PoliMedia - Connecting Politics and Media

Laura Hollink[1], Henri Beunders[2], Jaap Blom[3], Damir Juric[4], Max Kemman[2], Martijn Kleppe[2], and Johan Oomen[3]

[1] VU University Amsterdam
[2] Erasmus University Rotterdam
[3] Brunel University, London
[4] The Netherlands Institute for Sound and Vision

Abstract. Scholars of media and communication sciences study the role of media in our society. They frequently search through media archives to manually select items that cover a certain event. When this is done for large time spans and across media-outlets, this task can however be challenging and laborious. Therefore, up until now the focus has been on manual and qualitative analyses of newspaper coverage. PoliMedia aims to stimulate and facilitate large-scale, cross-media analysis of the coverage of political events. We focus on the meetings of the Dutch parliament, and provide automatically generated links between the transcripts of those meetings, newspaper articles, including their original lay-out on the page, and radio bulletins. Via a web application users are able to search through the debates and find related media coverage in various media outlets, facilitating a more efficient search process and analysis of the media coverage. Furthermore, the generated links are available via a SPARQL endpoint, allowing quantitative analyses with complex, structured queries.

1 Introduction

Analyzing media coverage across several types of media-outlets is a challenging task for academic researchers. Up until now, the focus has been on newspaper articles: researches search through media archives to manually select items that cover a certain event. When this is done for large time spans and across media-outlets, this task can however be challenging and laborious. Therefore, up until now the focus of has been on manual and qualitative analyses of newspaper coverage. PoliMedia aims to stimulate and facilitate large-scale, cross-media analysis of the coverage of political events, including both printed and audiovisual media, to provide a better overview of the choices that different media outlets make.

The PoliMedia project aims to facilitate cross-media analysis by connecting media items to the political events they are about. For example, when the parliament votes about a controversial proposal, several newspapers as well as the radio news will cover this event. PoliMedia links all these articles and bulletins to the particular parliamentary debate. Thus, the political event becomes the link that connects media items across their various archives and media-outlets. PoliMedia combines three open data sources: the digitized transcriptions of the

debates of the Dutch Parliament, news articles in over 20,000 newspapers in a historical newspaper archive, and around 1.8 Million radio bulletins of the Dutch National Press Agency (ANP). We discover and publish 3,804 links between speeches that were spoken by politicians in 9,294 debates, and newspaper articles and radio bulletins that cover them.

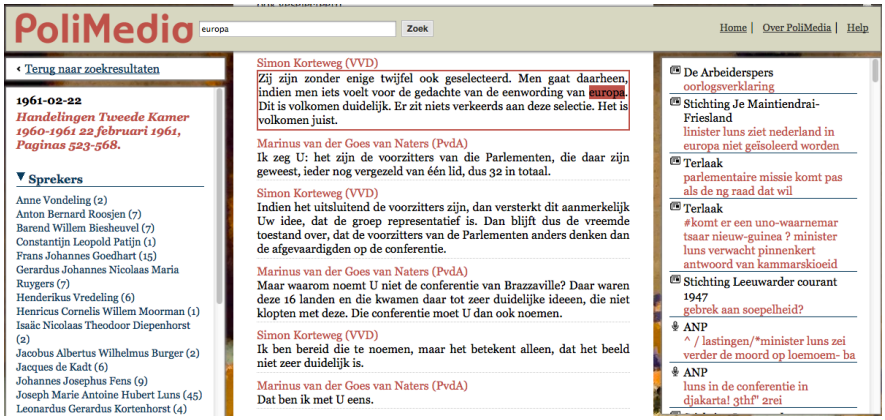
The web application was designed according to a requirements study among scholars of media and communication sciences. However, a much larger group can potentially use the application and the underlying data. To name a few: historians who study the changing relation between media and government, linguists interested in word use or framing in various media outlets over time, journalists seeking to find what other media have said about a political event. Recently, PoliMedia has won the LinkedUp Veni Competition for its use of open data for an education-related use case[5]. The PoliMedia web application allows these users to search and browse political debates by keyword, time span, name, role and party of a politician, and analyze the news items that cover it. Questions that can be answered include: *What choices do different media make in the coverage of people and topics while reporting on debates in the Dutch parliament? How did this change over time? What are the events that received most media attention?*

The PoliMedia web application is available from http://polimedia.nl. In addition, the data can be queried via a SPARQL endpoint at http://data.polimedia.nl/ .

