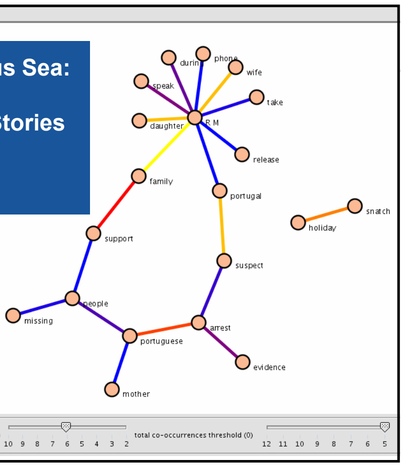
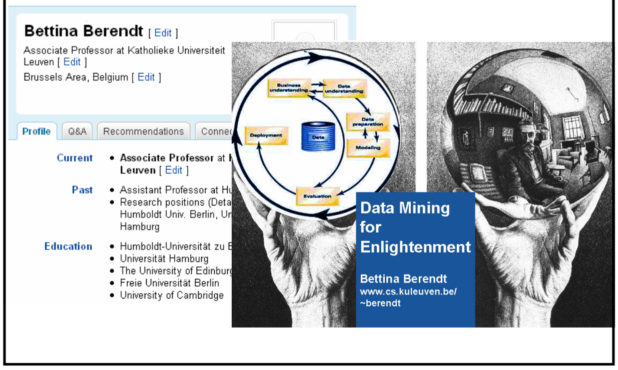
Sailing the Corpus Sea: Tools for Visual Discovery of Stories in Blogs and News



Bettina Berendt

www.cs.kuleuven.be/~berendt

About me



Bettina Berendt [Edit]

Associate Professor at Katholieke Universiteit Leuven [Edit]

Brussels Area, Belgium [Edit]

Current
- **Associate Professor** at K **Leuven** [Edit]

Past
- Assistant Professor at Hu
- Research positions (Deta Humboldt Univ. Berlin, Ur Hamburg

Education
- Humboldt-Universität zu E
- Universität Hamburg
- The University of Edinburg
- Freie Universität Berlin
- University of Cambridge

Data Mining for Enlightenment

Bettina Berendt
www.cs.kuleuven.be/~berendt

Thanks to ...

Daniel Trümper

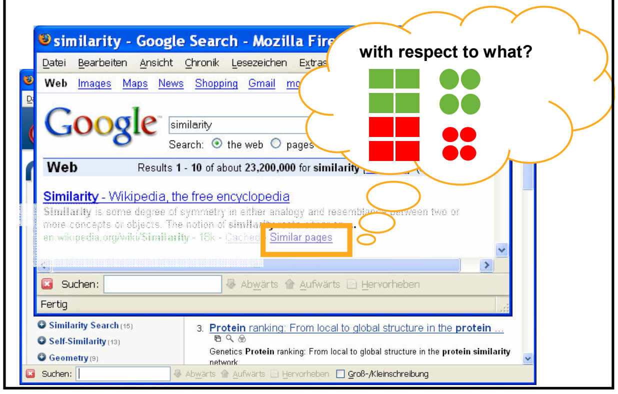
Tool at http://www.cs.kuleuven.be/~berendt/PORPOISE/

Ilija Subašić

Tool forthcoming;

all beta testers and experiment participants welcome!

First motivation: Global+local interaction; beyond "similar documents"



similarity - Google Search - Mozilla Fire
Datei Bearbeiten Ansicht Chronik Lesezeichen Extras
Web Images Maps News Shopping Gmail mo
Google
similarity
Search: ● the web ○ pages
Suchen:
Abwärts
Aufwärts
Hervorheben
Fertig
Similarity Search (15)
Self-Similarity (13)
Geometry (9)
□ Groß-/Kleinschreibung

Web Results **1** - **10** of about **23,200,000** for **similarity**

[Similarity - Wikipedia, the free encyclopedia]
Similarity is some degree of symmetry in either analogy and resemblance between two or more concepts or objects. The notion of **similarity** rests either on ...
en.wikipedia.org/wiki/Similarity - 16k - Cached - Similar pages

3. **Protein** ranking: From local to global structure in the **protein** ...
Genetics **Protein** ranking: From local to global structure in the **protein similarity** network

with respect to what?

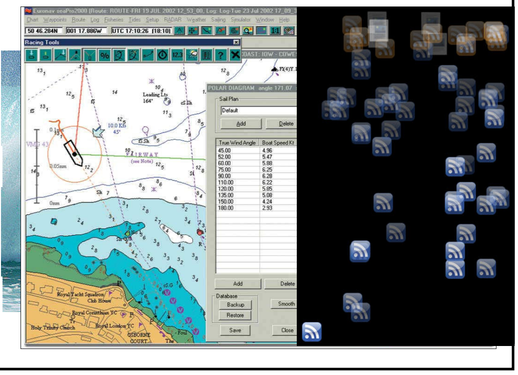
🟩🟩 ●●
🟩🟩 ●●
🟥🟥 ●●
🟥🟥 ●●

Solution vision: *Sailing the Internet*

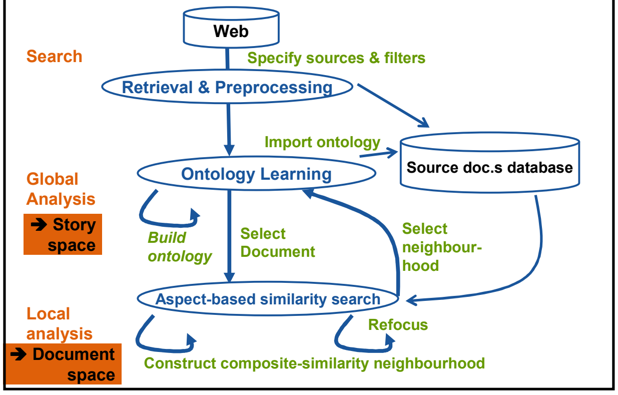
Search

Global Analysis

Local analysis



Solution approach: Architecture & states overview (version 1)



