

Long, often quite boring, notes of meetings

www.polidocs.nl

Maarten Marx

Universiteit van Amsterdam

February 2009

Outline

- **Context** Political Mashup Project
- **Techniques** Text Extraction, Semantic Annotation, XML
- **IR Aspects** Focused retrieval, result aggregation
- **Demo:** Exploit semantic XML structure



Political Mashup Project

- **Goal:** making political data easily accessible
- **Easy means**
 - good entry points and natural answer units
 - results ranked by **relevance**



Parliamentary Data

- Transcripts of meetings of Parliament
- 35 years of data until yesterday.
- Unit of retrieval (i.e. document) in the current system is
 - ★ notes of 1 day, =
 - ★ 80 pages 2 column PDF, corresponds to
 - ★ 10 ACM SIGIR articles,

One document is a small corpus

Information retrieval, not document retrieval

- Storage unit does **not** correspond to unit of information needs.
- Notes of one day cover
 - several **agenda topics**,
 - each divided in **speeches** by MP's.

Speeches are natural entry points to a debate

Structure of debates

- Example

```
<topic topic='...' page='..' ...>  
  <block lectern='Wilders' ...  
    <speaker name='Wilders' anchor='..' ...>  
      <p>  
        Mevrouw de voorzitter. Om te  
        beginnen mijn oprechte dank aan u persoonlijk omdat u  
        op mijn verjaardag vandaag een debat over de islam  
        heeft gepland. Een mooier cadeau had ik mij niet kunnen  
        wensen!  
      </p>  
      ...  
    </speaker>  
  </block>  
</topic>
```

From flat PDF to deep XML

1. `pdftohtml -xml,`
2. repair XML,
3. add semantic annotations using (regex) patterns
4. from flat to deep structure with `xsl:for-each-group`

Example

De heer **Van Gerven** (SP): Ik heb gezien dat de amendementen inmiddels in Parlendo staan.

vergeetende, staat in de steek van de minister.

De SP maakt zich zeer grote zorgen over de kwetsbare positie van kinderen. Punt is dat, wanneer ouders niet verzekerd zijn, zij zorg voor zichzelf, maar ook voor het kind mijden. Dit is niet in overeenstemming met het

Tweede Kamer

Zorgverzekeringswet

2 april 2008
TK 71

71-4975

```
<speech page="71-4975" anchor="568"
  party="SP" name="Van Gerven"
  PDCid="03116"
  role='MP'
  gender='m'>
```

Ik heb gezien dat de amendementen inmiddels in Parlendo staan.

```
</speech>
```

Deduplication of typos and changes

Aadsted-Madsen
Aasted Madsen
Aasted-Madsen
Aasted Madsen-Stiphout
Aasted Madsen - van Stiphout
Aasted-Madsen van Stiphout
Aasted-Madsen-van Stiphout
Aasted Madsen-Van Stiphout



Deduplication of OCR errors

FvdA FvdA FVdA
PcdA pdA PdA PF
P>dA Pvc)A PvciA
PvciA PvclA pvdA
Pvd.A PvdA PvDA
PviiA PvriA PvrJA
PvtiA PWD PydA

Exploitation

- Focused retrieval [Kamps, Geva, Trotman '08]
 - ★ Entry point retrieval
 - ★ Faceted search [Hearst '06, '08]
 - ★ Cheap Video Annotation
- Result aggregation [Murdock&Lalmas '08]
 - ★ TermClouds
 - ★ Structure summaries
 - ★ Topic tracking
 - ★ Attackogram

Entry point retrieval

- Search system returns ranked list of **speeches** which are **entry points** to a debate on a topic.
- Aggregates results by **year**, **party** and **politicians**

<http://www.polidocs.nl>

Video search

- **Golden opportunity**: rich, cheap, fine-grained annotations of video
- **Syncing** can be done by speech and speaker recognition
- Demo op <http://www.openkamer.nl/>

```
<spreker naam="De voorzitter" pagina="101-7178" anker="66"  
  timecode="00:07:07">
```

```
Wij gaan niet het cv van de heer Bosma doornemen.
```

```
</spreker>
```

TermClouds

- **Technique:** log-likelihood [Rayson and Garside 2000]

The screenshot shows a Firefox browser window displaying a document from PoliDocs.nl. The document text discusses cable television and includes a section titled "Jasper van Dijk (SP) zei:". A blue term cloud is overlaid on the text, containing the terms: "baan Bosma directeur", "multicultigehalte", "omroep uitstek", and "zender".

