# Semantic Document Selection

## Historical Research on Collections that Span Multiple Centuries

Daan Odijk[1], Ork de Rooij[1], Maria-Hendrike Peetz[1],
Toine Pieters[2], Maarten de Rijke[1], and Stephen Snelders[2]

[1] ISLA, University of Amsterdam, The Netherlands
{d.odijk, o.derooij, m.h.peetz, derijke}@uva.nl
[2] Descartes Center for the History and Philosophy of the
Sciences and the Humanities, Utrecht University, The Netherlands
{t.pieters, s.a.m.snelders}@uu.nl

**Abstract.** The availability of digitized collections of historical data, such as newspapers, increases every day. With that, so does the wish for historians to explore these collections. Methods that are traditionally used to examine a collection do not scale up to today's collection sizes. We propose a method that combines text mining with exploratory search to provide historians with a means of interactively selecting and inspecting relevant documents from very large collections. We assess our proposal with a case study on a prototype system.

## 1 Introduction

The availability of vast amounts of publicly accessible digital data motivates the development of techniques for large-scale data analysis and brings about methodological changes in disciplines that are shifting from data-poor to data-intensive. So far, the humanities have profited only marginally from large-scale digital collections that possibly span decades or even centuries. This may be a result of the distinct nature of data in the humanities, which are often historically specific, geographically diverse, and culturally ambiguous [6]. Hence, a qualitative analysis of documents is a first choice. In historical research one closely studies a manually determined sample of the material [5, 9, 10]. These traditional historical research methods can be used for studying large-scale digital libraries, but they impose substantial limitations. Adapting these research methods and combining them with computationally-based methods—for document selection and the analysis of the selected documents—may yield new research questions for historians.

In this paper we describe the intertwined development of a method to select documents and a system to access a digital archive spanning several centuries, designed to provide an historian with valuable insight. Specifically, we consider the digital newspaper archive of the Koninklijke Bibliotheek (KB).[3] Today, the archive comprises over three million pages from hundreds of titles from 1618–1995, consisting of more than 36 million articles, about a third of its planned final size of 100 million articles.

The document selection method we propose is based on a combination of text mining and exploratory search methods. Text mining is an umbrella term for a wide range of techniques for obtaining useful information from large textual data sources [2, 8]. Exploratory search is a form of information retrieval where users do not precisely know what they want beforehand, or where they can find it, but prefer to explore a collection

[3] The Dutch Royal Library, National library of the Netherlands, http://kranten.kb.nl/.

to uncover new associations between documents [7]. Here, users explore the collection, and iteratively fine-tune their queries until they find what they are looking for. Typically, exploratory search interfaces have methods that quickly provide an overview of large parts of collections and tools to quickly zoom into details [3]. Comparing selections from a collection was shown to benefit television history researchers [4]. Exposing and presenting temporal information, in particular using timelines, are considered to be open challenges in exploratory search systems [1]. Our combination of interactive exploratory search and text mining supports historians to set up systematic search trails: the tooling helps them interpret and contrast the result sets returned. By exploring word associations for a result set, inspecting the temporal distribution of documents, and by comparing selections historians can make a more principled document selection.

In Section 2 we describe traditional sampling-based methods for document selection. In Section 3 we describe semantic document selection and Section 4 puts this into practice. In Section 5 we discuss implications and conclude.

## 2 Traditional Historical Research Methodology: Manual Sampling

Corpora available for historical research are often too large to be examined entirely. Researchers select subsets and closely examine only the selection. This leads to the fundamental choice: what to select? In this section we describe the most common approach to do this in historical research: through sampling. We start with three examples.

Van Vree [9] studied Dutch public opinion regarding Germany in the period 1930–1939. This was one of the first studies that explored the possibility to use media history as a form of mentality history. Van Vree [9] assumed newspapers to be the most important mass media at that time and selected four newspapers that represented major population groups (such as Catholics and Protestants). All issues of these newspapers were browsed manually, yielding a selection almost 4000 articles expressing an opinion on the subject. Neutral press, with a marketshare of about 45%, were not considered. Witte [10] followed a similar approach to study the image of the nation in the Belgian Revolution. Six newspapers from different cities and political signatures were manually selected and browsed, yielding 350 articles that expressed an opinion, possibly omitting many more. Condit [5] studied public expositions on heredity and genetics; 650 articles were selected from a period of 95 years based on indexes provided by publishers. Considering the dynamic and often inaccurate nature of indexes, one can question the representativeness of this sample. Moreover, only public expositions were studied, implicit assumptions regarding heredity were not studied.