Search

Global Analysis

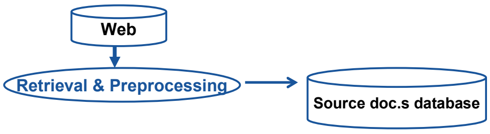
➔ Story space

Local analysis

➔ Document space

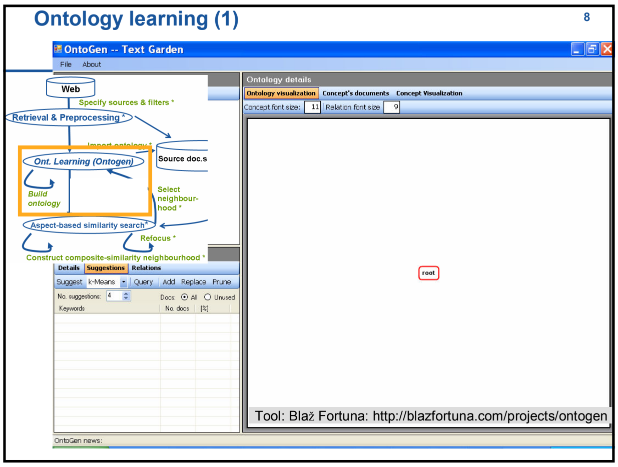
Web
Specify sources & filters
Retrieval & Preprocessing
Import ontology
Source doc.s database
Ontology Learning
Build ontology
Select Document
Select neighbour-hood
Aspect-based similarity search
Refocus
Construct composite-similarity neighbourhood

Retrieval and preprocessing



- Crawler / wrapper (uses Yahoo! / Google News; Blogdigger)
- Translator (uses Babelfish)
- Preprocessing (uses Textgarden, Terrier)
- Named-entity recognition (uses GATE, OpenCalais)
- Similarity Computation

Ontology learning (1)



No. suggestions: 4

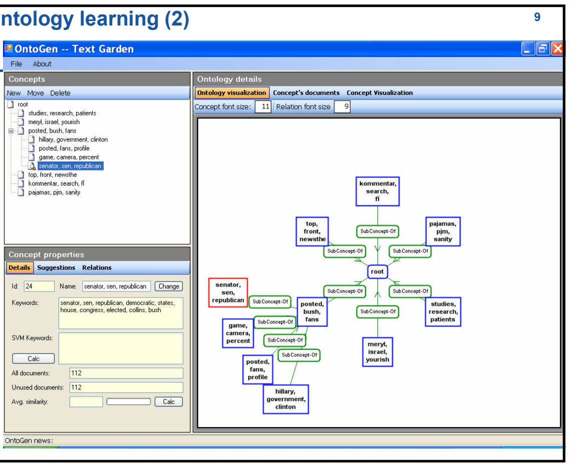
Docs: ● All ○ Unused

Keywords	No. docs	[%]

Concept font size: 11 Relation font size: 9

Tool: Blaž Fortuna: http://blazfortuna.com/projects/ontogen

Ontology learning (2)



Id: 24 Name: senator, sen, republican

Keywords: senator, sen, republican, democratic, states, house, congress, elected, collins, bush

SVM Keywords:

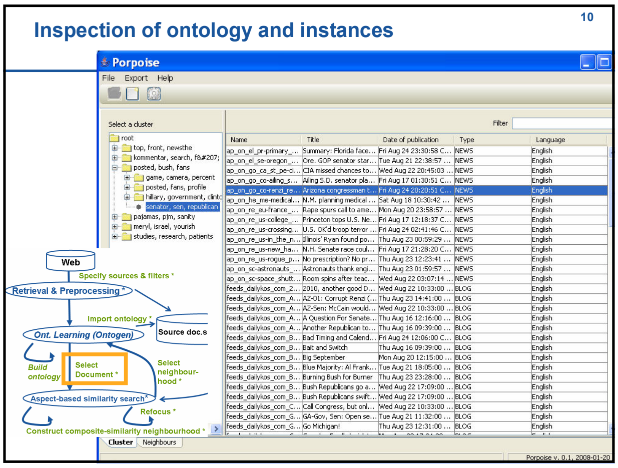
All documents: 112

Unused documents: 112

Avg. similarity:

Concept font size: 11 Relation font size: 9

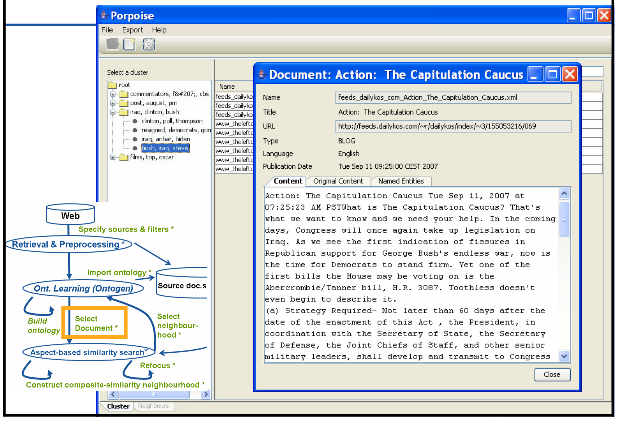
Inspection of ontology and instances



Name	Title	Date of publication	Type	Language
ap_on_el_pr-primary_…	Summary: Florida face…	Fri Aug 24 23:30:58 C…	NEWS	English
ap_on_el_se-oregon_…	Ore. GOP senator star…	Tue Aug 21 22:38:57 …	NEWS	English
ap_on_go_ca_st_pe-ci…	CIA missed chances to…	Wed Aug 22 20:45:03 …	NEWS	English
ap_on_go_co-ailing_s…	Ailing S.D. senator pla…	Fri Aug 17 01:30:51 C…	NEWS	English
ap_on_go_co-renzi_re…	Arizona congressman t…	Fri Aug 24 20:20:51 C…	NEWS	English
ap_on_he_me-medical…	N.M. planning medical …	Sat Aug 18 10:30:42 …	NEWS	English
ap_on_re_eu-france_…	Rape spurs call to ame…	Mon Aug 20 23:58:57 …	NEWS	English
ap_on_re_us-college_…	Princeton tops U.S. Ne…	Fri Aug 17 12:18:37 C…	NEWS	English
ap_on_re_us-crossing…	U.S. OK'd troop terror …	Fri Aug 24 02:41:46 C…	NEWS	English
ap_on_re_us-in_the_n…	Illinois' Ryan found po…	Thu Aug 23 00:59:29 …	NEWS	English
ap_on_re_us-new_ha…	N.H. Senate race coul…	Fri Aug 17 21:28:20 C…	NEWS	English
ap_on_re_us-rogue_p…	No prescription? No pr…	Thu Aug 23 12:23:41 …	NEWS	English
ap_on_sc-astronauts_…	Astronauts thank engi…	Thu Aug 23 01:59:57 …	NEWS	English
ap_on_sc-space_shutt…	Room spins after teac…	Wed Aug 22 03:07:14 …	NEWS	English
feeds_dailykos_com_2…	2010, another good D…	Wed Aug 22 10:33:00 …	BLOG	English
feeds_dailykos_com_A…	AZ-01: Corrupt Renzi (…	Thu Aug 23 14:41:00 …	BLOG	English
feeds_dailykos_com_A…	AZ-Sen: McCain would…	Wed Aug 22 10:33:00 …	BLOG	English
feeds_dailykos_com_A…	A Question For Senate…	Thu Aug 16 12:16:00 …	BLOG	English
feeds_dailykos_com_A…	Another Republican to…	Thu Aug 16 09:39:00 …	BLOG	English
feeds_dailykos_com_B…	Bad Timing and Calend…	Fri Aug 24 12:06:00 C…	BLOG	English
feeds_dailykos_com_B…	Bait and Switch	Thu Aug 16 09:39:00 …	BLOG	English
feeds_dailykos_com_B…	Big September	Mon Aug 20 12:15:00 …	BLOG	English
feeds_dailykos_com_B…	Blue Majority: Al Frank…	Tue Aug 21 18:05:00 …	BLOG	English
feeds_dailykos_com_B…	Burning Bush for Burner	Thu Aug 23 23:28:00 …	BLOG	English
feeds_dailykos_com_B…	Bush Republicans go a…	Wed Aug 22 17:09:00 …	BLOG	English
feeds_dailykos_com_B…	Bush Republicans swift…	Wed Aug 22 17:09:00 …	BLOG	English
feeds_dailykos_com_C…	Call Congress, but onl…	Wed Aug 22 10:33:00 …	BLOG	English
feeds_dailykos_com_G…	GA-Gov, Sen: Open se…	Tue Aug 21 11:32:00 …	BLOG	English
feeds_dailykos_com_G…	Go Michigan!	Thu Aug 23 12:31:00 …	BLOG	English

Porpoise v. 0.1, 2008-01-20

Inspection of documents



Name: feeds_dailykos_com_Action_The_Capitulation_Caucus.xml

Title: Action: The Capitulation Caucus

URL: http://feeds.dailykos.com/~r/dailykos/index/~3/155053216/069

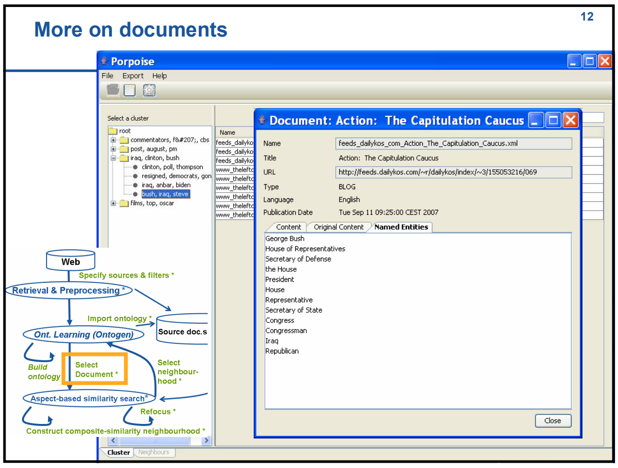
Type: BLOG

Language: English

Publication Date: Tue Sep 11 09:25:00 CEST 2007

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

More on documents



Name: feeds_dailykos_com_Action_The_Capitulation_Caucus.xml

Title: Action: The Capitulation Caucus

URL: http://feeds.dailykos.com/~r/dailykos/index/~3/155053216/069

Type: BLOG

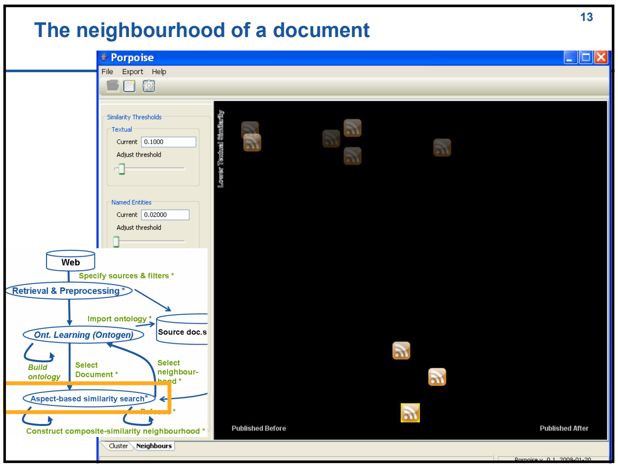
Language: English

Publication Date: Tue Sep 11 09:25:00 CEST 2007

George Bush
House of Representatives
Secretary of Defense
the House
President
House
Representative
Secretary of State
Congress
Congressman
Iraq
Republican

The neighbourhood of a document

13

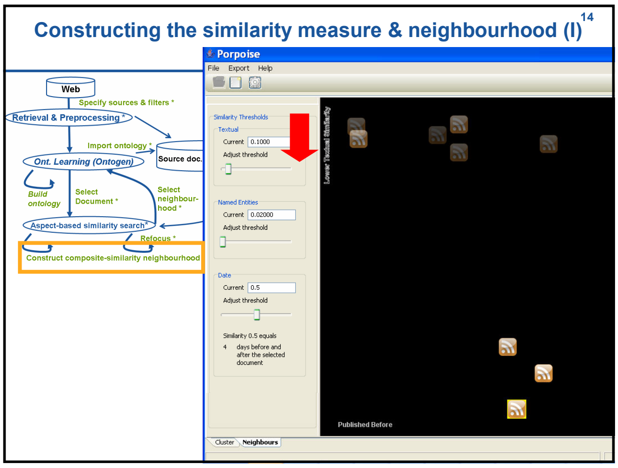


Textual: Current 0.1000, Adjust threshold

Named Entities: Current 0.02000, Adjust threshold

Constructing the similarity measure & neighbourhood (I)