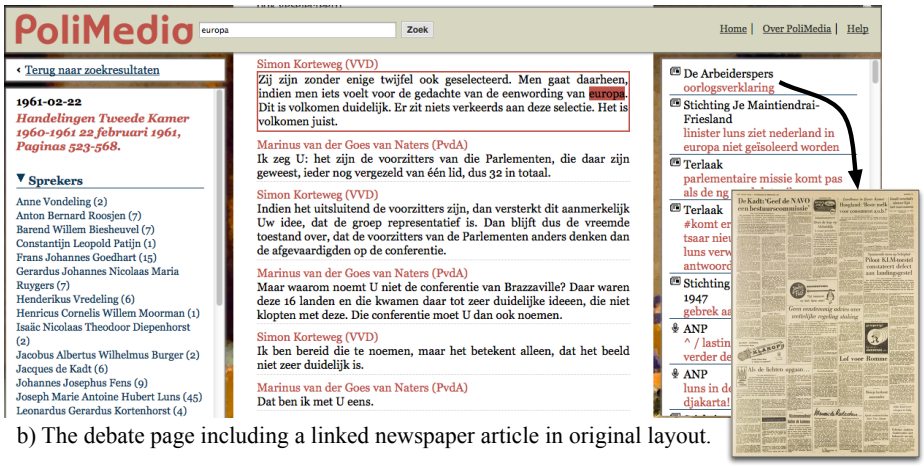
2 System description

The user interface consists of three main levels. (1) A user starts on the landing page where she enters search terms in a search field. (2) Second, she gets to the results page (Figure 1a) showing a list of the speeches that contain her query term(s). Using facets, the user can refine the results by choosing a time span, the role of the speaker (chairperson, member of the government or opposition), the name of the speaker and the political party of the speaker. For each facet, the application shows the number of results that would remain if the facet were selected. This gives the user an overview of the distribution of matching speeches over time, roles, parties or people. In Figure 1a, we have refined the results to be within the period 1960-1969. Next to each result the application shows the number of newspaper articles and radio bulletins that are linked to the debate. (3) When a result is clicked, the user comes to the debate page depicted in Figure 1b. The complete debate with its consecutive speakers is displayed. The query term(s) are highlighted. On the left of the screen, the user can refine the search further using facets. On the right, the application displays the titles and publishers of linked newspaper articles and radio bulletins. When a media item is clicked, the user is redirected to the website of the news archive, where a PDF image of the media item in its original layout is displayed. Seeing the original newspaper page is important to scholars of media and communication,

[5] http://linkedup-challenge.org/



a) The results page including search facets (left) and no. of links to media (right)



b) The debate page including a linked newspaper article in original layout.

Fig. 1. Two screenshots of the PoliMedia web application.

as font size, location on the page, headlines and images all contribute to an understanding of how the news was brought. For copyright reasons, the data in the PoliMedia application does not contain the media items themselves; only links to the URLs of the items in their original archives are included.

Since both the political debates and the media items are only available in Dutch, we user interface of the application is also in Dutch.

3 Open Data Used and Produced

PoliMedia combines three open data sources: parliamentary debates, a newspaper archive and a radio bulletin archive. The collection of Dutch parliamentary debates are published by the government in the form of complete transcripts of the speeches of politicians in parliamentary debates. For the period 1945-1995,

the transcripts of all 9,294 debates that were held are published in unstructured TXT and PDF format at http://www.statengeneraaldigitaal.nl/. The project "War in Parliament" has transformed them to a fine-grained XML structure. We build upon War in Parliament and translate their XML to RDF.

To store, query and link the debate data, we have created a semantic model in RDF which is a specialization of the more widely applicable Simple Event Model (SEM) [1], that enables us to express information associated with the debates such as topics, actors, and links to media. To increase re-usability of the data, we use FOAF and Dublin Core properties where appropriate, for example to denote names, dates, titles and publishers of debates. The model captures the structure of the debates: each debate consist of several parts (called *partOfDebate*) that are each introduced by a few words (called *debateContext*) to introduce the topic or name the report that is discussed. The design of the model is flexible - if another/future debate dataset contains several nested parts, this could just as easily be represented by the model. The RDF data also includes provenance information in the form of links to the corresponding document on the government site as well as to the XML document it is based on. Figure 2 shows the data model. For a more detailed description of the design decisions we refer to [2]. The RDF is available for download and via a SPARQL endpoint.

