informational unit of a newspaper, only a typographical or printing unit. Pages consist of articles or news items (and advertisements or notices of different kind, too), although length and form of them can be quite variable. Thus, separation of the article structure of digitized newspaper pages is an important step to improve usability of digital newspaper collections. As the amount of digitized historical journalistic information grows, also good search, browsing and exploration tools for harvesting the information are needed, as these affect usability of the collection. Contents of the collections are one of the key elements of usefulness of the collections, but also presentation of the contents for the user is important [2, 3]. According to Dengel and Shafait [4] "availability of logical structure facilitates navigation and advanced search inside the document as well as enables better presentation of the document in a possibly restructured format." Possibility to use article structure will also improve further analysis stages of the content, such as topic modeling

[1] https://digi.kansalliskirjasto.fi/etusivu?set_language=en

or any other kind of content analysis. Several digitized historical newspaper collections have implemented article extraction on their pages. Good examples are for example Italian La Stampa[2], British Newspaper Archive[3], and Australian Trove[4].

The historical digital newspaper archive environment of the NLF is based on commercial docWorks[5] software. The software is capable of article detection and extraction, but our material does not seem to behave well in the system in this respect. We have not been able to produce good article segmentation with docWorks, although such work has been accomplished e.g. in the Europeana Newspaper framework [5]. However, we have recently produced article separation and marking on pages of one newspaper, *Uusi Suometar,* by using article extraction software named PIVAJ developed in the LITIS laboratory of University of Rouen Normandy [1, 6, 7]. In this paper we describe intended use of the extracted articles in our digital library presentation system, digi.kansalliskirjasto.fi (Digi), as newspaper clippings which can be collected by the user out of the markings of the article extraction software.

2 Article Extraction

We have described results of article extraction using PIVAJ software in a recent conference paper [7]. The results we achieved with our training and evaluation collection were at least decent, if not remarkable, and we believe that they provide a useful way to introduce articles for users, too. Figure 1 describes evaluation results of a 14 issue and 56 page evaluation collection with three evaluation scenarios described in [8]. On average the three evaluation scenarios get success rates of 67.9, 76.1, and 92.2 for the whole data set of 56 pages. Same evaluation schema is used in the bi-annual digitized journalistic data evaluation campaign ICDAR [9].

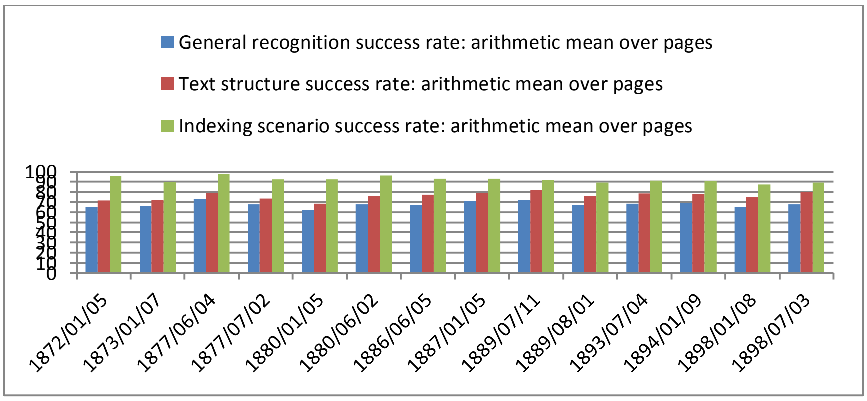

Figure 1. Area weighted success rates for the three evaluation scenarios. Mean average figures for the issues.

[2] http://www.archiviolastampa.it/

[3] https://www.britishnewspaperarchive.co.uk/

[4] https://trove.nla.gov.au/

[5] https://extranet.content-conversion.com/dW/_layouts/15/start.aspx#/SitePages/Home.aspx

3 Providing Articles for the User

Users of our digital presentation system Digi have been able to mark and collect so called clippings for several years [10]. The clipping feature has been quite popular and many users have collected hundreds and even thousands of clippings for their own collections on their user accounts[6]. The clippings can also be seen by other users. Researchers have used the clipping function to collect their data, too. So far the feature has been totally manual: the user has marked on the pdf representation of the page the textual area he/she is interested in and the image of the clipping has been stored with bibliographical information. The user can also add keywords, topic and title to the clippings. There has not been possibility of storing the OCRed text of the clipping so far, only an image file [10].

The procedure of creating articles automatically for the user utilizes the existing clipping functionality of Digi. PIVAJ uses the defined newspaper models of Uusi Suometar for article extraction, and it provides as its output an XML file (regions.xml) which contains the coordinates of the article regions for each recognized article on a page. In Digi's context these are the different parts of the clipping that are created in the order of the creation. After the regions have been entered to the presentation system, they are shown as individual clippings on the page.

Figure 2 illustrates the overall work flow of clipping production.