14



Textual: Current 0.1000, Adjust threshold

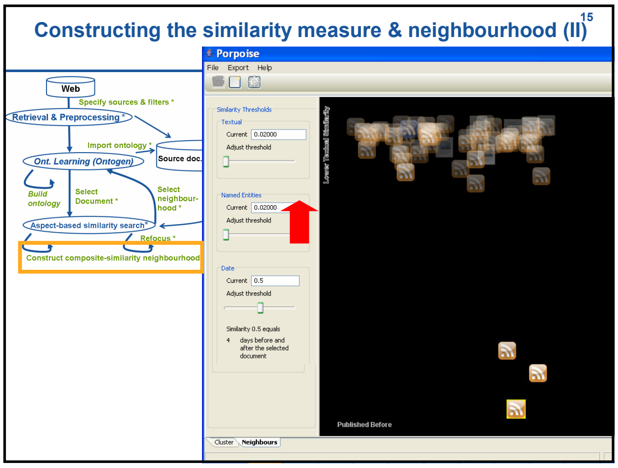
Named Entities: Current 0.02000, Adjust threshold

Date: Current 0.5, Adjust threshold

Similarity 0.5 equals 4 days before and after the selected document

Constructing the similarity measure & neighbourhood (II)

15



Textual: Current 0.02000, Adjust threshold

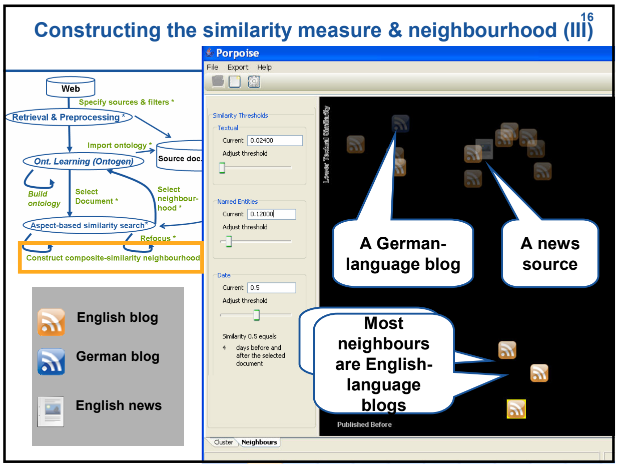
Named Entities: Current 0.02000, Adjust threshold

Date: Current 0.5, Adjust threshold

Similarity 0.5 equals 4 days before and after the selected document

Constructing the similarity measure & neighbourhood (III)

16



Textual: Current 0.02400, Adjust threshold

Named Entities: Current 0.12000, Adjust threshold

Date: Current 0.5, Adjust threshold

Similarity 0.5 equals 4 days before and after the selected document

English blog

German blog

English news

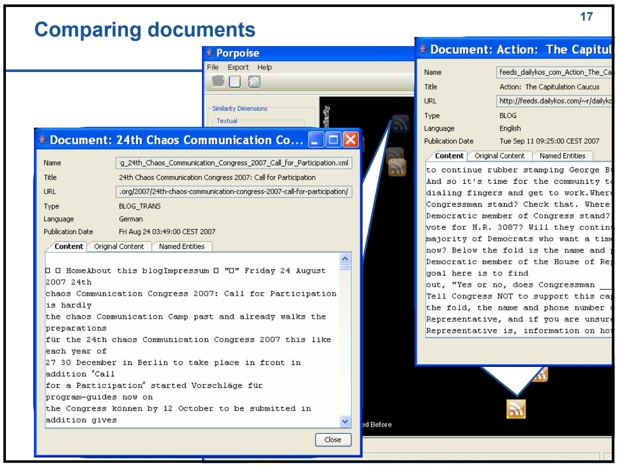
A German-language blog

A news source

Most neighbours are English-language blogs

Comparing documents

17



Document: 24th Chaos Communication Co...

Name: 'g_24th_Chaos_Communication_Congress_2007_Call_for_Participation.xml

Title: 24th Chaos Communication Congress 2007: Call for Participation

URL: .org/2007/24th-chaos-communication-congress-2007-call-for-participation/

Type: BLOG_TRANS

Language: German

Publication Date: Fri Aug 24 03:49:00 CEST 2007

```
□ □ HomeAbout this blogImpressum □ "□" Friday 24 August
2007 24th
chaos Communication Congress 2007: Call for Participation
is hardly
the chaos Communication Camp past and already walks the
preparations
für the 24th chaos Communication Congress 2007 this like
each year of
27 30 December in Berlin to take place in front in
addition "Call
for a Participation" started Vorschläge für
program-guides now on
the Congress können by 12 October to be submitted in
addition gives
```

Document: Action: The Capitul

Name: feeds_dailykos_com_Action_The_Ca

Title: Action: The Capitulation Caucus

URL: http://feeds.dailykos.com/~r/dailyko

Type: BLOG

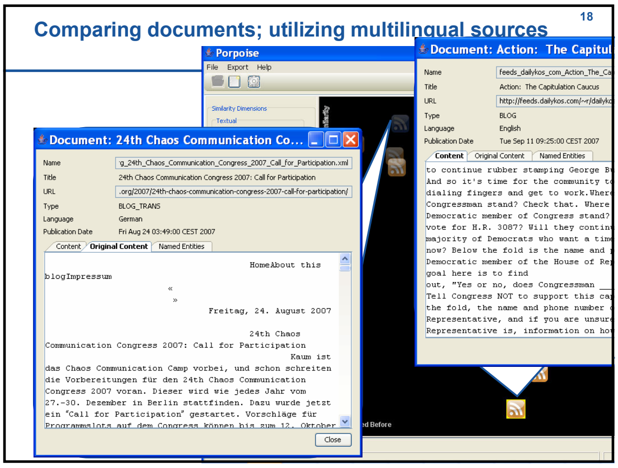
Language: English

Publication Date: Tue Sep 11 09:25:00 CEST 2007

```
to continue rubber stamping George Bu
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 contin
majority of Democrats who want a time
now? Below the fold is the name and p
Democratic member of the House of Rep
goal here is to find
out, "Yes or no, does Congressman ___
Tell Congress NOT to support this cap
the fold, the name and phone number o
Representative, and if you are unsure
Representative is, information on how
```

Comparing documents; utilizing multilingual sources

18



Document: 24th Chaos Communication Co...

