

Helping People to Choose for Whom to Vote. A Web Information System for the 2009 European Elections

Arjan Nusselder, Hendrike Peetz, Anne Schuth, and Maarten Marx
ISLA, University of Amsterdam
Amsterdam, The Netherlands
{anussel,hpeetz,aschuth,marx}@science.uva.nl

ABSTRACT

We demonstrate a web information system created for the European elections in June 2009. Based on their speeches in the EU parliament and their written questions, we created language models for each of the 736 members of the EU parliament. These language models were used to search for politicians responsible for a given topic, similar to expert search applications. Users prefer to see some kind of evidence for returning a hit after a search. We created a profile of each EU parliamentarian by comparing her personal language model to the language model created from all EU parliamentarians. The top 50 words best separating the individual from the average were shown as a wordcloud. These top 50 words and their scores were derived from a parsimonious language model.

Keywords

Information Extraction, Expert Search, EGovernance

General Terms

Experimentation

Categories and Subject Descriptors

H.4 [Information Systems Applications]: Miscellaneous

1. INTRODUCTION

Motivation for this research. The Dutch ministry of foreign affairs asked us to create a website to inform the public about the European parliament. This website was intended for the European elections in June 2009. Low turnout is a serious problem at European elections (in many countries it is lower than 40%). This is partly attributed to the unfamiliarity of voters with the European parliament, its parties and its members. We decided to use expert search and profiling technology developed within the TREC Enterprise Track to give voters the opportunity to find out which politicians are working on which topics.

Our approach. To match politicians to topics an approach named *expert finding* [1] was used. We used the parliamentary proceedings from 2004 till April 2009 of the European parliament to build language models for the members of parliament. For each member, we downloaded all her

speeches in Parliament and all her written questions, in all available languages. In this period, on average, documents from the EU parliament are translated into 12 languages. For each language, we concatenated all these documents and turned them into language models. We used the Indri system for this.

We assume topics are represented by news-articles, and we also build language models for these. We use the Kullback-Leibler divergence to match topics to politicians [5, 2].

Expert profiling is done using parsimonious language models [3]. We use a unigram language model to estimate probabilities and generate word clouds, where we assume that the most probable words are the most informative. The parsimonious probabilities are estimated using *Expectation-Maximization*:

$$\text{E-step: } e_t = tf(t, S) \cdot \frac{\lambda P(t|S)}{\lambda P(t|S) + (1 - \lambda)P(t|D)}$$

$$\text{M-step: } P(t|S) = \frac{e_t}{\sum_t e_t}, \text{ i.e. normalize the model}$$

where S is either a speech or a set of interruptions, and D is the complete debate i.e. our background model. In the initial E-step, maximum likelihood estimates are used. For λ we use a value of 0.01. In the M-step the words that receive a probability below our threshold of 0.0001 are removed from the model. In the next iteration the probabilities of the remaining words are again normalized. The iteration process stops after a fixed number of iterations. An earlier user-study showed that these wordclouds provide useful information about the topics that politicians are working on [4].

Demonstration. We demonstrate the described expert finding system for politicians, now available at <http://www.kieswijzer.eu/personen/expertsearch.php?lang=en&country=all&q=information+retrieval>. Search is possible in 6 European languages. The profiles are available in almost all languages, but, due to data sparseness, of best quality in the main languages (English, German, French). Visitors can experiment with the system and ask queries. We will also demonstrate our data-collection system.

Key technologies used and related work Building the system consists of three phases: (1) data collection and data-extraction, (2) creating the expert search engine, and (3) creating the profiles of the politicians, the countries and the parties. The first phase is basically an Extract-Transfer-Load (ETL) [6] process in which special attention has to be given to Named Entity Recognition and Normalization. The second phase is largely based on work done by Balog [1]. We used his *Model 1*, which describes the idea of

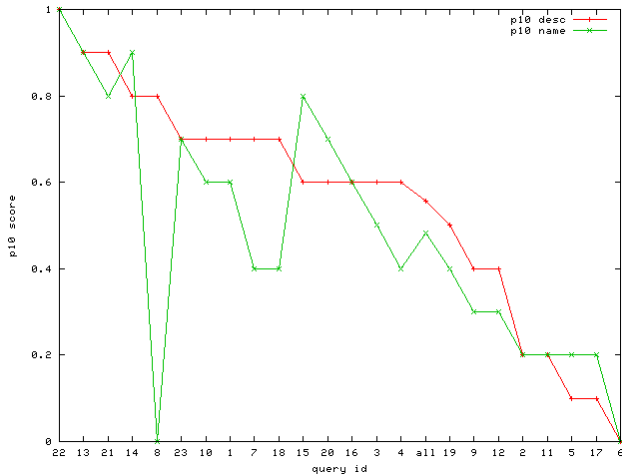


Figure 1: P@10 for the full description (desc) and the committee-names (name).

representing experts –politicians in our case– as single documents. Searching for experts then can be reduced to a document-comparison task. The latter task is implemented using language modelling techniques [5, 2]. Expert profiles are created from parsimonious language models [3].