These studies provide important insights on public opinion, but there are important practical and theoretical disadvantages to the methodological approach they employ. One can argue that not all relevant articles were selected, yet checking this would mean redoing the entire laborious selection process. Insight gained from inspecting the collection cannot be used to obtain a more representative sample. Even though the fields of digital libraries and information retrieval have come a very long way, this arguably subjective and rigid method of manual sampling is common practice—the selection of documents from a collection spanning over four centuries calls for an alternative.

## 3 Revisiting the Document Selection Process

Without dismissing the benefits of traditional document sampling methods, we advance document selection by including all relevant documents in a qualitative analysis.

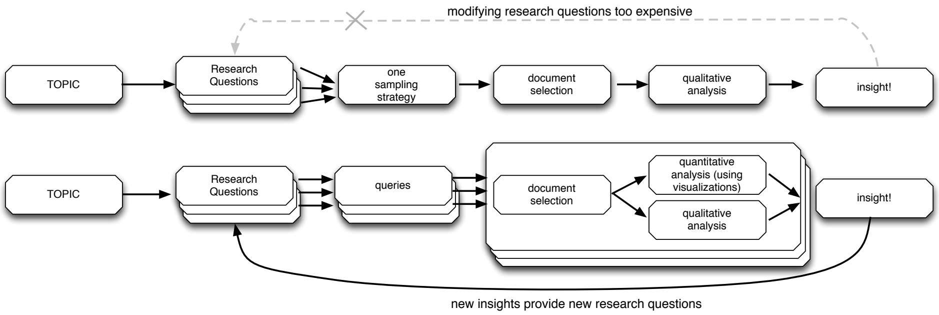


Fig. 1: Two implementations of the document selection process. We compare a manual sampling method (top) with semantic document selection method (bottom).

We have modeled manual sampling in the top part of Fig. 1. Given a research topic, a historian poses research questions that are then used to form a single sampling strategy. This strategy determines which documents to include in the selection. The researcher then develops an insight into the topic based on a qualitative analysis of the selection. This is a one way process: exploring new research questions means redoing the entire laborious process of sampling.

In semantic document selection (bottom part of Fig. 1) research questions are associated with queries against a collection. This takes away the limitation of having a single sampling strategy. A query can consist of keywords, a specific time period, a particular document source, or any combination. Each query yields a document selection, with no laborious sampling needed. Through text mining and visualizations, new insights can be gained from an initial selection. This can lead to an improved query and, therefore, a more representative document selection. This can be done by exploring word associations and metadata and through a visualization of the number of documents over time. This quantitative analysis leverages the knowledge of the historian. A clear benefit is that the historian can use the gained insight to investigate new research questions. Moreover, comparing document selections using quantitative analysis helps to validate these selections, making them less biased and more representative. With manual sampling validating the document selection is impractical or even impossible as replicating the manual selection process is too time intensive.

To support the document selection process, we developed ShoShin, an exploratory search interface that guides the user to interesting bits of information, leveraging the fact that the users are experts on the topic of interest. Due to the scale of the collection ShoShin does not store all articles locally, but collects articles on-demand and processes these on-the-fly. Fig. 2 details the architecture of ShoShin. Fig. 3 shows both an abstract overview and a screenshot of the interface. ShoShin provides the user not just with a list of relevant documents, but also with visualizations that allow inspection of, and navigation through, the document selection. Next, we describe the visualizations of word associations and temporal distribution of documents in more detail.

First, the word association visualization allows historians to glance over the content of the document selection. This visualization is a term cloud based on the relative frequencies of the words occurring in documents within a selection. Clicking on words in the cloud modifies the selection. The historian can thus inspect individual documents that contain both the clicked word and the original query.