Name: 'g_24th_Chaos_Communication_Congress_2007_Call_for_Participation.xml

Title: 24th Chaos Communication Congress 2007: Call for Participation

URL: .org/2007/24th-chaos-communication-congress-2007-call-for-participation/

Type: BLOG_TRANS

Language: German

Publication Date: Fri Aug 24 03:49:00 CEST 2007

```
                                        HomeAbout this
blogImpressum
                        «
                         »
                                Freitag, 24. August 2007

                                        24th Chaos
Communication Congress 2007: Call for Participation
                                                Kaum ist
das Chaos Communication Camp vorbei, und schon schreiten
die Vorbereitungen für den 24th Chaos Communication
Congress 2007 voran. Dieser wird wie jedes Jahr vom
27.-30. Dezember in Berlin stattfinden. Dazu wurde jetzt
ein "Call for Participation" gestartet. Vorschläge für
Programmslots auf dem Congress können bis zum 12. Oktober
```

Document: Action: The Capitul

Name: feeds_dailykos_com_Action_The_Ca

Title: Action: The Capitulation Caucus

URL: http://feeds.dailykos.com/~r/dailyko

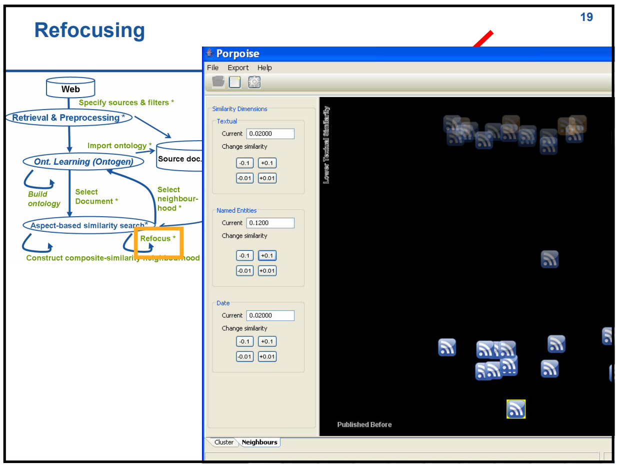
Type: BLOG

Language: English

Publication Date: Tue Sep 11 09:25:00 CEST 2007

```
to continue rubber stamping George Bu
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 contin
majority of Democrats who want a time
now? Below the fold is the name and p
Democratic member of the House of Rep
goal here is to find
out, "Yes or no, does Congressman ___
Tell Congress NOT to support this cap
the fold, the name and phone number o
Representative, and if you are unsure
Representative is, information on how
```

Refocusing



Web

Specify sources & filters *

Retrieval & Preprocessing *

Import ontology *

Source doc.

Ont. Learning (Ontogen)

Build ontology

Select Document *

Select neighbour-hood *

Aspect-based similarity search*

Refocus *

Construct composite-similarity neighbourhood

Porpoise

File Export Help

Similarity Dimensions

Textual

Current 0.02000

Change similarity

-0.1 +0.1 -0.01 +0.01

Named Entities

Current 0.1200

Change similarity

-0.1 +0.1 -0.01 +0.01

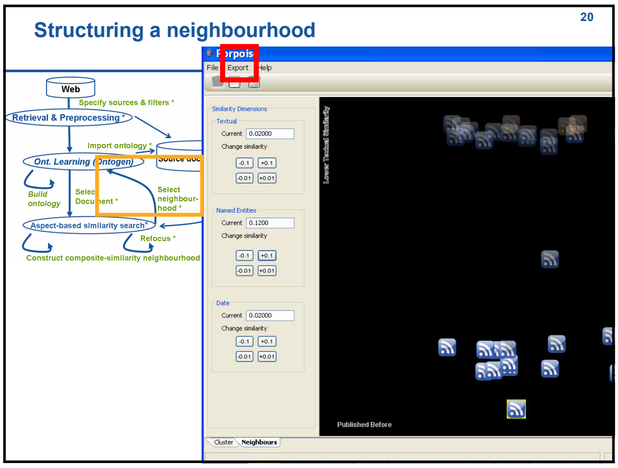
Date

Current 0.02000

Change similarity

-0.1 +0.1 -0.01 +0.01

Structuring a neighbourhood



Web

Specify sources & filters *

Retrieval & Preprocessing *

Import ontology *

Source doc.

Ont. Learning (Ontogen)

Build ontology

Select Document *

Select neighbour-hood *

Aspect-based similarity search*

Refocus *

Construct composite-similarity neighbourhood

Porpoise

File Export Help

Similarity Dimensions

Textual

Current 0.02000

Change similarity

-0.1 +0.1 -0.01 +0.01

Named Entities

Current 0.1200

Change similarity

-0.1 +0.1 -0.01 +0.01

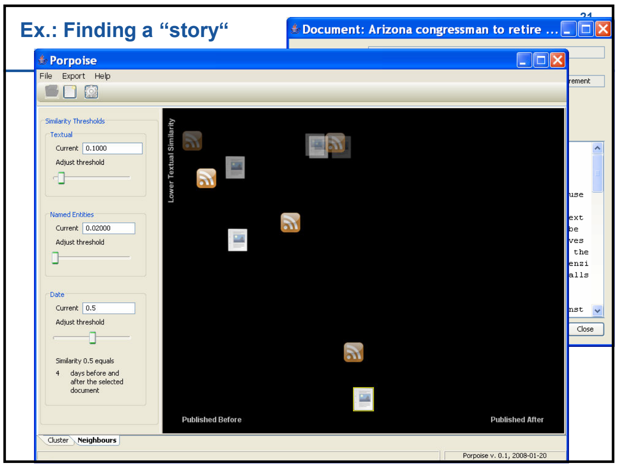
Date

Current 0.02000

Change similarity

-0.1 +0.1 -0.01 +0.01

Ex.: Finding a “story“



Document: Arizona congressman to retire ...

