Structuring Political Documents for Importance Ranking

Alexander Hogenboom, Maarten Jongmans, and Flavius Frasincar

Erasmus University Rotterdam
P.O. Box 1738
NL-3000 DR Rotterdam
The Netherlands
{hogenboom,frasincar}@ese.eur.nl,
maarten.jongmans@gmail.com

Abstract. Today's parliamentary information systems make political data available to the public in an effective and efficient way by moving from the classical document-centric model to a rich information-centric model for political data. We propose a novel approach to exploiting the rich information sources available through such parliamentary information systems for ranking results of a typical query for debates in accordance with their importance, for which we have developed several proxies. Our initial evaluation indicates that debate intensity and key players have an important role in signaling the importance of a debate.

1 Introduction

Since the beginning of this century, people's on-line hunger for information related to politics has increased dramatically. While 18% of all adults in the United States was estimated to consume political news on-line in 2000, 44% of all American adults searched the Web for political news in 2009 [9]. This phenomenon has been taking place in the Netherlands as well, which appears to have caused Dutch political parties to start embracing this new era in which information technology plays an increasingly important role in the political space.

Given the increasing importance of the Web in the political space, one of the main problems is that anyone, e.g., citizens, politicians, political (pressure) groups, or qualified journalists, can publish anything at any time, in such a way that the published information is accessible to anyone. This can result in vast amounts of on-line "news" in which opinions may be represented as facts. As such, it is crucial to have a reliable information source on the Web allowing easy access to information produced in parliament, e.g., meeting notes or voting records. This is of paramount importance, as existing parliamentary information platforms are mainly visited by the public (44%) and businesses (24%) [8]. Such information sources can thus help bridging the gap between the public and the government, while forming a concrete foundation of democracy and providing a starting point for a true e-democracy.

G. Bouma et al. (Eds.): NLDB 2012, LNCS 7337, pp. 345–350, 2012.
© Springer-Verlag Berlin Heidelberg 2012

Existing systems like PoliDocs [4] have already taken promising steps towards making political information easily accessible to the public. The nature of data published through such systems introduces new challenges for ranking results of queries on political data. As political documents form a natural collection for search tasks in which answers do not typically consist of documents but rather of information hidden in one or more documents [6], recent developments exhibit a tendency of moving from a document-centric to a structured, information-centric model for political data. Such a model enables linking concepts in various political documents, yet the typical incompleteness of these references thwarts the applicability of well-known query result ranking methods such as PageRank [1] and HITS [7]. In our current endeavors, we propose a way of using the information obtained from structured political documents in order to rank debates in accordance with their importance as perceived by political experts.

The remainder of this paper is structured as follows. First, we discuss some related work on political information systems in Sect. 2. Then, Sect. 3 demonstrates how well-structured political information can be used in order to rank debates in accordance with their importance, which is evaluated in Sect. 4. Last, in Sect. 5, we draw conclusions and propose directions for future work.

2 Parliamentary Information Systems

In order for parliamentary information systems to effectively and efficiently disclose political information to the public, political data must be available through an information-centric data model. Unfortunately, most political data has until now been published as unstructured natural language text. Governments have only fairly recently begun to become aware of the need for political data in a more structured data format like XML. Therefore, in order to be able to disclose the wealth of information hidden in past, present, and future parliamentary publications in natural language text, a principal method for converting natural language text into structured political information is of paramount importance.

One of the first attempts of structuring political data is conceptually surprisingly simple. Gielissen and Marx [3] take a collection of political documents in PDF format and convert these documents into an XML format in which each line of the original text is annotated with the properties of its bounding box. By applying some information retrieval heuristics on the text, the resulting XML file is subsequently enriched with additional annotations for concepts of interest in the political domain, e.g., political parties, members of parliament, etcetera.

With parliamentary data being available in a structured format, the possibilities are numerous. For instance, some researchers have explored structured political data for traces of sentiment, by automatically determining subjectivity in parliamentary publications and by subsequently determining the semantic orientation of the identified subjective parts [5]. The potential of structured data in parliamentary information systems has also been demonstrated by the utilization of such data in order to facilitate faceted search [4].

As political data is predominantly queried for information that does not typically consist of full documents but rather of information hidden in one or more documents [6], ranking original political documents for relevance with respect to a query is a far from trivial task. Well-established ranking techniques like PageRank [1] and HITS [7] are suitable for documents or concepts with many interconnections, but the incompleteness of the links in today's structured political data requires a different ranking approach for search results in political data. Ontology-based approaches [2], in which ranking is based on the (lack of) appearance of domain concepts, may seem a suitable alternative. Yet, in political recordings, it may not so much be the concepts that are discussed that make a document relevant, but rather the way in which these concepts are discussed. A ranking method that takes into account this phenomenon is yet to be proposed.

