# Electoral Search Using the VerkiezingsKijker*

## An Experience Report

Valentin Jijkoun Maarten Marx Maarten de Rijke Frank van Waveren
ISLA, University of Amsterdam
Kruislaan 403, 1098 SJ Amsterdam, The Netherlands
jijkoun,marx,mdr,fwaveren@science.uva.nl

## ABSTRACT

The Netherlands had parliamentary elections on November 22, 2006. We built a system which helped voters to make an informed choice among the many participating parties. One of the most important pieces of information in the Dutch election and subsequent coalition government formation is the *party program*, a text document with an average length of 45 pages. Our system provides the voter with focused access to party programs, enabling her to make a topic-wise comparison of parties' viewpoints. We complemented this type of access ("What do the parties promise?") with access to news ("What happens around these topics?") and blogs ("What do people say about them?"). We describe the system, including design technical details, and user statistics.

## Categories and Subject Descriptors

H.4.m [**Information Systems**]: Miscellaneous; D.2 [**Software**]: Software Engineering

## General Terms

Design, Experimentation

## Keywords

Elections, democracy, domain specific search

## 1. INTRODUCTION

We describe the VerkiezingsKijker ("election watcher"), an electoral search engine aimed at helping the general public in its electoral decision making.[1] Based on interest from real users, on user feedback and media coverage, we believe that this application of search and language technology is one of wide interest. We motivate the choices made in our design, describe the technical challenges and our solutions, and argue that our findings are applicable to the more general problem of making decisions when faced with competing solutions (offered as poorly structured textual information).

*This research was supported by the Netherlands Organization for Scientific Research (NWO) under project numbers 017.001.190, 220-80-001, 264-70-050, 354-20-005, 600.065.120, 612-13- 001, 612.000.106, 612.066.302, 612.069.006, 640.001.501, 640.002.501, and by the E.U. IST programme of the 6th FP for RTD under project MultiMATCH contract IST-033104.

[1]VerkiezingsKijker is available at `www.verkiezingskijker.nl/2006`.


*WWW 2007,* May 8–12, 2007, Banff, Alberta, Canada.
ACM 978-1-59593-654-7/07/0005.

## 2. REQUIREMENTS AND DESIGN

The most recent Dutch parliamentary elections were held on November 22, 2006. Members of the Dutch parliament (150 seats) are chosen according to the principle of proportional representation. In 2006, 65,591 votes were needed for a single seat (`http://www.kiesraad.nl`). This system leads to a proliferation of political parties; in general, some twenty parties participate in the national elections (in 2006: 24 parties), each with its own party program (in 2006: with an average length of 45 pages).

Asked to set up a search engine for party programs by the Instituut voor Publiek en Politiek (IPP[2]), a public-private non-profit organization aimed at bringing politics and the general public closer together, we identified three groups of requirements. User's requirements included paragraph-based access to party programs, providing both thematic search (with themes based on previous elections and current issues) and free-text search; facilities to compare parties' viewpoints on topics; integration with additional sources of information (news and blogs), and ways of identifying important events and trends in the latter sources. Developer's requirements concerned the gathering of domain knowledge (specifically, the themes for the thematic search facility) and the data preprocessing effort (with the party programs becoming available at a late stage). System requirements boiled down to the use of open source, off-the-shelf technology, the provision of a simple API to the search engine, and robustness. We decided on the design given in Figure 1.

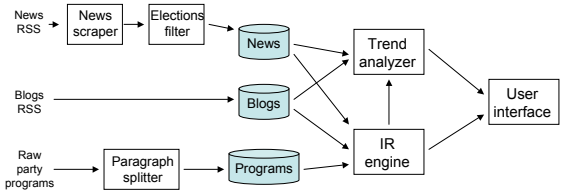


**Figure 1: Architecture of the system.**

## 3. IMPLEMENTATION

We describe the implementation of the VerkiezingsKijker search engine in two steps: *components* and *overall*.

[2]`http://www.publiek-politiek.nl/english`

*Preprocessing and Indexing Components.* We processed three types of data—party programs, news feeds and blog feeds—and had to generate expansion terms to enable thematic search.

Sixteen (of the 24 participating) parties made available their programs, 2 in HTML, and the rest in Word or PDF. Programs were automatically split into paragraphs based on layout, yielding a corpus of 4618 paragraphs. On average, half an hour of extra manual work per program was needed.

