

Political Mashup Position Paper

Maarten Marx Maarten de Rijke
marx@science.uva.nl mdr@science.uva.nl

ISLA, University of Amsterdam
Kruislaan 403, 1098 SJ Amsterdam

ABSTRACT

A mashup is a web application that combines data from more than one source into an integrated experience. A political mashup as we envisage it brings together political information produced by political parties (their promises and actions: election manifestos, party websites, parliamentary proceedings) and information on the reception of political promises and actions (news as well as user generated content from blogs, discussion fora and reactions to online news-stories). Such a political mashup will enable novel historical research, creates challenging opportunities for computational linguists, and offers valuable and innovative testing-grounds for information retrieval researchers.

We aim to collect, standardize in digital format, and integrate the four types of political data listed above. Integration of these data sources consists of creating connections between them along three natural dimensions:

temporal happening in the same period, being about the same event;

political actors involving the same parties and/or members of parties and/or other politically influential actors;

political issue being about the same political issue.

The elements in these dimensions link the data items to each other. The main challenge is to detect and normalize these elements in the textual data that we have. For this task we will use existing event- and news-story clustering, as well as named entity recognition and normalization techniques. The innovative and challenging aspect of our task is the type of data we use: rough and dirty user-generated content on the one hand and camouflaged content created by politicians on the other.

Categories and Subject Descriptors

H.4.3 [Information Systems]: Information Systems Applications

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.

Copyright 2008 ACM 978-1-59593-597-7/07/0007 ...\$5.00.

We describe our vision of a very specific set of highly linked data, on whose realization we have recently begun working, a political mashup that is aimed at organizing a broad range of political information in semantically meaningful ways.

Introduction.

Diachronic comparative research of election manifestos is an established research tradition which has created an extensive digital collection [3, 8]. Election manifestos have traditionally been seen as a key source of information on the standpoints of parties on political issues [9], however, the growth of the web has reduced their importance for this purpose: party websites, weblogs, discussion fora, appearance in parliament and in the press have increasingly gained importance [12]. Large-scale comparative research into political issues using a combination of these data sources is at present very difficult because there is no single entry point in which these sources are integrated. An entry point containing this data would integrate the promises of parties, their actions and how these are conceived by the press and the general public.

Our envisioned search and discovery infrastructure will thus facilitate large scale comparative research on this broad range of politically relevant data. Examples of queries the system will be able to answer are: “What are the promises made by the parties on immigration issues in 2006?”, “How did they vote in the years preceding 2006 on this issue?”, “Which party first put this issue on the political agenda; in their own material (website, manifesto), and in Parliament?”, “How did the press and public react when Minister Rita Verdonk refused to sign the General Amnesty law?”.

Linked data.

We want to create a true mashup in which users do not even experience that they are switching between sources. The different pieces of textual information are linked by semantically meaningful items like events, political issues, persons, and parties. The main challenge is to detect and normalize these items in the textual data that we have. For this task we will use existing event- and news-story clustering methods [1], as well as named entity recognition and normalization techniques [11, 5].

An obvious choice for the normalized forms are Wikipedia URLs: they satisfy the four principles of linked data from [2] and they have proven to be effective anchors for named entity normalization [4]. The innovative and challenging aspect of our task is the type of data we use: rough and dirty

user-generated content on the one hand and camouflaged content created by politicians on the other.

Use case: Historical sciences.

The integrated dataset which we intend to create could cause a major breakthrough in comparative diachronic research: the digital interlinked corpus that we plan to create can facilitate research which was virtually impossible before. The internet as a whole has gained great importance in election campaigns [12] and can be very well studied with our planned infrastructure and its integrated corpus. Here's an example research topic:

The programmatic impact of new parties on established parties. The rise of new parties is strongly linked to the theme of the contestedness of modern democracies [6]. New parties can only be successful when enough voters are dissatisfied with the functioning of a democracy and with the role that established parties are playing in it [10]. As a consequence, many new parties try to change the way democracy functions. In addition, the study of new parties can clarify how established parties cope with change and how they react to new challengers. By studying the programmatic reactions of established parties to new parties we can explain which choices are offered to voters and to what extent these new issues prevail.

Other envisaged scientific users beside historians are political scientists, corpus linguists and social scientists doing content analysis. Potential non-academic professional users are media-analysts, journalists, spin doctors and speech writers.