3 Importance Ranking of Debates

In order to make political documents accessible to the public through parliamentary information systems, the information contained by such documents needs to be processed first. We propose to follow a state-of-the-art method [3] by converting political documents into an XML format containing a structure which primarily models the documents' layout and subsequently using heuristics in order to add information into the structure of the XML such that a query mechanism like XQuery can be used to retrieve information to answer specific questions.

The data can be queried by means of techniques exploiting the newly introduced structure. An interesting application is to retrieve information on debates that are possibly interesting, given a user's query. Common methods of ranking query results rely on graphs of interlinked items. Although debating report documents have some references to documents containing, e.g., motions or amendments, these references are typically incomplete, thus rendering such methods less applicable to ranking debates. Therefore, we propose a novel method for ranking debates in accordance with their importance.

We define importance as the probability that the public finds a debate of great significance. We model importance as a three-dimensional construct. The first dimension is the intensity of a debate. In intense debates, people argue and interrupt one another a lot. When this happens, a debate often receives a lot of attention from the media, which suggests its importance to the public.

The second dimension of our importance measure is constituted by the quantity and quality of key players in a debate. Here, quantity refers to the number of participants in a debate, whereas quality is constituted by the people participating in a debate. For instance, the presence of the (deputy) prime minister or the number of participating floor leaders may signal the importance of a debate.

Last, the third dimension of our construct is formed by the debate length. This length refers to how much time is required for a debate. A debate can take very long when parties do not agree with each other or with the executive branch. Alternatively, a debate may consume a lot of time when a topic is so complex and important that it requires a lot of time to be discussed properly.

We propose to operationalize the three dimensions of importance by means of several attributes that can be extracted from the structured parliamentary data by means of a query mechanism like XQuery. We assume the importance I_d of a debate d to be a function of its n attributes $x_{d1}, \ldots, x_{dn}$ and their associated weights $\beta_1, \ldots, \beta_n$, i.e.,

$$I_d = \sum_{i=1}^{n} \beta_i x_{di}, \quad 0 \leq I_d \leq 1, \quad 0 \leq x_{di} \leq 1. \tag{1}$$

The attributes in (1) are distributed over the three dimensions of our importance construct. The intensity of a debate is operationalized as the amount of switches between speakers in a debate. The attributes constituting the dimension of key players are the percentage of attending members of parliament, the presence of the (deputy) prime minister, the percentage of speaking floor leaders, the number of members of parliament speaking in a debate, and the amount of members of the executive branch of the government speaking in a debate. The debate length is operationalized by means of the total number of words spoken in a debate, the time a debate closes, the number of blocks in a debate, and the number of times the executive branch refers to a possible second term.

The range of each variable is limited to the interval $[0, 1]$. We apply min-max normalization to the number of interruptions, such that its minimum is mapped to 0, and its maximum is mapped to 1. The percentage of members of parliament attending a debate is a value between 0 and 1 and as such does not need further normalization. The presence of the (deputy) prime minister is encoded as either 0 (not present) or 1 (present). Furthermore, the percentage of floor leaders speaking is recoded as 1 (percentages higher than 0.9), 0.3 (percentages between 0.6 and 0.9) or 0 (all other cases). If there are less than nine speakers, this recoded value is multiplied with 0.5. Additionally, we apply min-max normalization to the number of members of parliament speaking in a debate. The number of speaking members of the executive branch is recoded to 1 (three or more speaking members), 0.35 (two speaking members), or 0 (all other cases). Furthermore, we apply min-max normalization to the word count as well as to the number of minutes a debate ends after midnight. The number of blocks is encoded as either 0 (ten blocks or less) or 1 (more than ten blocks). Last, second term references are encoded as 0 (no references) or 1 (one or more references).

4 Evaluation

In order to evaluate how our proxies contribute to a debate's importance, we consider five different methods for distributing the attribute weights. First, we consider giving each attribute an equal weight of 0.091. Alternatively, we consider distributing all weight over attributes related to either the intensity dimension (a weight of 1.000 for the single attribute related to this dimension), the key players (where the six attributes are assigned a weight of 0.166), or the debate length (resulting in four attributes with a weight of 0.250). Last, we perform a Multivariate Linear Regression (MLR) analysis in order to optimize the weights.

Our evaluation is performed on a set of parliamentary recordings of 100 debates held in the Netherlands between January 1, 2009, and March 31, 2010, retrieved through PoliDocs [4]. These debates have been structured by means of a state-of-the-art approach [3]. Through a survey, 9 out of 17 approached Dutch political experts (analysts, scientists, etcetera) have assigned each of our 100 selected debates to the *Top 10*, *Top 11–20*, or *Top 21–30* debates, or an *Unordered* category for the remaining 70 debates. We have aggregated the expert rankings by first distributing 100 points over all debates for each survey. Each *Top 10*, *Top 11–20*, and *Top 21–30* debate received four, three, and two points, respectively, and the remaining points were equally distributed over the unranked debates. We have then averaged each debate's score over all nine surveys.