The newspaper archive as well as the radio bulletin archive reside at the National Library of the Netherlands. The newspaper archive stores the content of over 20,000 newspapers, both regional and national and from the former Dutch colonies, from the period 1618-1995. For copy-right reasons, the National Library does not publish archives of more recent newspapers. The radio archive contains around 1.8 Million news bulletins broadcasted between 1937 and 1984. Both archives provide full texts as well as metadata. In addition, PDF images of the original layout of the documents are provided - for radio bulletins these are in the form of copies of the typed notes that the news reader read from. PoliMedia uses the news archives as-is.

4 Creating Links between Debates and Media

We create links between speeches in debates and media items that cover them. Note that these are very different in nature: debates are long and contain spontaneously spoken words, while news articles and bulletins are short and professionally written. To find links between them, we employed a two-step method. For each speech, we first identify candidate news items by querying the two media archives for items that (a) contain the name of the speaking politician and (b) were published within seven days of the date of the debate. The fetched candidate items are tokenized, stripped of stop words, and indexed. Each media item d is represented as a term vector t of length n, where n is the length of the total number of terms in our corpus of candidate items. The elements of t are term frequency-inverse document frequency (TF-IDF) scores.

Second, we detect topics and named entities in the debates, so that we can use them to determine similarity between the speeches in the debate and the

Fig. 2. RDF model of parliamentary debates and links to news items.

candidate media items. For topic detection, we use an off-the-shelve tool called Mallet [3]. The topics and entities are used as term vectors to represent the speeches, which are compared to the media item vectors using cosine similarity. We create an explicit link between a speech and a media item if the similarity score is above a threshold. We have experimented with using named entities and topics detected in only the current speech or also in the other speeches that belong to the same *partOfDebate* (see section 6 for details on these experiments). This process resulted in 3,804 links between speeches in debates to newspaper articles or radio bulletins.

5 Web application architecture and SPARQL endpoint

The architecture of the web application is based on a Django web service, an OWLIM Lite RDF store and a Apache SOLR index. In the design of the architecture, we have strived to achieve response times that are acceptable for real use, not only for demoing purposes. The OWLIM store contains all RDF data as described in section 3 and 4: the political debates and the links to media. The same OWLIM store that is used by the web application are also behind a SPARQL endpoint. The endpoint allows users with a technical background a greater flexibility for their queries. The SOLR index was optimized using Dutch stemmers and stoppers, and was populated with documents that contain the text of a speech and 4 additional fields: the title of the debate, the description text of the *partOfDebate* the speech belongs to, the name and the party of the speaker. The contents of all fields were taken from the RDF data in the OWLIM store, which means that the SOLR index essentially contains a (less structured) subset of the RDF data. While we could in theory have queried all data directly from the OWLIM store, the use of SOLR greatly improves the speed of the full text search. As discussed in section 2, the interaction with the web application starts with a user entering one or more search terms. It is in this step that we perform a full text search on the SOLR index. Next, a user selects a speech from the result list, upon which all speeches in the relevant debate are shown. As this requires knowledge of the part-of structure of debates, the data is now retrieved from the OWLIM RDF store. For this step, we have improved the response time by splitting the RDF data in six parts and storing it in six separate stores. Each store contains the data for one decade. This of course complicates queries that require knowledge over more than one decade. However, these types of queries are not used by the web application, and the response time of the application was more important to use than the ease of use of the SPARQL endpoint. The front-end of the web application was build using HTML5, CSS, and JavaScript.

6 Evaluation of data and user interface

The development of the application was based on a requirements study with five scholars in history and political communication. The user interface was evaluated and further refined using an eye tracking study [4], in which 24 participants

performed five known item search tasks and three exploratory search tasks. Participants were divided over two conditions: one with and one without search facets. We found that the length of the debates caused problems in the no-facet condition: users had trouble gaining an overview of the topics and actors in the debate. We therefore include the facets in the final version of the web application. For this version, both the eye tracking results and post-experiment interviews showed that the application enabled users to perform both tasks satisfactory, and that users were able to analyze media coverage of a topic over time.