Porpoise

File Export Help

Similarity Thresholds

Textual

Current 0.1000

Adjust threshold

Named Entities

Current 0.02000

Adjust threshold

Date

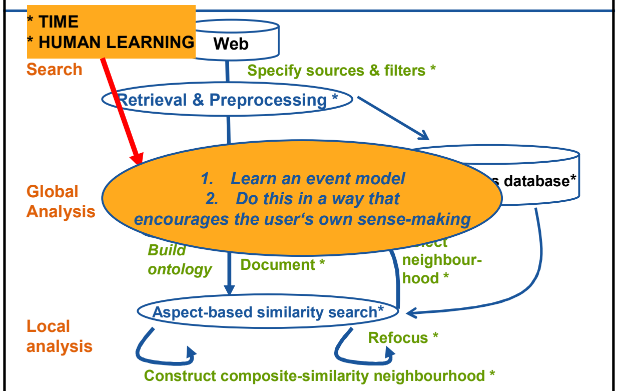
Current 0.5

Adjust threshold

Similarity 0.5 equals 4 days before and after the selected document

Porpoise v. 0.1, 2008-01-20

Going further: Towards STORIES



* TIME

* HUMAN LEARNING

Web

Search

Specify sources & filters *

Retrieval & Preprocessing *

s database*

Global Analysis

1. *Learn an event model*
2. *Do this in a way that encourages the user's own sense-making*

Build ontology

Document *

ect neighbour-hood *

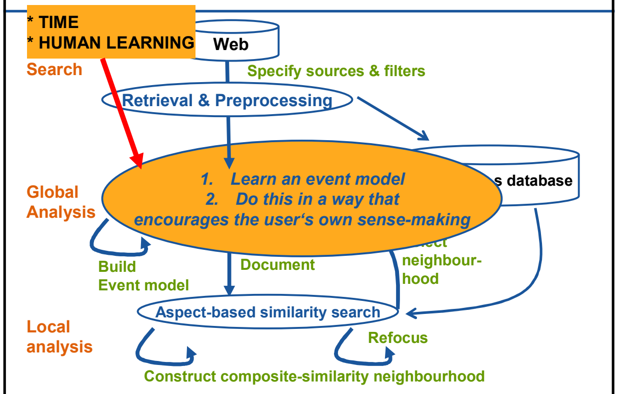
Aspect-based similarity search*

Local analysis

Refocus *

Construct composite-similarity neighbourhood *

Going further: Architecture and states (version 2)



* TIME

* HUMAN LEARNING

Web

Search

Specify sources & filters

Retrieval & Preprocessing

s database

Global Analysis

1. *Learn an event model*
2. *Do this in a way that encourages the user's own sense-making*

Build Event model

Document

ect neighbour-hood

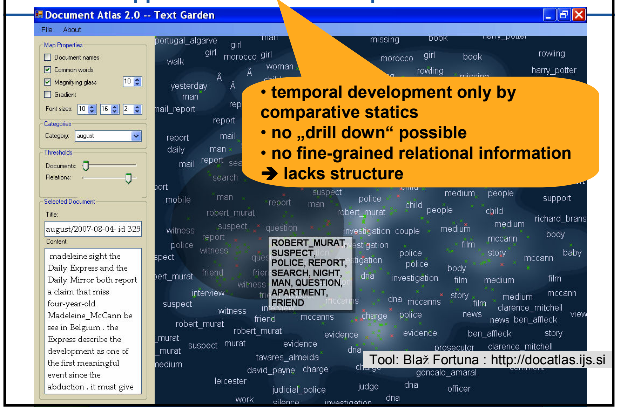
Aspect-based similarity search

Local analysis

Refocus

Construct composite-similarity neighbourhood

Solution approach 1: Find latent topics



Document Atlas 2.0 -- Text Garden

File About

Map Properties

☐ Document names

☑ Common words

☑ Magnifying glass 10

☐ Gradient

Font sizes: 10 16 2

Categories

Category: august

Thresholds

Documents:

Relations:

Selected Document

Title:

august/2007-08-04- id 329

Content:

madeleine sight the Daily Express and the Daily Mirror both report a claim that miss four-year-old Madeleine_McCann be see in Belgium . the Express describe the development as one of the first meaningful event since the abduction . it must give

ROBERT_MURAT, SUSPECT, POLICE, REPORT, SEARCH, NIGHT, MAN, QUESTION, APARTMENT, FRIEND

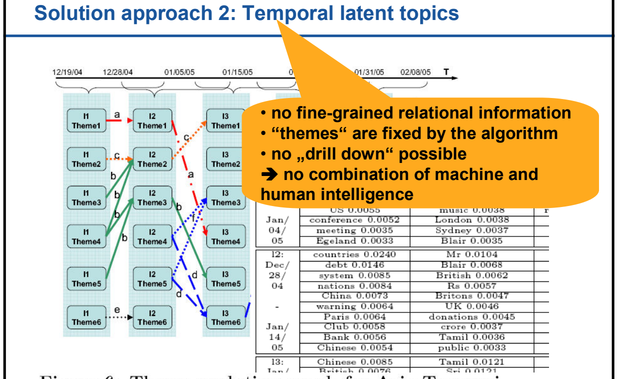
- temporal development only by comparative statics
- no „drill down“ possible
- no fine-grained relational information

➔ lacks structure

Tool: Blaž Fortuna : http://docatlas.ijs.si

25

Solution approach 2: Temporal latent topics



	US 0.0055	music 0.0038
Jan/ 04/ 05	conference 0.0052	London 0.0038
	meeting 0.0035	Sydney 0.0037
	Egeland 0.0033	Blair 0.0035
I2: Dec/ 28/ 04	countries 0.0240	Mr 0.0104
	debt 0.0146	Blair 0.0068
	system 0.0085	British 0.0062
	nations 0.0084	Rs 0.0057
	China 0.0073	Britons 0.0047
-	warning 0.0064	UK 0.0046
	Paris 0.0064	donations 0.0045
Jan/ 14/ 05	Club 0.0058	crore 0.0037
	Bank 0.0056	Tamil 0.0036
	Chinese 0.0054	public 0.0033
I3: Jan/	Chinese 0.0085	Tamil 0.0121
	British 0.0076	Sri 0.0121