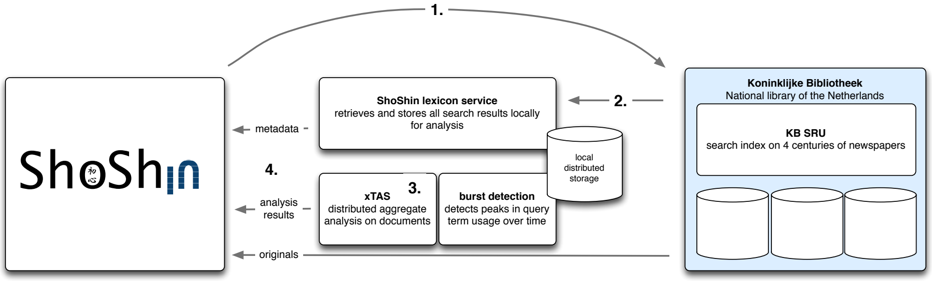
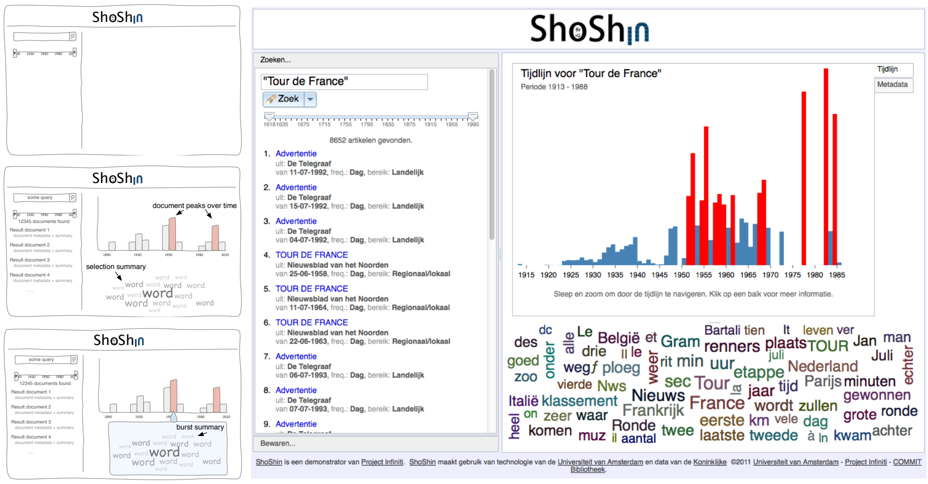

1.
2.
3.
4.
ShoShin
metadata
analysis results
originals
ShoShin lexicon service
retrieves and stores all search results locally for analysis
local distributed storage
xTAS
distributed aggregate analysis on documents
burst detection
detects peaks in query term usage over time
Koninklijke Bibliotheek
National library of the Netherlands
KB SRU
search index on 4 centuries of newspapers

Fig. 2: (1) When a user enters a query, the query is sent to the KB. (2) The resulting document set is sent to the ShoShin lexicon service. (3) Here, the documents are analyzed by xTAS. (4) The documents and analysis results are visualized in the user interface.

Second, the temporal distribution visualization allows historians to discover patterns in documents's publication dates. This visualization is a histogram of publication dates that can be explored interactively; it provides more fine-grained data when zooming in on specific parts of the histogram. To enable quick recognition of atypical patterns, bursts within the histogram—time periods where significantly more documents were published compared to neighboring periods—are highlighted. Clicking on a burst yields a visualization of word associations of that burst alone and a list of documents contained within that burst. This allows the historian to get an in-depth understanding of what each burst is about. Together, these interactions facilitate exploration of the document selection in order to detect patterns, improving the representativeness of the selection.


ShoShin
some query
12345 documents found
Result document 1
Result document 2
Result document 3
Result document 4
document peaks over time
selection summary
burst summary
word
Zoeken...
"Tour de France"
Zoek
1618 1635 1675 1715 1755 1795 1835 1875 1915 1955 1995
Bewaren...
Tijdlijn voor "Tour de France"
Periode 1913 - 1988
Tijdlijn
Metadata
1915 1920 1925 1930 1935 1940 1945 1950 1955 1960 1965 1970 1975 1980 1985
Sleep en zoom om door de tijdlijn te navigeren. Klik op een balk voor meer informatie.
des goed zoo onder alle Le België et Gram Bartali tien lt leven ver man Jan Juli echter drie il le weg ploeg weer rit min uur juli etappe renners plaats TOUR Nederland sec Tour jaar tijd Parijs minuten gewonnen vierde Nws Nieuws Italië klassement France wordt zullen Frankrijk on zeer waar Ronde eerste km vele dag grote ronde heel komen muz il aantal twee laatste tweede à in kwam achter

8652 artikelen gevonden.

