

Expert Finding of Dutch Politicians

Arjan Nusselder
ISLA, University of Amsterdam
arjannusselder@uva.nl

Maarten Marx
ISLA, University of Amsterdam
KRuislaan 403 1098SJ Amsterdam, The
Netherlands
maartenmarx@uva.nl

ABSTRACT

A system is proposed and implemented that creates a language model for each member of the Dutch parliament, based on the official transcripts of the meetings of the Dutch Parliament. Using expert finding techniques, the system allows users to retrieve a ranked list of politicians, based on queries like news messages.

1. INTRODUCTION

Motivation for this research. The Dutch House of Representatives (*Tweede Kamer*) is supplied with information about current and past affairs by its information department, the *Dienst Informatievoorziening* (DI). The DI often pro-actively collects information about topics and events when they suspect one of the politicians will show a special interest in this topic. The DI also performs recommendations of “hot topics” to politicians likely to show an interest in that topic. The DI asked us to implement a system that automates this recommendation process.

Our approach. To match politicians to topics an approach named *expert finding* was used. This approach is detailed in section 4, and based on work by Balog [1]. We used the parliamentary proceedings to build a profile of each politician. A description of the data is given in section 3. The resulting system can be seen as answering the question: “Given the words spoken in parliament by a politician, how well does she match a given text?”

The current approach shows that, given well-structured parliamentary proceedings, it is possible to construct a good performing retrieval system using out-of-the-box information retrieval techniques. Our evaluation using committee descriptions suggests that the current approach has merit and could be explored further, incorporating more advanced techniques.

2. RELATED WORK

The current approach to the retrieval of politicians is based largely on work done by Balog [1]. We used his

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.

DIR2009 Enschede
Copyright 2009 ACM ...\$5.00.

Model 1, which describes the idea of representing experts –politicians in our case– as single documents.¹ This model itself is based on language modelling techniques [4][3].

3. DATA

We created a language model of each politician in the Dutch parliament (*Tweede Kamer*) in the summer of 2008. As textual input data we took the parliamentary proceedings which record everything being said in parliament. Through the PoliticalMashup project [2], this data is available in XML in a format which is excellent for our task: every word is annotated with the name of its speaker, her party and the date.

Besides these primary data sources we used biographical data about our politicians available at www.parlement.com.

4. METHOD

What needs to be expressed somehow, is the chance that a politician is knowledgeable on –or at least interested in– the topic expressed by a query. To do so, each politician must be represented with a profile. We first define such a profile as a document in which all text related to that politician is concatenated. This way, the politician–topic matching problem can be reduced to an instance of ranked document retrieval. To calculate the probabilities and ranking, the query is compared to all politicians, each represented as a language model of the concatenation of the related texts.

The measure used for comparison is the Kullback-Leibler divergence. We take $Q : Word \rightarrow Wordcount$ as the function over the words in the query, and $P : Word \rightarrow Wordcount$ as the function over the words in a document representing a politician. The basic formula to calculate the chance of a query given a politician is expressed in equation (1).

$$KL(Q|P) = \sum_i Q(i) \log \frac{Q(i)}{P(i)} \quad (1)$$

The result of a query is a ranked list of document identifiers, corresponding to the politician the texts belong to. To create an accessible and usable interface, the results are embedded in a block of additional information. At the time of writing, an interface is available at <http://zookma.science.uva.nl/politiciansearch/search.php>

For the actual implementation, the Lemur Toolkit was used.² The important Lemur parameters are *Simple KL* as

¹See Balog, section 3.2.1.

²See: <http://www.lemurproject.org>

6	Commissie voor de Verzoekschriften en de Burgerinitiatieven
8	Financien

Table 1: Names of topics 6 and 8, as they were used as query-text for the evaluation.

Commissie voor de Verzoekschriften en de Burgerinitiatieven Commissie Verzoekschriften en Burgerinitiatieven De commissie voor de Verzoekschriften en de Burgerinitiatieven heeft twee taken: het voorbereiden van een beslissing van de Kamer over een individuele aangelegenheid (waar een burger in een verzoekschrift om heeft gevraagd) en het voorbereiden van een beslissing van de Kamer over de ontvankelijkheid van een burgerinitiatief . . .