Firefox File Edit View History Bookmarks Tools Window Help
PoliDocs.nl - Mediawet
http://staff.science.uva.nl/~marx/pub/www/HAN8308A10.output.N.xml#51
Most Visited We-think Postbank marx wiki Blackboard Marx TWiki frmget
Podcasten enzo Stichting Rinco mdr-bw.jpg (JPEG Image, 127x... Boris van der Ham PoliDocs.nl - Mediawet

van kabelstroom. De kabelnetten zijn inderjare voor een upper en een er verkocht aan grote bedrijven. Deze netten, die zijn opgebouwd met gemeenschapsgeld, zijn nu een bron van megawinsten. Mijnheer John Malone lacht zich in Arizona een gat in zijn broek over de "windfall profits" die hij met UPC boekt. Uit berekeningen blijkt dat de kostprijs per kabelhuishouden op ongeveer € 6 à € 8 per gezin per maand liggen, terwijl een abonnement ongeveer € 16 kost. Dat is een winst van meer dan 100%. Het enige wat in deze situatie verandering kan brengen, is concurrentie. En alleen de kabel kan op analoge basis concurreren met een kabel, want kijkers willen nu eenmaal meerdere televisietoestellen in huis. Daarbovenop willen zij digitale keuzemogelijkheden. Het zou een mooi beleidsdoel zijn voor deze minister om te zorgen voor concurrentie op de kabel door meerdere aanbieders toe te laten die samen de prijs naar beneden gaan concurreren. De huidige kabel moet met andere woorden veranderen van distributeur naar transporteur. Dat is technisch mogelijk. Het zou vervelend zijn voor mijnheer Malone, maar goed voor de Nederlandse consument.

Link: <http://www.polidocs.nl/XML/HAN/HAN8308A10.xml#60>

Jasper van Dijk (SP) zei:
De heer Bosma is niet ... omroep en al helemaal niet over het door hem geconstateerde multicultigehalte van di ... aan als directeur van de zender Colorful Radio? Dat was namelijk bij uitstek er ...

Link: <http://...>

baan Bosma directeur
multicultigehalte
omroep uitstek zender

der Imca Media. Onder de Nederlandse Radio Groep vielen ... SP meldt dat ik directeur was van deze zender, maar dat klopt dus ... zender, maar een R&B-zender. Na de nieuwe verdeling van de FM ... valt in ... onderdaad verklaard een multiculturele zender te zijn, maar dat

Link: <http://...>

De voorzitter (I
Zullen wij een poging doen, het niet te persoonlijk te maken?

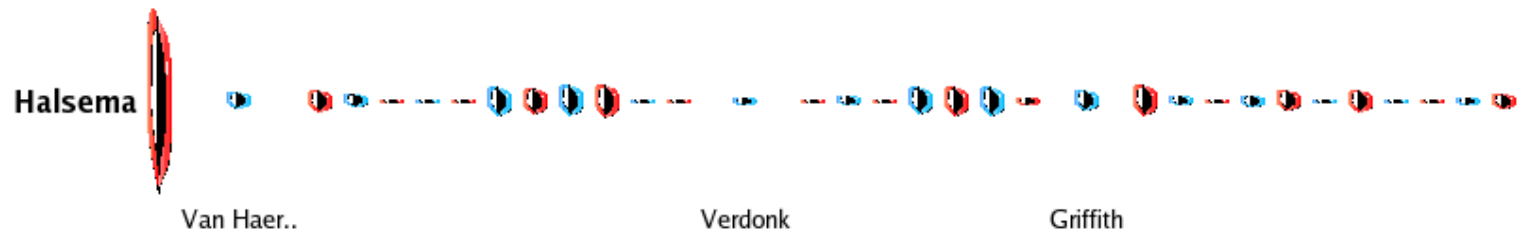
Link: <http://www.polidocs.nl/XML/HAN/HAN8308A10.xml#63>

Find: knett Next Previous Highlight all Match case

Result summarization: 3 examples

1. Debate timelines
2. Topic tracking
3. Attackogram

Debate timelines

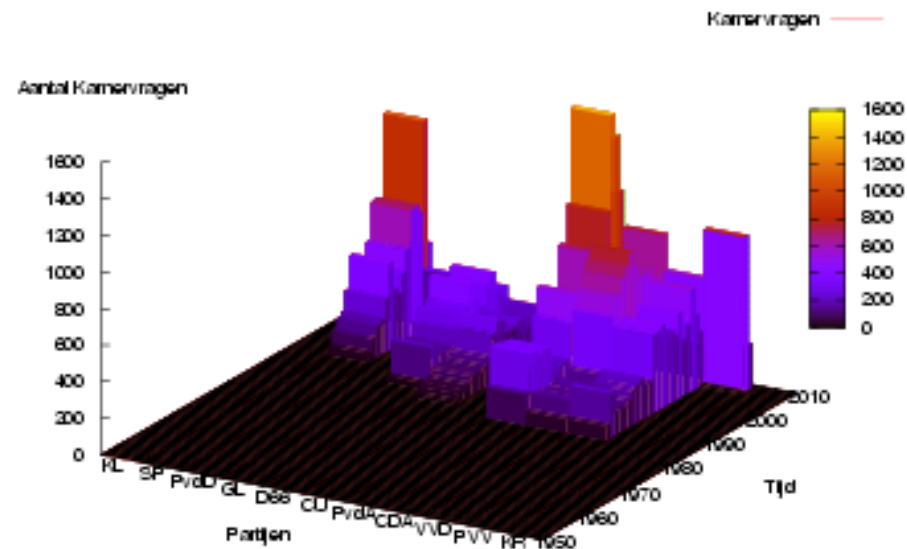


1. Summarizes structure
2. Functions as a hyperlinked table of contents
3. Termcloud comes out of the mouth by mouse-rollover