Secondly, we evaluated the quality of the created links [5]. We performed a manual assessment of a sample of the links. We randomly selected 20 debates from our total dataset of 9,294 debates. Second, we randomly selected 50 speeches from these 20 debates, and assessed the quality of the links to media items. In reality one speech can be linked to multiple articles, but for evaluation purposes we randomly selected one linked article per speech. We repeated this experiment three times for three versions of the linking algorithm based on (1) named entities, (2) named entities and topics or (3) named entities and topics in the larger *partOfDebate*.

The results showed the commonly observed inverse relation between precision and recall; the versions that returned most links gave the lowest precision. In our case, a high precision was more important than high recall to answer the questions of our target user group. Therefore, we used version (3) of the algorithm, which gave the highest precision (and thus lowest recall): a precision of 80% and a recall of 62%.

References

1. Willem Robert van Hage, Véronique Malaisé, R.S.L.H., Schreiber, G.: Design and use of the simple event model (SEM). Journal of Web Semantics **9**(2) (2011) 128–136
2. Juric, D., Hollink, L., Houben, G.J.: Bringing parliamentary debates to the semantic web. In: Proceedings of the DeRiVE workshop at ISWC2013, Boston, USA (2012)
3. McCallum, A.K.: Mallet: A machine learning for language toolkit. http://mallet.cs.umass.edu (2002)
4. Kemman, M., Kleppe, M., Maarseveen, J.: Eye tracking the use of a collapsible facets panel in a search interface. In: Proceedings of TPDL2013, Malta (2013)
5. Juric, D., Hollink, L., Houben, G.J.: Eye tracking the use of a collapsible facets panel in a search interface. In: Proceedings of ICWE2013, Aalborg, Denmark (2013)

Minimal Requirements (MR) & Additional Features (AF)

MR *The application is an end-user application.* PoliMedia was designed for scholars in media and communication studies, but we foresee a wider user group.

MR *The information sources are under diverse ownership or control, heterogeneous, and contain substantial quantities of real world data.* PoliMedia searches for links between heavily used media archives with articles of over 20,000 newspapers and 1.8M radio bulletins, and over 9,000 political debates.

The sources are under the control of two different bodies, namely the National Library and the Dutch government. They are different in nature (e.g. spoken words vs. written text) and have different metadata and schema's.

MR *The meaning of data plays a central role.* (1) The debate data was translated to RDF to enable more structured queries. (2) New semantic relations between media and politics are discovered and published in RDF .

AF *The application provides an attractive and functional Web interface for human users.* A requirements study and eye-tracking study ensured an attractive and intuitive interface with functionality for non-technical scholars.

AF *The application should be scalable (in terms of the amount of data used and in terms of distributed components working together). Ideally, the application should use all data that is currently published on the Semantic Web.* The PoliMedia data contains all publicly accessible newspaper data in the Netherlands, and all parliamentary debates held between 1945-1995. The only thing keeping us from including more recent material is strict copyrights on more recent news items.

AF *Rigorous evaluations have taken place.* The newly created links have been systematically evaluated. The user interface has been evaluated with an eye-tracking study.

AF *Novelty, in applying semantic technology to a domain or task that have not been considered before.* To the best of our knowledge, semantic technology has not yet been used to disclose the relation between media and politics.

AF *Functionality is different from or goes beyond pure information retrieval* Next to keyword based search in the debate data, the application offers facetted search and browsing of links to other archives. Our main contribution is newly created links between politics and media, which goes far beyond information retrieval in either archive.

AF *The application has clear commercial potential and/or large existing user base* Insight into how news covers politics and how the two influence each other is an essential part of democracy, and interesting for a much larger audience than only scholars - i.e everyone who consumes news.

AF *Contextual information is used for ratings or rankings* We use the context in which a speech is made - i.e. the larger debate - to improve the discovery of links to media.

AF *Multimedia documents are used in some way* We link to radio bulletins as well as printed media.

AF *There is a use of dynamic data (e.g. workflows)* At the moment we do not use dynamic data. We aim to do this in the future, by including media outlets that are publicly available in real time.

AF *The results should be as accurate as possible (e.g. use a ranking of results according to context)* (1) The links to media are ranked according to their similarity with the speech. (2) Search results can be tweaked to more accurately reflect the user's need in the facetted search interface.

AF *There is support for multiple languages and accessibility on a range of devices* The political and media datasets are language specific - in our case in Dutch.