Online Religious Studies: A Pilot

Jonneke Bekkenkamp
Religious Studies
University of Amsterdam
j.j.bekkenkamp@uva.nl

Edgar Meij
ISLA
University of Amsterdam
edgar.meij@uva.nl

Maarten de Rijke
ISLA
University of Amsterdam
derijke@uva.nl

ABSTRACT

Data transitions have revolutionized many scientific disciplines, starting with the exact sciences, then the life sciences, and now the social sciences and humanities are in the process of making the transition to becoming data intensive sciences, with descriptions through quantitative measurements. New analysis tools [1], and publicly accessible utterances, opinions, transactions and interactions resulting from widespread internet and social media usage facilitate new, data-intensive research methods in disciplines that have so far relied on small-scale literature and/or panel-based studies [3]. To illustrate the new possibilities, we report on a pilot carried out by a cross-disciplinary team consisting of computer scientists and researchers in religious studies. In the latter area, research is often focused on mapping out the convictions, hopes, and beliefs of groups of people, be it within certain religions or within any other group, such as those defined by a political party.

In the pilot, religious scholars examined the core keywords in a left-wing political party in order to determine their hopes and beliefs. Rather than following their standard way-of-working, they were equipped with a search engine with an index of content crawled from discussion forums, the party's web site plus a range of online publications relating to the party and going back to 1990. In this paper we focus on lessons learned and on methodological innovations for religious scholars as well as for computer scientists building the enabling technology.

Categories and Subject Descriptors

H.4 [**Information Systems Applications**]: H4.m Miscellaneous

General Terms

Theory, Experiments

1. INTRODUCTION

In politics, "who is who?" is an important question. Religious scholars, on the other hand, are more interested in the question: "who is we?" This "we" of politics encompasses not only the "we" of the political parties, but also the "we" of the societies they envision, the *polis* of politics. The polis of politics is like the god of theology: all of politics is centered around it, yet everyone has their own different interpretation of it. Which particular representation of society (polis) appeals to the public differs from party to party.

In the field of religious studies, the aim is to map out hopes and beliefs [2]. The old approach—dividing and subdividing religions—simply does not work anymore. There is a gap between the professional articulations of belief and the actual beliefs of the people themselves. Asking people directly what they believe does not help, since hardly anyone seems to be able to articulate his or her own hopes and beliefs. Roughly speaking, there are two ways out of the impasse. One can either concentrate on observing the actions that people take—from small gestures to big manifestations—or one can concentrate on what people say when they are not explicitly articulating their beliefs. From such patterns of speaking (including typical terms, characteristic words, recurring themes, etc.) one can infer people's belief-systems. Using the latter approach one can, for example, infer vocabularies of violence, memory, creativity, death, or belonging.

Almost all people are directly inspired by modern texts. From their readings of the canonical religious texts one can retrieve the books they read for themselves, or from the films they watch the television programs they select. Hence, in the project of mapping out actual hopes and beliefs it is only natural to focus on these "holy texts" of modern culture. All of this can be done by a process known as "close reading," which is the careful, sustained interpretation of a brief passage of text. This methodology places an emphasis on the particular over the general, paying close attention to individual words and the order in which ideas unfold as they are read. However, due to the enormous volumes of text available through new media (such as websites, blogs, Twitter, etc.), close reading simply isn't humanly feasible anymore. As such, we require the aid of intelligent information access systems to help us make sense of this data deluge.

2. PILOT

A cross-disciplinary team consisting of computer scientists and researchers in religious studies performed a pilot research study together. In the pilot, religious scholars examined the notions of "we" in the context of a Dutch left-wing political party called GroenLinks. The computer scientists participating in the pilot are researchers in the field

Figure 1: Screenshot of the main search interface, showing the results for the query "islam" on two sources.

of language technology and information retrieval. In these research areas, algorithms, models, and methods are developed to improve information access on a broad range of textual data.[1] Rather than following their standard research methodology, the religious scholars were provided by the computer scientists with a search engine using an index of content crawled from discussion forums, the party's web site, and a range of online publications relating to the party and going back to 1990.