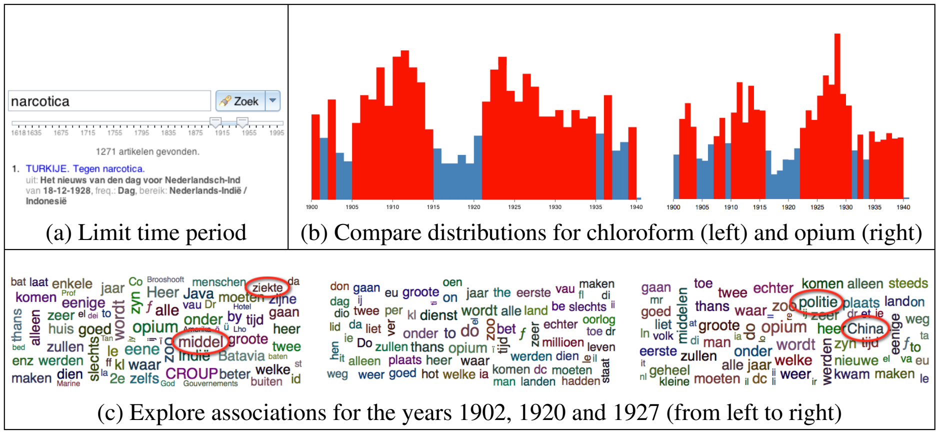
1. Advertentie
uit: De Telegraaf
van 11-07-1992, freq.: Dag, bereik: Landelijk
2. Advertentie
uit: De Telegraaf
van 15-07-1992, freq.: Dag, bereik: Landelijk
3. Advertentie
uit: De Telegraaf
van 04-07-1992, freq.: Dag, bereik: Landelijk
4. TOUR DE FRANCE
uit: Nieuwsblad van het Noorden
van 25-06-1958, freq.: Dag, bereik: Regionaal/lokaal
5. TOUR DE FRANCE
uit: Nieuwsblad van het Noorden
van 11-07-1964, freq.: Dag, bereik: Regionaal/lokaal
6. TOUR DE FRANCE
uit: Nieuwsblad van het Noorden
van 22-06-1963, freq.: Dag, bereik: Regionaal/lokaal
7. Advertentie
uit: De Telegraaf
van 06-07-1993, freq.: Dag, bereik: Landelijk
8. Advertentie
uit: De Telegraaf
van 07-07-1993, freq.: Dag, bereik: Landelijk
9. Advertentie

ShoShin is een demonstrator van Project Infiniti. ShoShin maakt gebruik van technologie van de Universiteit van Amsterdam en data van de Koninklijke Bibliotheek. ©2011 Universiteit van Amsterdam - Project Infiniti - COMMIT

Fig. 3: Interface sketches of ShoShin. A user enters a query (top left), resulting in a list of relevant documents, a term cloud and temporal distribution visualization (center left). Users can click on bursts to see word associations of that burst (bottom left). The screenshot on the right is the actual ShoShin user interface.

## 4 A Worked Example

A full-blown assessment of the ShoShin demonstrator is work-in-progress. Below, we report on one of several case studies with individual historians. Our subject, a senior his-


narcotica

1271 artikelen gevonden.

1. TURKIJE. Tegen narcotica.
uit: Het nieuws van den dag voor Nederlandsch-Ind
van 18-12-1928, freq.: Dag, bereik: Nederlands-Indië / Indonesië

Zoek
1618 1635 1675 1715 1755 1795 1835 1875 1915 1955 1995
1900 1905 1910 1915 1920 1925 1930 1935 1940
1900 1905 1910 1915 1920 1925 1930 1935 1940
bat laat enkele jaar Co Brooshooft menschen ziekte da komen Prof eenige Heer Java moeten zijne thans alleen zeer huis goed wordt zyn f alle vau Dr onder by tijd gaan heer opium middel groote twee zullen le eene zoo Indië Batavia enz werden slechts ff kl waar CROUP beter welke maken dien 2e zelfs God Gouvernements buiten id
don gaan eu groote oen on jaar the eerste vau maken dag dio twee per kl dienst wordt alle land be slechts lid da liet ver onder to do zoo bet f zeer echter oorlog toe dr Do zullen thans opium tijd millioen leven hen alleen plaats heer waar werden dien weg weer goed hot welke komen dc moeten man landen hadden staat
gaan mr toe twee echter komen alleen steeds goed middelen thans waar zoo politie plaats landon liet volk groote do opium heer China zyn tijd weg eerste zullen man onder wordt f to geheel alle jaar welke nieuwe va eu kleine moeten dc weer werden kwam maken le

(a) Limit time period

(b) Compare distributions for chloroform (left) and opium (right)

(c) Explore associations for the years 1902, 1920 and 1927 (from left to right)

Fig. 4: Case study of a historian.

torian, wants to analyze the public opinion on drugs, drug trafficking, and drug users, as represented in newspapers, in the early twentieth century (1900–1940). He wants to know whether the view on drugs is predominantly based on medical aspects (addictions, health benefits) or on social aspects (crime). How does our subject use ShoShin? First, he needs to create a lexicon of terms related to drugs. The term *narcotica* is a Dutch umbrella term for several narcotics. The actual terms describing narcotics may have changed over time and not all may be known to the historian. For a high recall of documents, a lexicon is required that captures all possible relevant terms. When a researcher uses his *domain knowledge* to create a list of words, ShoShin supports the researcher to find terms that are not readily available to him by showing a *term cloud* based on all retrieved documents (see Fig. 4c for term clouds for drug related queries). The historian can expand the original query with terms he recognizes as drugs.