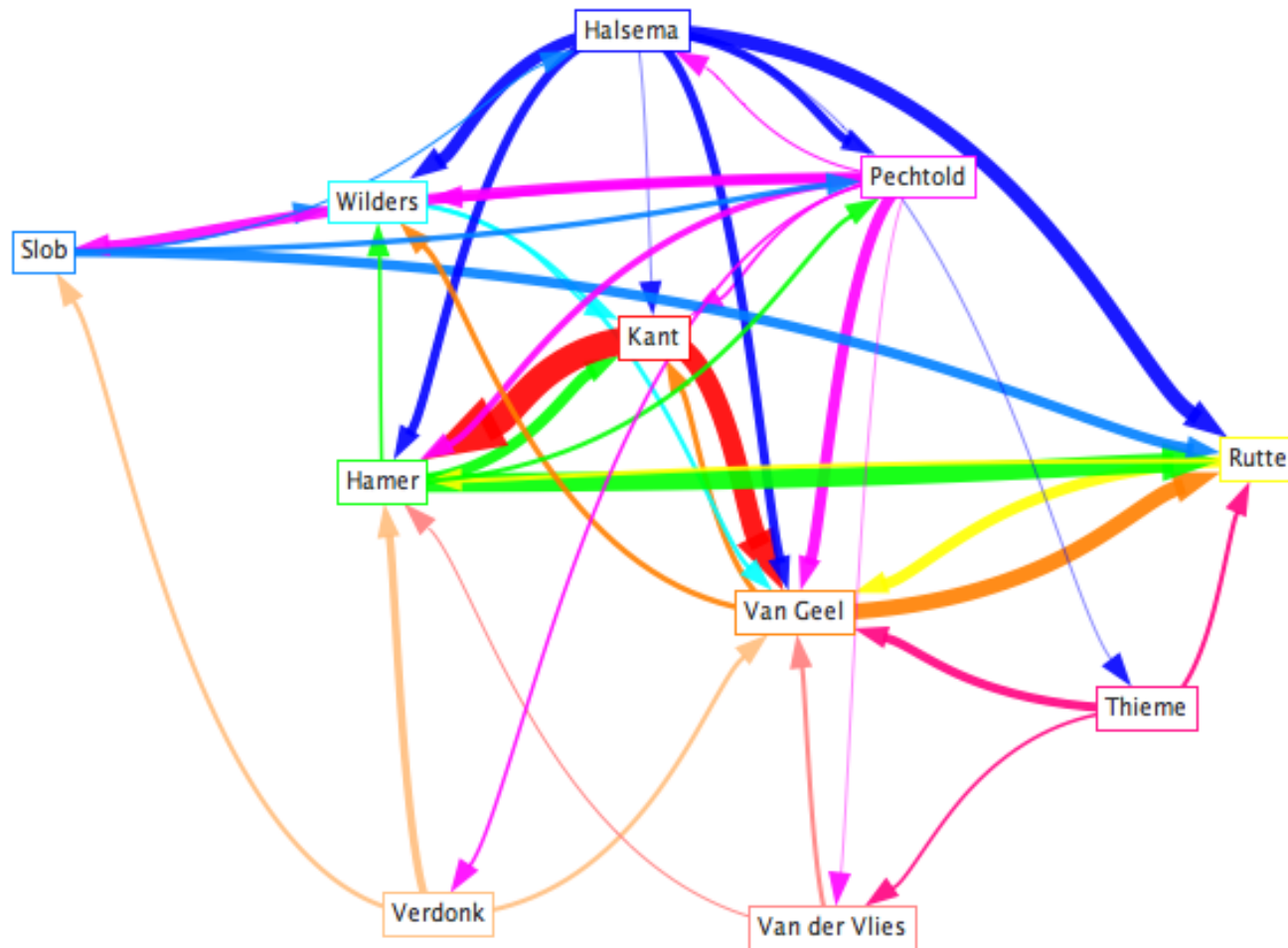
Algemene Beschouwingen 17 Sep 2008

Topic tracking

- development of a topic over time and across parties
- 3D browser



Summarization



Attackogram by [Rianne Kaptein](#).

Conclusions

- Parliamentary debates are full of **extractable structure**
- Same techniques and analysis applies to other countries
- Current digitization efforts provide many opportunities
- A lot of structure to exploit ...
- **Next steps:**
 1. integrate data from different EU countries
 2. **Just XML, nothing else** use MonetDB/XQuery and PFTijah as backend.

The final step