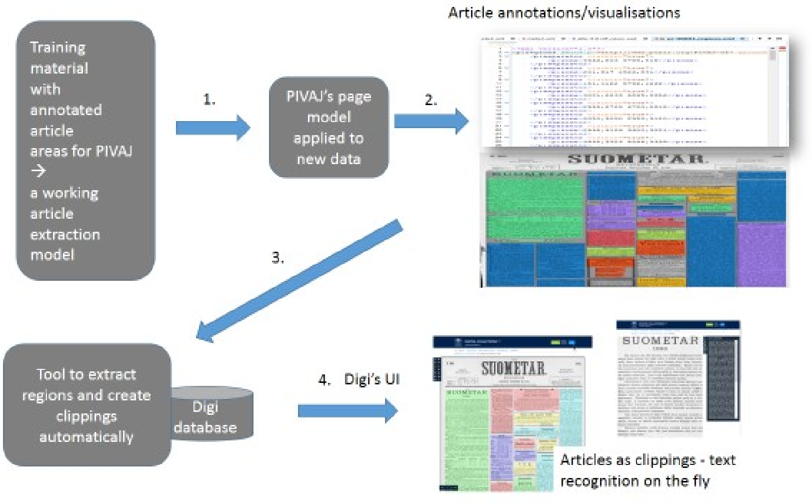
Article annotations/visualisations
Training material with annotated article areas for PIVAJ → a working article extraction model
1.
PIVAJ's page model applied to new data
2.
SUOMETAR
3.
Tool to extract regions and create clippings automatically
Digi database
4. Digi's UI
SUOMETAR
SUOMETAR
Articles as clippings - text recognition on the fly

Figure 2. Flow of article and clipping creation

After choosing the article from the automatic pre-selection of PIVAJ, the user can store the article in his/her collection. The user is also able to store the OCRed text along the clipping.

The new functionality will appear in our presentation system during the year 2019 with the data of Uusi Suometar 1869-1918. This newspaper is one of the most used in our collection and consists of 86 068 pages.

[6] In early June 2019 the number of stored clippings in our system was 165 494.

4 Conclusion

This paper has described utilization of automatic article extraction on one historical Finnish newspaper, Uusi Suometar, in the journalistic collection of The National Library of Finland. The new enhanced functionality of the digital presentation system has been implemented by using an article detection and extraction tool PIVAJ and a clipping functionality already available in the user interface of our presentation system. The user can collect automatically marked articles for his/her own use both as images and OCRed text.

Acknowledgment

This work is funded by the European Regional Development Fund and the program Leverage from the EU 2014-2020. We thank LITIS laboratory of The University of Normandy Rouen for the possibility of using PIVAJ software.

References

1. D. Hebert, T. Palfray, T. Nicolas, P. Tranouez, T. Paquet (2014). PIVAJ: displaying and augmenting digitized newspapers on the Web Experimental feedback from the "Journal de Rouen" Collection. In Proceeding DATeCH '14 Proceedings of the First International Conference on Digital Access to Textual Cultural Heritage, 173–178. http://dl.acm.org/citation.cfm?id=2595217
2. N. Fuhr, G. Tsakonass, T. Aalberg, M. Agosti, P. Hansen, S. Kapidakis, C.-P. Klas, L. Kovács, M. Landoni, A. Micsik, C. Papatheodorou, C. Peters, I. Sølvberg (2007). Evaluation of digital libraries. International Journal on Digital Libraries 8(1), 21–38.
3. H.I. Xie (2008). Users' evaluation of digital libraries (DLs): Their uses, their criteria, and their assessment. Information Processing and Management, 44(3), 1346–1373.
4. A. Dengel, F. Shafait (2014). Analysis of the Logical Layout of Documents. In Doerman, D., Tombre, K. (eds.) Handbook of Document Image Processing and Recognition, 177–222. Springer. DOI 10.1007/978-0-85729-859-1
5. M. Willems, R. Atanassova, Rossitza (2015). Europeana Newspapers: searching digitized historical newspapers from 23 European countries. Insights 28(1).
6. D. Hebert, T. Palfray, T. Nicolas, P. Tranouez, T. Paquet (2014). Automatic article extraction in old Newspapers Digitized Collections. In Proceeding DATeCH '14 Proceedings of the First International Conference on Digital Access to Textual Cultural Heritage, 3–8. http://dl.acm.org/citation.cfm?id=2595195
7. K. Kettunen, T. Ruokolainen, E. Liukkonen, P. Tranouez, D. Antelme, T. Paquet (2019). Detecting Articles in a Digitized Finnish Historical Newspaper Collection 1771–1929: Early Results Using the PIVAJ Software. DATeCH 2019.
8. C. Clausner, S. Pletshacher, A. Antonacopoulos (2011). Scenario Driven In-Depth Performance Evaluation of Document Layout Analysis Methods. 2011 International Conference on Document Analysis and Recognition (ICDAR). DOI: 10.1109/ICDAR.2011.282
9. A. Antonacopoulos, C. Clausner, C. Papadopoulos, S. Pletschacher (2013). ICDAR2013 Competition on Historical Newspaper Layout Analysis – HNLA2013. DOI: 10.1109/ICDAR.2013.293
10. T. Pääkkönen (2015). Crowdsourcing metrics of digital collections. Liber Quarterly, https://www.liberquarterly.eu/article/10.18352/lq.10090/