GroenLinks (Dutch for "GreenLeft") is a Dutch left-wing party.[2] It is relatively young, having been formed on 1 March 1989 as a merger of four left-wing political parties: the Communist Party of the Netherlands, Pacifist Socialist Party, the Political Party of Radicals, and the Evangelical People's Party. This particular party is of interest to us because communist, socialist, Christian, and ecological thinking where woven together during the fusion of the original parties out of which GroenLinks came into existence in 1990.

We harvested, crawled, and indexed a large number of online and offline publications relating to GroenLinks. The offline publications mainly consist of official party programs, including those for both national and European elections and going back to 1990. The bulk of the online resources came from the party itself, including the party's website,[3] a community website where various blogs of people affiliated with the party are aggregated,[4] and the official Twitter and blog posts of GroenLinks politicians (harvested through PentaPolitica).[5] We also included all tweets containing the name of the party leader and/or the name of the party, harvested over a period of several months.[6]

In the remainder of this section we first describe the tools that were developed and we then report on some of the main findings.

2.1 Tools

In order to facilitate the research done by the religious scholars, we have implemented a number of text indexing and analysis tools as well as a number of web-based visualization tools. First, a search engine was created that enables searching through each of the sources listed above. Using this engine, it is possible to search through the documents at varying levels of granularity, including sentences, paragraphs, and whole documents. Figure 1 shows an example screenshot of this tool. Being able to select particular sources, the accompanying web-interface facilitates easy comparisons between them. Furthermore, using an in-house developed text analytics toolkit called Fietstas,[7] custom text visualizations (so-called term clouds) were generated on the fly based on the displayed results. In these clouds, the size of a term is proportional to the frequency of that term in

[1] See http://ilps.science.uva.nl/.
[2] See http://en.wikipedia.org/wiki/GreenLeft.
[3] See http://www.groenlinks.nl/.
[4] See http://planeetgroenlinks.nl/.
[5] See http://pentapolitica.nl/.
[6] See http://zookma.science.uva.nl/VK2010-monitor/.
[7] See http://fietstas.science.uva.nl/.

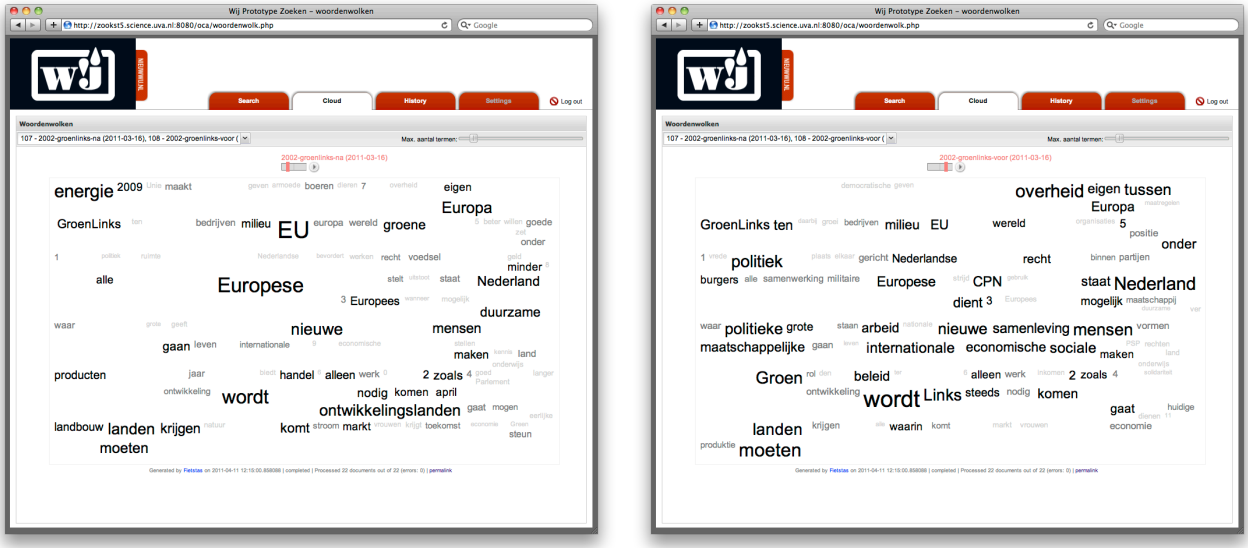

Figure 2: Screenshots of two *epochs* in a dynamic term cloud (in this case of all official GroenLinks documents published before and after 2002).

