Why and how is this a "related document"?: Semantics-based analysis of and navigation through heterogeneous text corpora

Bettina Berendt & Daniel Trümper (KU Leuven / HU Berlin)
Blaž Fortuna, Marko Grobelnik & Dunja Mladenić (JSI Ljubljana)
www.cs.kuleuven.be/~berendt
ICT Motivation:
Global+local interaction; beyond “similar documents“

with respect to what?
Application motivation: Beyond dedicated search engines

1. News and blogs

2. Multilingual sources
   - Good results in semi-automatic ontology learning based on simple machine translation

(Lloyd et al., *Proc. CAAW* 2006; Berendt et al., *Kommunikation, Partizipation und Wirkungen im Social Web*, 2008; Berendt, Fortuna et al., in prep.)
PASCAL motivation:
Re-use Textgarden’s bread & butter and advanced tools

- Text to bag-of-words
- Ontogen

http://www.textmining.net
http://ontogen.ijs.si/
Solution vision: 
**PORPOISE – Sailing the Internet**
Solution approach: Architecture & states overview

Search

Web

Specify sources & filters *

Retrieval & Preprocessing *

Global Analysis

Ont. Learning (Ontogen)

Import ontology *

Select Document *

Local analysis

Aspect-based similarity search*

Select neighbourhood *

Construct composite-similarity neighbourhood *

Refocus *

Source docs database*

Data / tool
External
Textgarden tool
User action
Created in this project *
Retrieval and preprocessing

- Crawler / wrapper * (uses Blogdigger)
- Translator * (uses Babelfish)
- Preprocessing (*Txt2Bow*)
- NER (GATE)
- Similarity Computation *

Retrieval & Preprocessing

Source doc.s database

Web
Ontology learning (1)
Ontology learning (2)
Ontology learning (3)
Inspection of ontology and instances
Inspection of documents

Action: The Capitulation Caucus
Tue Sep 11, 2007 at 07:25:23 AM PST

What is The Capitulation Caucus? That's what we want to know and we need your help. In the coming days, Congress will once again take up legislation on Iraq. As we see the first indication of fissures in Republican support for George Bush's endless war, now is the time for Democrats to stand firm. Yet one of the first bills the House may be voting on is the Abercrombie/Tanner bill, H.R. 3087. Toothless doesn't even begin to describe it.

(a) Strategy Required—Not later than 60 days after the date of the enactment of this Act, the President, in coordination with the Secretary of State, the Secretary of Defense, the Joint Chiefs of Staff, and other senior military leaders, shall develop and transmit to Congress

Close
More on documents
The neighbourhood of a document
Constructing the similarity measure & neighbourhood (I)
Constructing the similarity measure & neighbourhood (II)
Constructing the similarity measure & neighbourhood (III)

- Construct composite-similarity neighbourhood
- Specify sources & filters
  - Import ontology
  - Select document
  - Build ontology
  - Aspect-based similarity search
- Refocus

A news source

A German-language blog

Most neighbours are English-language blogs

English blog
German blog
English news
Comparing documents

to continue rubber stamping George Bush and so it's time for the community to dialing fingers and get to work. Where Congressman stand? Check that. Where Democratic member of Congress stand? vote for H.R. 3087? Will they continue majority of Democrats who want a time now? Below the fold is the name and phone number.

Democratic member of the House of Rep. goal here is to find out, "Yes or no, does Congressman ___ Tell Congress NOT to support this cash. The fold, the name and phone number of Congress Representative, and if you are unsure Representative is, information on how
Comparing documents; utilizing multilingual sources

to continue rubber stamping George Bush. And so it's time for the community to dialing fingers and get to work. Where Congressman stand? Check that. Where Democratic member of Congress stand? Will they continue majority of Democrats who want a time now? Below the fold is the name and phone number of the Democratic member of the House of Representatives. Tell Congress NOT to support this cap. The fold, the name and phone number of the Representative, and if you are unsure Representative is, information on how
Refocusing
Structuring a neighbourhood
Ex.: Finding a “story“

Evaluation? User studies!
“Pump-priming“: PORPOISE as catalyst

Using PASCAL software for analyzing social-media docs.

Using PASCAL software for analyzing multilingual social-media docs.

Supporting constructive search

Analyzing blogs and news

PORPOISE+

More fine-grained sailing

STORYGROWTH:

Tracking concept and community evolution

DM4E:

“More constructive search“
Finally ... could I express it better?

Mood: creative

_Ahem._

Another thing. If you do happen to read my book for whatever reason... Be so kind as to at least comment so I know. I like feedback. Thank you very much.