- no fine-grained relational information
- "themes" are fixed by the algorithm
- no „drill down" possible

➔ no combination of machine and human intelligence

Figure 6: Theme evolution graph for Asia Tsunami

Mei & Zhai, *PKDD 2005*

26

The ETP3 problem

Evolutionary theme patterns discovery, summary & exploration

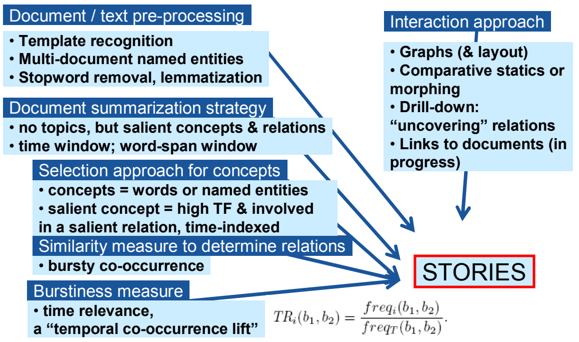
1. identify topical sub-structure in a set (generally, a time-indexed stream) of documents constrained by being about a common topic
2. show how these substructures emerge, change, and disappear (and maybe re-appear) over time
3. give users intuitive and interactive interfaces for exploring the topic landscape and the underlying documents

and for their own sense-making

– use machine-generated summarization only as a starting point!

27

Ingredients of a solution to ETP3 = Evolutionary theme patterns discovery, summary and exploration



Document / text pre-processing

- Template recognition
- Multi-document named entities
- Stopword removal, lemmatization

Document summarization strategy

- no topics, but salient concepts & relations
- time window; word-span window

Selection approach for concepts

- concepts = words or named entities
- salient concept = high TF & involved in a salient relation, time-indexed

Similarity measure to determine relations

- bursty co-occurrence

Burstiness measure

- time relevance, a "temporal co-occurrence lift"

$TR_i(b_1, b_2) = \frac{freq_i(b_1, b_2)}{freq_T(b_1, b_2)}.$

Interaction approach

- Graphs (& layout)
- Comparative statics or morphing
- Drill-down: "uncovering" relations
- Links to documents (in progress)

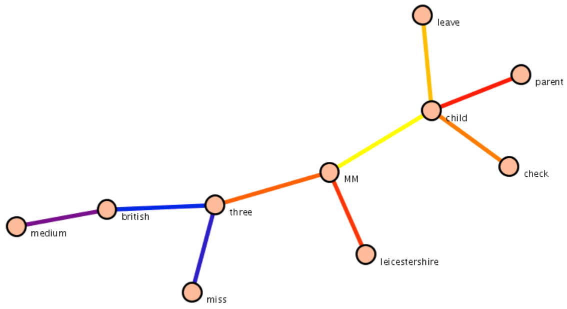
STORIES

28

"Powerpoint demo"

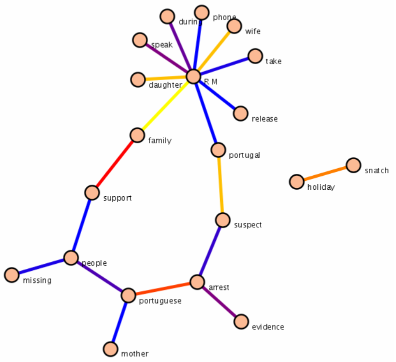
29

An event: a missing child



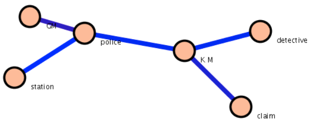
30

A central figure emerges in the police investigations



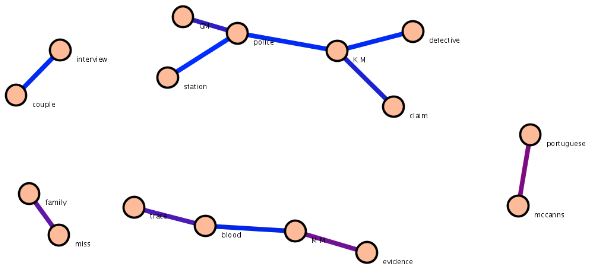
31

Uncovering more details



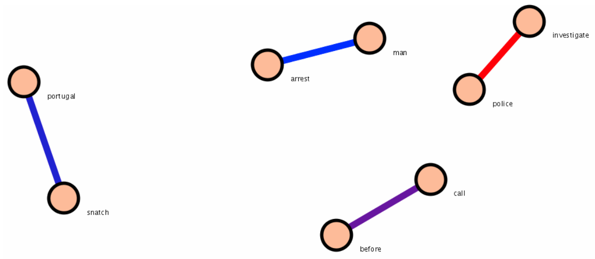
32

Uncovering more details



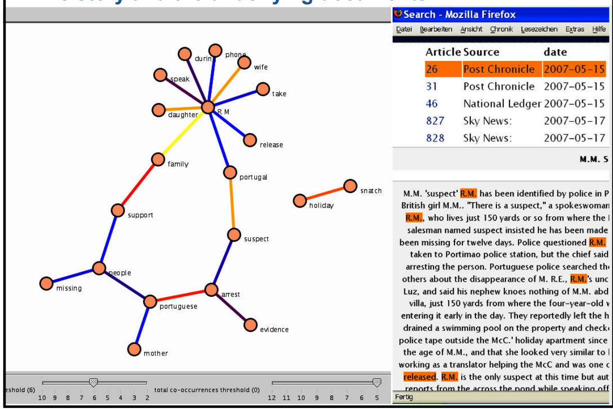
33

An eventless time



34

The story and the underlying documents



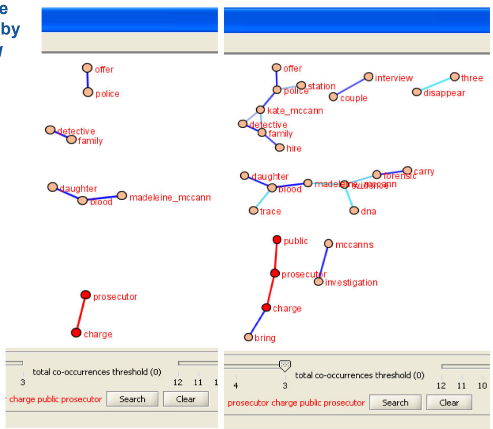
Article	Source	date
26	Post Chronicle	2007–05–15
31	Post Chronicle	2007–05–15
46	National Ledger	2007–05–15
827	Sky News:	2007–05–17
828	Sky News:	2007–05–17