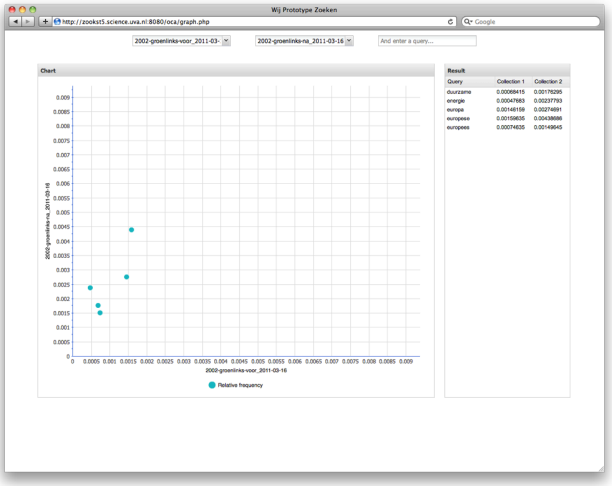

Figure 3: Screenshot showing the relative term frequency for five terms in two sources.

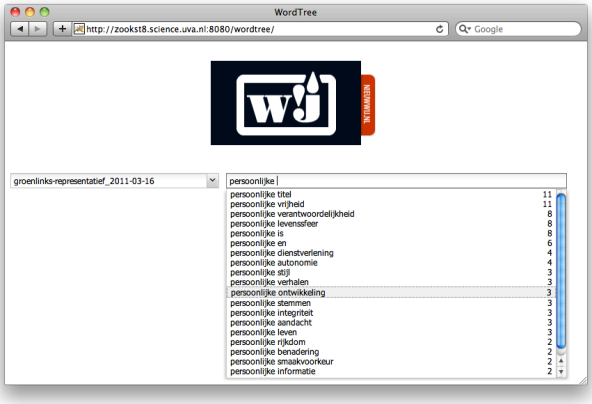

Figure 4: Screenshot of the n-gram tool.

a piece of text, such as a document or—in this case—a set of retrieval results. This type of visualization thus makes it possible to aggregate and visualize the most important terms in a set of retrieval results and observe the main theme in the glance of an eye. The terms in the generated clouds are also clickable; when clicked, the term is added to the query in order to "zoom in" on this particular aspect.

Using the same text analytics toolkit, we also provided term clouds of all documents in individual sources. This kind of analysis facilitates summarizing an entire document collection and highlighting the most important terms. Fietstas also provides so-called dynamic term clouds. In these, two (or more) sets of documents are analyzed and a term cloud is generated for each. Then, the terms in each cloud are aligned and the term clouds are presented after each other with a fluent animation. This kind of analysis makes it possible to observe which terms are frequent in one set but absent in the other (and vice versa). In contrast, when terms have a roughly equal relative frequency of occurrence, the size of these terms do not change much. Figure 2 shows an example of such dynamic term clouds using two sources.

In another tool, one can enter two sources and a term of interest (see Figure 3 for an example). Using a simple scatterplot, the tool displays the relative term frequency in all of the documents in each source collection (one on each axis). Using this method, the two sources can be easily compared on the relative frequency of terms provided by a user. When the term occurs the same amount of times in both sources, the scatterplot shows data points that are positioned roughly on the diagonal. When data points appear closer to one of the axes, this means that the particular term is less prevalent in that particular source.

The final tool is highly similar to the query suggestion facili-

ties provided by most major web search engines. One selects a source and starts typing a phrase. The tool then displays all phrases starting with that particular term, including the count of each, see Figure 4 for an example.