Table 2: Beginning of the description of topic 6. retrieval model, and for smoothing a *Dirichlet prior* set at the total number of word types.

Some additional ideas focussing more on the presentation of the results have been implemented. It is possible to not only collect texts on a per person basis, but also split the aggregations on a temporal or party level. Using a log-likelihood comparison, politicians can then be described as opposed to other politicians, or in a specific time-frame. Extensions like these could improve the usefulness of a system, but are left for future evaluation.

5. EVALUATION

To see how well our approach performs, an experimental evaluation similar to the TREC 2005 W3C enterprise search task was devised.³ The Dutch parliament has 23 committees, each focussed on a policy topic, roughly corresponding to the existing ministries⁴. Each committee consists of about eight to twenty-five members, and an equal or smaller number of reserve members. For each committee its name, a short description and its members (all MP's) are known. We used the both the committee names and their descriptions as topics. 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 (reserve members were not counted). The total number of candidates is 150, which is the number of current members of parliament.

Thus we do two evaluation runs, one with the names of the committees as topics, and one with the descriptions of the committees. 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).⁵ Table 1 gives two examples of committee names; Table 2 contains a part of the description of committee with topic id 6.

These longer descriptions match the purpose of our recommendation system more closely.

Results. We measured the mean average precision (MAP) and precision at 10 (P@10) over two times 23 topics. The results are in Table 3.

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 the committee-name only), with the topics ordered by

³See: <http://trec.nist.gov/>

⁴See: <http://www.tweedekamer.nl/kamerleden/commissies/index.jsp>

⁵The description can be found at <http://www.tweedekamer.nl/kamerleden/commissies/FIN/sub/index.jsp>.

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

Table 3: MAP and P@10 of our experiments.

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.

6. CONCLUSION

As asked for by the information department of the Dutch parliament, we created a recommendation system which matches current members of parliament to hot topics being described by a piece of text. These are typically news articles. We used an out-of-the-box expert search system based on Model 1 of [1] which showed promising performance using an evaluation similar to that of the TREC 2005 W3C enterprise search task.

A small evaluation (3 topics) which mimics exactly the use-case in mind (finding politicians likely to be interested in a news-story) gave even better results: all topics got a P@10 of .6 or higher. These results can be found at <http://zookma.science.uva.nl/politiciansearch/search.php>. Here the reader can also evaluate the system herself. Interesting queries are “ik” (*I*), “Nederland” (*The Netherlands*) and “vrede” *peace*.

7. REFERENCES

- [1] K. Balog. *People Search in the Enterprise*. PhD thesis, University van Amsterdam, September 2008.
- [2] T. Gielissen and M. Marx. Exemelification of parliamentary debates. In *Proc. DIR*, 2009.
- [3] D. Hiemstra. *Using Language Models for Information Retrieval*. PhD thesis, University of Twente, 2001.
- [4] J. Ponte and W. Croft. A language modelling approach to information retrieval. *Proc. SIGIR '98*, 1998.