Our data is split into a training set for the MLR analysis (60%) and a test set (40%) for comparing the ranking of debates as produced by our five methods with the ranking made by political experts. This comparison is done by means of a P@k test, where we assess the percentage of the top k debates of our methods occurring in the experts' top k debates, for $k \in \{10, 20, 30\}$.

The weights of our models are reported in Table 1. The number of speaker switches has a relatively high positive correlation with a debate's importance. The presence of the (deputy) prime minister and parliament members, the number of floor leaders speaking, and the debate closing time exhibit a rather positive correlation too. Interestingly, the number of executive branch members speaking exhibits a negative correlation with a debate's importance in our data set.

Table 2 reports the precision of the top 10, 20, and 30 documents returned by our considered models. Of our four dimension-based models, both the model focusing on the intensity dimension and the model focusing on the dimension of key players perform comparably well. This is in line with the observed weight distribution found by our MLR analysis, which emphasizes attributes related to both of these dimensions. However, our MLR model appears to have a more stable overall performance. It outperforms all other models in terms of precision on the top 10 and 20 documents, while exhibiting a performance that is comparable with the other models on the top 30 documents.

Table 1. Attribute weight configurations per model

Attribute	Equal	Intensity	Players	Length	MLR
Speaker switches	0.091	1.000	0.000	0.000	0.195
Parliament members present	0.091	0.000	0.166	0.000	0.133
Prime minister	0.091	0.000	0.166	0.000	0.181
Deputy prime minister	0.091	0.000	0.166	0.000	0.144
Floor leaders speaking	0.091	0.000	0.166	0.000	0.135
Parliament members speaking	0.091	0.000	0.166	0.000	0.046
Executive branch speaking	0.091	0.000	0.166	0.000	-0.132
Word count	0.091	0.000	0.000	0.250	0.107
Closing time	0.091	0.000	0.000	0.250	0.149
Block count	0.091	0.000	0.000	0.250	-0.106
Second term	0.091	0.000	0.000	0.250	-0.050

Table 2. Importance ranking performance per model

Precision	Equal	Intensity	Players	Length	MLR
P@10	20%	30%	30%	10%	40%
P@20	55%	50%	55%	55%	65%
P@30	70%	70%	73%	77%	73%

5 Conclusions and Future Work

We have proposed a novel way of exploiting an information-centric model for political data in order to rank results of a typical query for parliamentary debates in accordance with their importance. To this end, we have developed several proxies for a debate's importance. Our results indicate that debate intensity and key players have an important role in signaling the importance of a debate.

A more extensive survey, in which political experts provide a more detailed ranking of more debates, may bring additional insights into what constitutes an important debate. Another direction for future work is to investigate the possibility of a non-linear relation between our proxies and a debate's importance.

Acknowledgments. The authors are partially supported by the Dutch national program COMMIT.

References

1. Brin, S., Page, L.: The Anatomy of a Large-Scale Hypertextual Web Search Engine. In: 7th International World-Wide Web Conference (WWW 1998), pp. 107–117. Elsevier (1998)
2. Frasincar, F., Borsje, J., Levering, L.: A Semantic Web-Based Approach for Building Personalized News Services. International Journal of E-Business Research 5(3), 35–53 (2009)
3. Gielissen, T., Marx, M.: Exemelification of Parliamentary Debates. In: Ninth Dutch-Belgian Workshop on Information Retrieval (DIR 2009), pp. 19–25 (2009)
4. Gielissen, T., Marx, M.: The design of PoliDocs: a Web Information System for the Disclosure of Dutch Parliamentary Publications. In: Sixth International Workshop on Web Information Systems Modeling (WISM 2009), vol. 461. CEUR-WS (2009)
5. Grijzenhout, S., Jijkoun, V., Marx, M.: Sentiment Analysis in Parliamentary Proceedings. In: Workshop From Text to Political Positions, T2PP 2010 (2010)
6. Kaptein, R., Marx, M.: Focused Retrieval and Result Aggregation with Political Data. Information Retrieval 13(5), 412–433 (2010)
7. Kleinberg, J., Kumar, R., Raghavan, P., Rajagopalan, S., Tomkin, A.: The Web as a Graph: Measurements, Models, and Methods. Springer (1999)
8. Marcella, R., Baxter, G., Moor, N.: The Effectiveness of Parliamentary Information Services in the United Kingdom. Government Information Quarterly 20(1), 29–46 (2003)
9. Smith, A.: The Internet's Role in Campaign 2008. Tech. rep., Pew Internet and American Life Project (2009)