2.2 Findings

In this section we present some of the main findings of the religious scholars, obtained using the data and tools described above. We started by using the words presented as core keywords in the first election program for the parlementary elections of 1989 (so just before the fusion of the parties that now make up GroenLinks was formalized). We then contrasted these with the actual keywords in the sources, focusing on "we" words and word-shifts. Finally, from this combination we can infer GroenLinks' hopes and beliefs. In the election program of 1989, GroenLinks described the quality of the society they envisioned in six keywords:

- Persoonlijke zelfstandigheid (personal autonomy)
- Democratisering (democratization)
- Geweldloosheid (non-violence)
- Creativiteit (creativity)
- Verdraagzaamheid (tolerance)
- Emancipatie (emancipation)

Interestingly, each of the keywords hardly appear in any of the GroenLinks sources—with the exception of democratization and emancipation. Strikingly absent from this list of keywords are "Green" and "Left." Maybe these qualities were too obvious, or maybe naming them would invoke internal discussions as to what they mean and about which should come first. The two explanations do not exclude one another, however. In a certain sense they come together in a third explanation, namely the presence of a third shared value: "Freedom." Indeed, when you look at the terms associated with "Samenleving" (society) you find terms such as "Groen/Links" (green/left-wing) "Duurzaam/Sociaal" (sustainable/social), "Politiek" (politics), "Mensen" (people), "Goed" (good), and "Open" (open). When you look at the sources, there are two senses of the term "Open." The first one is quite literal, as in "accessible, not closed," whereas the other sense is "unfulfilled, indeterminate." The second sense features more prominently, mainly in the context of a fear of fundamentalism that discourages people to discuss the society as a community.

We also looked at the difference in vocabulary before and after 2002, the year Pim Fortuyn (a controversial Dutch politician[8]) was murdered. The most prominent change is the drop in usage of the terms "Samenleving" (community) and "Democratie" (democracy). More nuanced are the changes for the terms "Multi-cultureel"(multi-cultural) and "Ecologisch" (ecological); the former disappears and is replaced by "Open" (open), whereas the latter is replaced by "Duurzaam" (sustainable). Albeit for different reasons, "Europe" is also a term that gains in popularity after 2002.

[8]See http://en.wikipedia.org/wiki/Pim_Fortuyn.

3. CONCLUSION AND OUTLOOK

Given the increase in information available online, traditional methods of doing research may no longer be feasible for some areas of research. In this paper we have reported on a cross-disciplinary pilot project involving scholars from religious studies as well as computer scientists, aiming to aid the research done by the religious scholars using modern text searching and analysis tools.

In the course of the pilot, the participating researchers identified a number of novel requests, features, and techniques that would facilitate further research. First, more data is needed, especially with respect to the websites. For the current pilot we crawled the listed websites only at two different dates and we believe that a more regularly scheduled crawler would be able to provide us with the means of performing longitudinal analyses. Another possible angle for obtaining more data is through initiatives such as PoliDocs,[9] where parliamentary proceedings are collected and indexed automatically. Second is the term selection strategy in the term clouds. These are currently based on the simple term frequency of each term in a given set of texts. However, more elaborate variants are possible and will be examined in the future. Finally, various other measures, techniques, and methods that are common in the language technology field may be applied here. For examples, measures of complexity, style, diversity, or the "richness" of the language could be used to analyze and compare different sources. Additionally, both sentiment analysis and named entity recognition are areas that could be of particular value.

Acknowledgements

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 PROMISE Network of Excellence co-funded by the 7th Framework Programme of the European Commission, grant agreement nr 258191, the DuOMAn project carried out within the STEVIN programme which is funded by the Dutch and Flemish Governments under project nr STE-09-12, the Netherlands Organisation for Scientific Research (NWO) under project nrs 612.-061.814, 612.061.815, 640.004.802, 380.70.011, the Center for Creation, Content and Technology (CCCT), the Dutch Ministry of Economic Affairs and Amsterdam Topstad under the Krant van Morgen project, the Hyperlocal Service Platform project funded by the Service Innovation & ICT program, and the WAHSP project funded by the CLARIN-nl program.

REFERENCES

[1] M. Krakovsky. Deus ex machina. *Communications of the ACM*, 54(5):22, 2011.

[2] S. K. M. and P. Harvey. Everywhere and nowhere: Recent trends in american religious history and historiography. *Journal of the American Academy of Religion*, 78: 129–162, 2010.

[3] A. Steiner. Personal readings and public texts: Book blogs and online writing about literature. *Culture unbound*, 2(28):471–494, 2010.

[9]See http://www.polidocs.nl/.