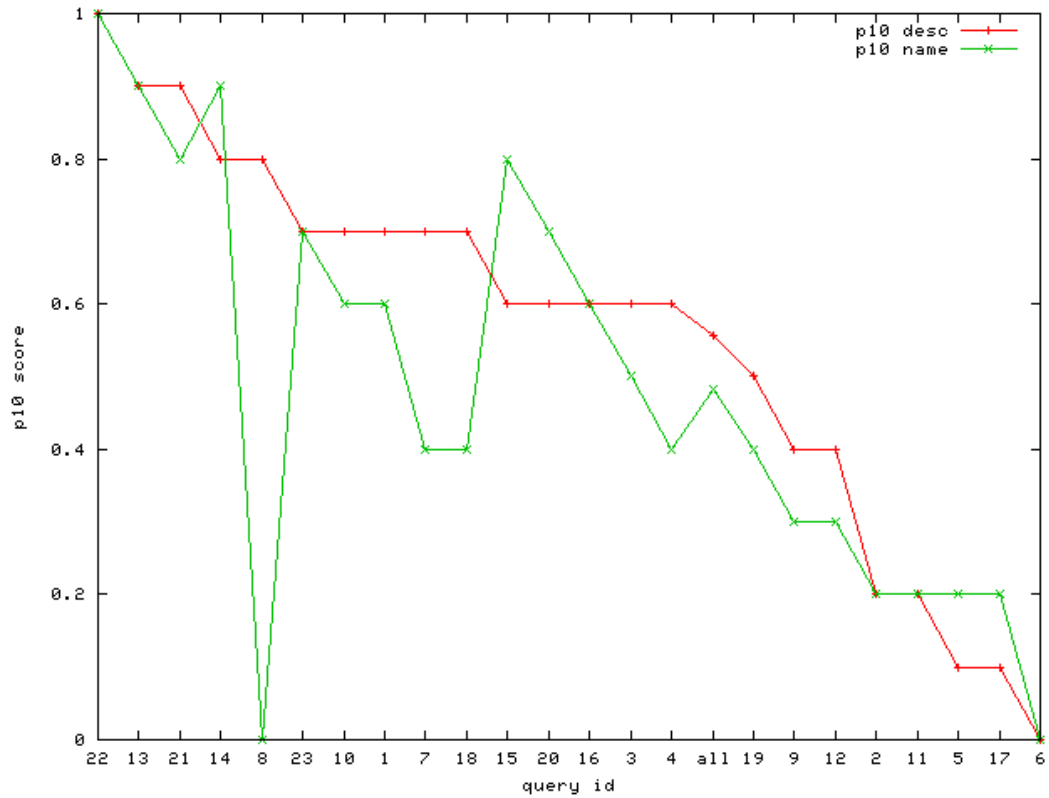


Figure 1: Precision at ten for the full description (desc) and the committee-names (name).

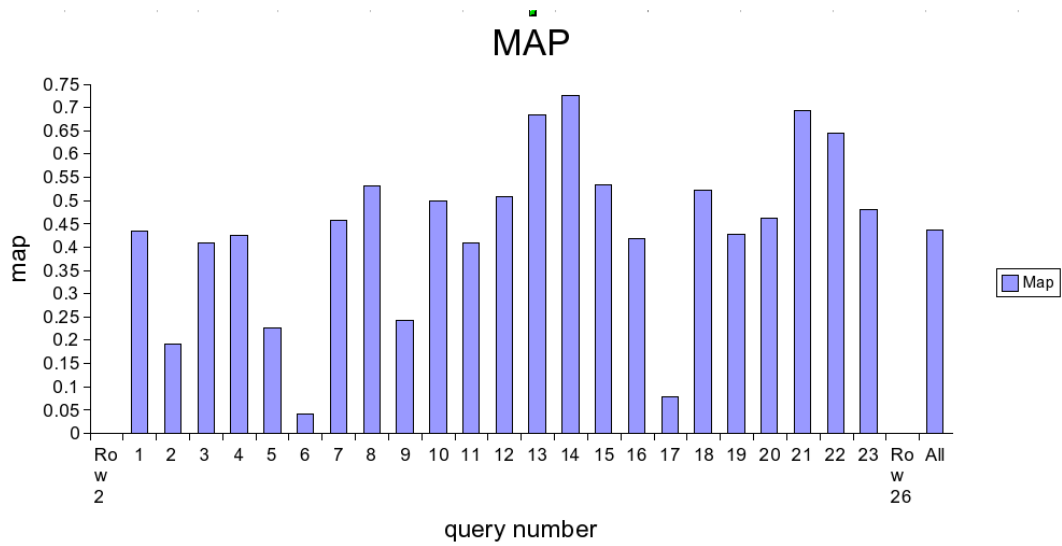


Figure 2: Mean average precision for each full text query.