M.M. S

M.M. 'suspect' R.M. has been identified by police in P
British girl M.M.. "There is a suspect," a spokeswoman
R.M., who lives just 150 yards or so from where the
salesman named suspect insisted he has been made
been missing for twelve days. Police questioned R.M.
taken to Portimao police station, but the chief said
arresting the person. Portuguese police searched th
others about the disappearance of M. R.E., R.M.'s unc
Luz, and said his nephew knoes nothing of M.M. abd
villa, just 150 yards from where the four–year–old
entering it early in the day. They reportedly left the h
drained a swimming pool on the property and check
police tape outside the McC.' holiday apartment since
the age of M.M., and that she looked very similar to
working as a translator helping the McC and was one
released. R.M. is the only suspect at this time but aut
reports from the across the pond while speaking off

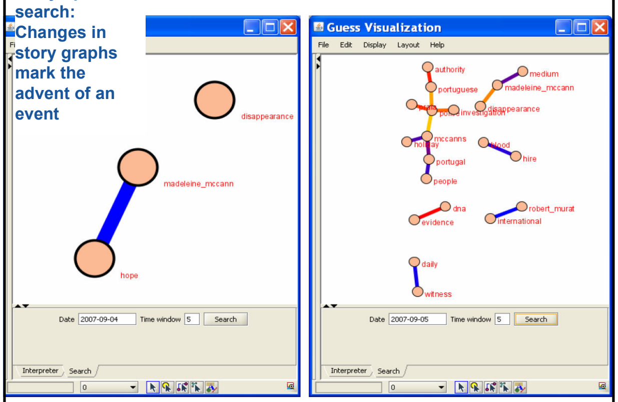
35

Story-space navigation by *uncovering* and advanced document search



36

Story-space search: Changes in story graphs mark the advent of an event



Data collection and preprocessing

- Articles from Google News 05/2007 – 11/2007 for search term “madeleine mccann“
  - *(there was a Google problem in the December archive)*
- Only English-language articles
- For each month, the first 100 hits
- Of these, all that were freely available ➔ 477 documents

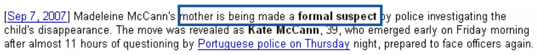
- Preprocessing:
  - HTML cleaning
  - tokenization
  - stopword removal

Story elements

- *content-bearing words*
  - the 150 top-TF words without stopwords

$$TF_{ij} = \frac{n_{ij}}{|d_i|}$$

Story stages: co-occurrence in a window



[Sep 7, 2007] Madeleine McCann's mother is being made a **formal suspect** by police investigating the child's disappearance. The move was revealed as **Kate McCann**, 39, who emerged early on Friday morning after almost 11 hours of questioning by Portuguese police on Thursday night, prepared to face officers again.

“mother“ and “suspect“ co-occur
- in a window of size ≥ 6 (all words)
- in a window of size ≥ 2 (non-stopwords only)

Salient story elements

1. Split whole corpus T by week (17 = 30 Apr + until 44 = 12 Nov +)
2. For each week
   - Compute the weights for corpus t for this week
3. Weight =
   - *Support* of co-occurrence of 2 content-bearing words w_1, w_2 in t =

   (# articles from t containing both w_1, w_2 in window) / (# all articles in t)
4. Threshold
   - Number of occurrences of *co-occurrence(w_1, w_2)* in t ≥ θ_1 (e.g., 5)
   - *Time-relevance* TR of *co-occurrence(w_1, w_2)* =

     support(*co-occurrence(w_1, w_2)*) in t / support(*co-occurrence(w_1, w_2)*) in T ≥ θ_2 (e.g., 2) *
5. Rank by TR, for each week identify top 2
6. Story elements = peak words = all elements of these top 2 pairs (# = 38)

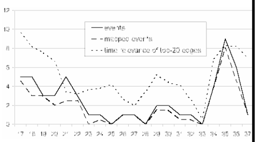
Salient story stages, and story evolution

7. Story stage = co-occurrences of peak words in t
   - For each week t: aggregate over t-2, t-1, t ➔ moving average

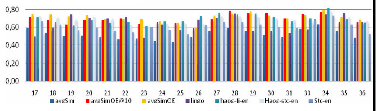
8. Story evolution = how story stages evolve over the t in T

Evaluations *(so far ...)*

1. Information retrieval quality
   - Challenge: What is the ground truth
   - ➔ Build on Wikipedia
   - Edges – events: up to 80% recall, ca. 30% precision

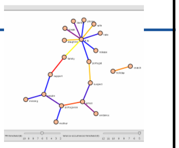


2. Search quality
   - Story subgraphs index coherent document clusters

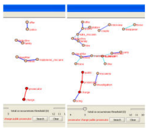


3. Learning effectiveness
   - Document search with story graphs leads to averages of
   - 75% accuracy on judgments of story fact truth
   - 3.4 nodes/words per query

Summary



Navigation in story space → story building



+

Document search

+

Navigation in document space

lead to understandable, useful + intuitive interfaces

The Future

- Better language processing
- Linkage information!
- Opinion mining

Thanks!

References

Subašić, I. & Berendt, B. (2008). Web Mining for Understanding Stories through Graph Visualisation. In *Proceedings of ICDM 2008*. IEEE Press.

Berendt, B. and D. Trümper (in press). Semantics-based analysis and navigation of heterogeneous text corpora: the Porpoise news and blogs engine. In I.-H. Ting & H.-J. Wu (Eds.), *Web Mining Applications in E-commerce and E-services*, Berlin etc.: Springer.

Berendt, B. & Subašić, I. (in press). Measuring graph topology for interactive temporal event detection. To appear in *Künstliche Intelligenz*.

Berendt, B. & Subašić, I. (under review). Discovery of interactive graphs for understanding and searching time-indexed corpora.

Please see http://www.cs.kuleuven.be/~berendt/ for these papers.