We decided to implement thematic search as regular free text search, but with queries consisting of a theme (e.g., "education") and a number of additional terms. For each of the 179 themes proposed by IPP (our customer) we asked a domain expert to (use the search engine to) identify at least 5 relevant paragraphs. For each theme we collected the top-15 overused terms as characteristic for the topic. "Overusage" was determined using the log-likelihood statistical test [2], comparing the paragraphs marked relevant with the set of all paragraphs in the index. Terms likely to lead to topic drift were manually removed.

The other data sources used by the VerkiezingsKijker (news and blogs) needed frequent and repeated harvesting, extraction and indexing. As we were dealing with national elections, we restricted ourselves to feeds of nation-wide daily newspapers, and included eight such newspapers, covering the entire political spectrum. For these we obtained the HTML articles, extracted the text content from the HTML, classified the contents into election-related vs non-election-related, and indexed it. For extraction, a robust, unsupervised method based on block length was used [4]. For classificaton, we used a Naive Bayes classifier, which helped us increase the proportion of election-related articles from around 20% (prior to classification) to well over 90%.

As our source of (Dutch) blogs we used `http://web-log.nl`, one of the largest Dutch weblog hosts. At the time of the elections, 43,984 blogs were hosted with an average of 4,179 postings per day. Within the measured time-period there were 7,768 active bloggers (having at least one post a week). We did not perform election-related filtering on blog posts. Because we obtained clean data from the blog host, no additional cleanup was needed. Similar to news items, blog posts were indexed for retrieval and stored in a database, along with meta-data (blogger, URL, publication date and time, etc.).

*Putting Things Together.* VerkiezingsKijker allows users to search the three sources (party programs, news, and blogs), either by theme or free text. In addition, the system provides "trend" functionality for news and blogs: a visualization of the volume of news items or blogposts relevant to a theme or query, peak detection and explanation.

For search on programs, VerkiezingsKijker responds with a list of paragraphs ordered by party (all, or a selection) or relevance. For news and blogs, results can be ordered by relevance or publication date. The system is implemented using Lucene [3] for retrieval in programs, news and blogs, and a MySQL database for data storage. As to trends, the system displays counts of news items or blogposts relevant to topics, identifies peaks (comparing actual counts against expected counts based on earlier observations), and provides explanations of unusual peaks in blogpost counts on a topic by generating links from blogposts in peak periods to related news items, using the method described in [1].

| Query | English | # |
|---|---|---|
| kinderbijslag | child allowance | 5314 |
| minister-president gekozen | elected prime-minister | 5252 |
| kinderopvang | kindergarten | 4464 |
| Turkije | Turkey | 3969 |
| ontslagrecht | law governing dismissal | 3284 |
| bijstand | social security | 3123 |
| meningsuiting | freedom of speech | 3069 |
| dieren | animals | 2754 |
| nationaliteit | nationality | 2664 |

**Table 1: Ten most popular free-text search queries.**

# 4. RESULTS

VerkiezingsKijker went online on October 23, 2006 (about a month before the elections). Here are some statistics for the period of five weeks between October 23, 2006 and November 30, 2006:

- 109,954: the number of unique IP hosts accessing the system; 20,624 unique hosts (19% of the total) accessed the system on the day of the elections;
- 76,360: the number of unique IP hosts that used the search facilities of the system;
- 148,026: the total number of searches made in the system, in particular:
  - 117,132: the number of free text searches;
  - 28,025: the number of thematic searches;
  - 2,788: the number of free text trend requests;
  - 81: the number of thematic trend requests;
- 6,014: the number of distinct free text queries;
- 175: the number of distinct thematic queries (out of 179 available themes).

The distribution of the actual frequencies of search queries follows a power law and, moreover, the 40 most frequent free text queries (1% of all distinct queries) account for 80% of all free text searches in the system. Table 1 lists frequencies of the most popular free text queries (targeting party programs).

# 5. CONCLUSION

The main contribution of the poster is best summarized as a recipe describing how to use *off-the-shelf technology* to quickly build a web accessible search engine for cases which resemble our scenario: i.e., to create support for users that need to make an informed choice among several competitors which drown the choice-maker in textual, mostly unstructured, information, with multiple perspectives.

# 6. REFERENCES


[1] K. Balog, G. Mishne, and M. de Rijke. Why are they excited? In *Proceedings EACL 2006*, April 2006.

[2] T. Dunning. Accurate methods for the statistics of surprise and coincidence. *Comput. Ling.*, 19(1):61–74, 1993.

[3] Lucene. The Lucene search engine. `http://lucene.apache.org/`.

[4] F. van Waveren. Extracting and classifying election-related news items from the world wide web. Master's thesis, University of Amsterdam, 2006.