To inspect the representativeness of the document selection, the historian looks at the temporal distribution of documents. He sets the time period to 1900–1940 (see Fig. 4a) and queries for several names of drugs and compares the resulting temporal distributions. With his domain knowledge he marks key events, like the Opium Treaty from Shanghai (1912), the introduction of Dutch Opium laws (1920) and the tightening thereof (1928). He then concludes that before the Dutch Opium laws came into effect the term *chloroform* was dominantly used; afterwards, the terms *opium*, *heroïne*, and *cocaïne* are more prominent (see Fig. 4b).

To gain a better understanding of the aspects associated with drugs, the historian looks at what terms were associated with the drugs over time, by examining the associated term cloud. He compares term clouds of several time periods at several scales in time. These associated term clouds (Fig. 4c) show a shift from health issues (*geneesmiddelen*, *vergiften*, *wetenschap*, *apotheken*[4]) to crime related issues (*politie*, *smokkelhandel*, *gearresteerd*[5]). By inspecting the actual word counts, the historian can find quantitative evidence for an increased use of the terms associated with *narcotica* after the Dutch Opium laws came into effect and that they are decreasingly associated with health related terms while becoming increasingly associated with crime related terms.

[4] English: medications, poison, science, pharmacies

[5] English: police, smuggling, arrest

## 5 Conclusion

We have described semantic document selections as a methodology for historical research on large repositories; it addresses three problems of traditional manual sampling: representativeness, reproducibility and rigidness. Word associations improve representativeness of the document selection as these associations are produced from the data, not from prior knowledge. Comparing selections and inspection of specific timespans in the data further support understanding the representativeness of a document selection.

The use of text mining and exploratory search allows document selections to be reproducible and remove the rigidness that stems from a single sampling strategy. Associations, longitudinal search and comparisons allow the researcher to return to document selection with new insights, at any time. In small-scale trials, we found that semantic document selection fits well in historical research methodology as an alternative to manual sampling, improving the representativeness and reproducibility of document selection and, thereby, the validity of the conclusions drawn.

**Acknowledgments.** This research was supported by the European Union's ICT Policy Support Programme as part of the Competitiveness and Innovation Framework Programme, CIP ICT-PSP under grant agreement nr 250430, the European Community's Seventh Framework Programme (FP7/2007-2013) under grant agreements nr 258191 (PROMISE) and 288024 (LiMoSINe), the Netherlands Organisation for Scientific Research (NWO) under project nrs 612.061.814, 612.061.815, 640.004.802, 380-70-011, 727.011.005, 612.001.116, the Center for Creation, Content and Technology (CCCT), the Hyperlocal Service Platform project funded by the Service Innovation & ICT program, the WAHSP and BILAND projects funded by the CLARIN-nl program, the Dutch national program COMMIT, and by the ESF Research Network Program ELIAS.

## 6 References

[1] O. Alonso, J. Strötgen, R. Baeza-Yates, and M. Gertz. Temporal information retrieval: Challenges and opportunities. In *TWAW Workshop, WWW '11*. 2011.

[2] C. Au Yeung and A. Jatowt. Studying how the past is remembered: towards computational history through large scale text mining. In *CIKM '11*, 2011.

[3] M. Bron, J. van Gorp, F. Nack, and M. de Rijke. Exploratory search in an audio-visual archive. In *EuroHCIR '11*, 2011.

[4] M. Bron, J. van Gorp, F. Nack, M. de Rijke, A. Vishneuski, and S. de Leeuw. A subjunctive exploratory search interface to support media studies researchers. In *SIGIR '12*, 2012.

[5] C. Condit. *The meanings of the gene: Public debates about human heredity*. University of Wisconsin Press, 1999.

[6] P. Courant, S. Fraser, M. Goodchild, et al. Our cultural commonwealth: The report of the american council of learned societies commission on cyberinfrastructure for humanities and social sciences, 2006.

[7] G. Marchionini. Exploratory search: from finding to understanding. *Commun. ACM*, 49(4):41–46, 2006.

[8] J. Michel, Y. Shen, A. Aiden, et al. Quantitative analysis of culture using millions of digitized books. *Science*, 331(6014):176, 2011.

[9] F. van Vree. *De Nederlandse pers en Duitsland 1930–1939*. Historische Uitgeverij, 1989.

[10] E. Witte. *De constructie van België: 1828–1847*. Lannoo Uitgeverij, 2006.