2. EVALUATION

We carried out an experimental evaluation similar to the TREC 2005 W3C enterprise search task. We evaluated the expert retrieval system on Dutch data using 23 descriptions of committees as topics. The Dutch parliament has 23 committees, each focussed on a policy topic. Each committee consists of 8–25 members. For each committee its name, a short description and its members (all MP’s) are known. We did two evaluation runs: one with the committee names and one with the descriptions as topics. Committee names consist of 1 to 5 words (excluding stopwords); descriptions are between 500 and 1000 words. For instance, the description for the finance committee is 638 words (including stopwords).¹ A result (i.e., a politician) is correct (“relevant”) on a topic iff it is an active member of the committee described by that topic.

We measured the mean average precision (MAP) and precision at 10 (P@10) over two times 23 topics. The results are in Table 1.

	MAP	P@10
committee names	.38	.48
committee descriptions	.44	.56

Table 1: MAP and P@10 of our experiments.

Precision at ten is taken as an appropriate measure for two reasons. First, some committees have little more than ten members, which would make precision over ten difficult to evaluate. Second, the intended use of the application foresees a human-readable resultset. Figure 1 shows the P@10 for each topic for both evaluation runs (full description and

¹The description can be found at <http://www.tweedekamer.nl/kamerleden/commissies/FIN/sub/index.jsp>.

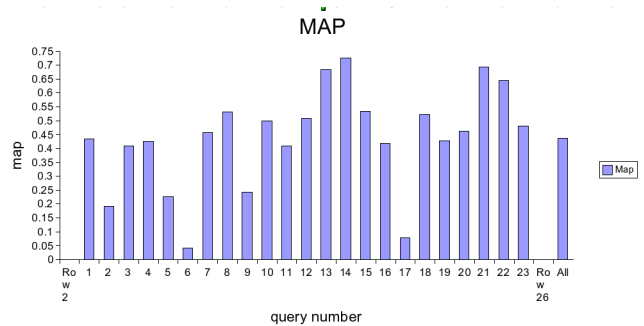


Figure 2: Mean average precision for each full text query.

the committee-name only), with the topics ordered by their P@10 for the description run. Figure 2 additionally shows the MAP score of each topic, ordered by topic id, for the full descriptions topics.

For the majority of topics –or committees– more than 6 from the first ten results were correct when we used the full description. Looking at figure 1, some possible problems can be identified. Query 8 shows a large discrepancy between the full description and the name only. This may be due to the fact that the topic –just the single word finance– can be and probably is used in virtually all contexts. The full text of the finance topic is descriptive enough to allow for a match between politicians focused on this area and the committee. The fact that almost all politicians will talk about financial issues however, could make the committee name by itself insufficient. Because the focus of the application lies on a search for more verbose text, this is not necessarily a problem.

Query 6 performs worse both with the full description and only the committee name. Several problems may be the cause of this. First, the committee itself consists –as an exception– of only eight members, which makes it harder to correctly retrieve the correct politicians. Also the topic of the committee is relatively new as compared to others, meaning there is probably less data available to create a profile that acknowledges this specific interest of the members. Third, the topic is pretty vague and seems rather specialized.

3. REFERENCES

- [1] K. Balog. *People Search in the Enterprise*. PhD thesis, University van Amsterdam, September 2008.
- [2] D. Hiemstra. *Using Language Models for Information Retrieval*. PhD thesis, University of Twente, 2001.
- [3] D. Hiemstra, S. Robertson, and H. Zaragoza. Parsimonious language models for information retrieval. In *Proceedings SIGIR 2004*, pages 178–185. ACM Press, New York NY, 2004.
- [4] R. Kaptein, J. Kamps, and M. Marx. Who said what to whom? Capturing the structure of debates. In *Proceedings SIGIR*, 2009.
- [5] J. Ponte and W. Croft. A language modelling approach to information retrieval. *Proc. SIGIR ’98*, 1998.
- [6] E. Rahm and H. Do. Data cleaning: Problems and current approaches. *IEEE Techn. Bulletin on Data Engineering*, 23(4), 2000.