But there's one more group of users of political mashups that we should mention. For the 2006 Dutch parliamentary election we created a mashup (the *verkiezingskijker*, or "election watcher") for the general public in which we integrated manifestos, news and blogs [7], which was very well received and attracted over 125,000 visitors in a three week period.¹

Description of the data.

The envisaged political mashup system will provide access to, and links between, the following data available:

- 1) election manifestos, and party websites;
- 2) parliamentary proceedings
- 3) daily news from news feed, national newspapers and web-news-sites;
- 4) user-generated content: weblogs, discussion-fora posts, and online reactions to news-articles.

The free availability of this data is of course crucial for the project. We have investigated this for the European situation and obtaining the data looks feasible:

1. Through the Comparative Manifesto Project [3, 8] party manifestos of over 25 European countries for almost all elections held in those countries in at least the past 25 years are digitally available and extensively manually annotated. Party websites can be crawled [12].

¹As an example <http://verkiezingskijker.nl/2006/search.php?topic=94> presents paragraphs from the manifestos of all parties about the political theme "Luchthavens" (airports). Note that this keyword does not appear in most results. The paragraphs of the different manifestos are linked on the semantic item of airports (introducing uncertainty as can be seen in the data).

2. Proceedings of the European Parliament are available online. The Dutch parliamentary proceedings are available on microfilm starting in 1814 and are currently being digitized. As of 2005 they are available in XML format.
3. News from all EC countries is gathered by the European Media Monitor <http://press.jrc.it/NewsExplorer>. Historical data can be obtained from Lexis Nexis.
4. For Dutch, we have been collecting comments on news articles since November 2006. Due to the uniform structure of sites robust wrappers are easy to build. Even easier, comment data is increasingly made available through RSS.

Conclusion.

The political mashup is a challenging but technically feasible case of textual data interlinked by semantically rich concepts. From our experience in creating a website linking manifestos, news-stories and blogs on election topics [7] we have learned that 1. users find the links between data very important; 2. creating the links requires serious effort; and 3. a mashup of the kind we foresee does not scale without a well-designed data-structure for the links. The Linked Data philosophy seems well suited for the described political mashup project and we look forward to discussing just how well it fits at the workshop.

1. REFERENCES

- [1] J. Allan, editor. *Topic Detection and Tracking: Event based Information Organization*. Kluwer Academic Publishers, 2000.
- [2] T. Berners-Lee. Design Issues: Linked Data. <http://www.w3.org/DesignIssues/LinkedData.html>.
- [3] I. Budge et al. *Mapping Policy Preferences. Estimates for Parties, Electors, and Governments 1945-1998*. Oxford University Press, 2001.
- [4] S. Cucerzan. Large-scale named entity disambiguation based on wikipedia data. In *EMNLP-CoNLL '07*, pages 708-716, 2007.
- [5] A.K. Elmagarmid, P.G. Ipeirotis, and V.S. Verykios. Duplicate record detection: A survey. *Transactions on Knowledge and Data Engineering*, 19(1):1-16, 2007.
- [6] S. Fisher. The decline of parties thesis and the role of minor parties. In P. H. Merkl, editor, *Western European Party Systems: Trends and Prospects*, pages 609-613. New York, 1980.
- [7] V. Jijkoun, M. Marx, M. de Rijke, and F. van Waveren. Electoral search using the *verkiezingskijker*: an experience report. In *Proc. WWW '07*, pages 1155-1156, 2007.
- [8] H.-D. Klingemann et al. *Mapping Policy Preferences II. Estimates for Parties, Electors, and Governments in Eastern Europe, the European Union and the OECD, 1990-2003*. Oxford University Press, 2006.
- [9] I. Lipschits. *Verkiezingsprogramma's 1977 : verkiezingen voor de Tweede Kamer der Staten-Generaal*. SDU Staatsuitgeverij, 1977. Subsequent editions appeared for the elections in '81, '86, '89, '94, '98.
- [10] Th. Rochon. Mobilizers and challengers. towards a theory of new party success. *International Political Science Review*, 6(4):419-439, 1985.
- [11] Y. Song, J. Huang, I.G. Councill, J. Li, and C.L. Giles. Efficient topic-based unsupervised name disambiguation. In *Proceedings of the 2007 Joint Conference on Digital Libraries*, pages 342-351, 2007.
- [12] G. Voerman, A. Keyzer, F. den Hollander, and H. Druiven. Archiving the web: political party web sites in the netherlands. *European Political Science*, 2(1):68-75, 2002.