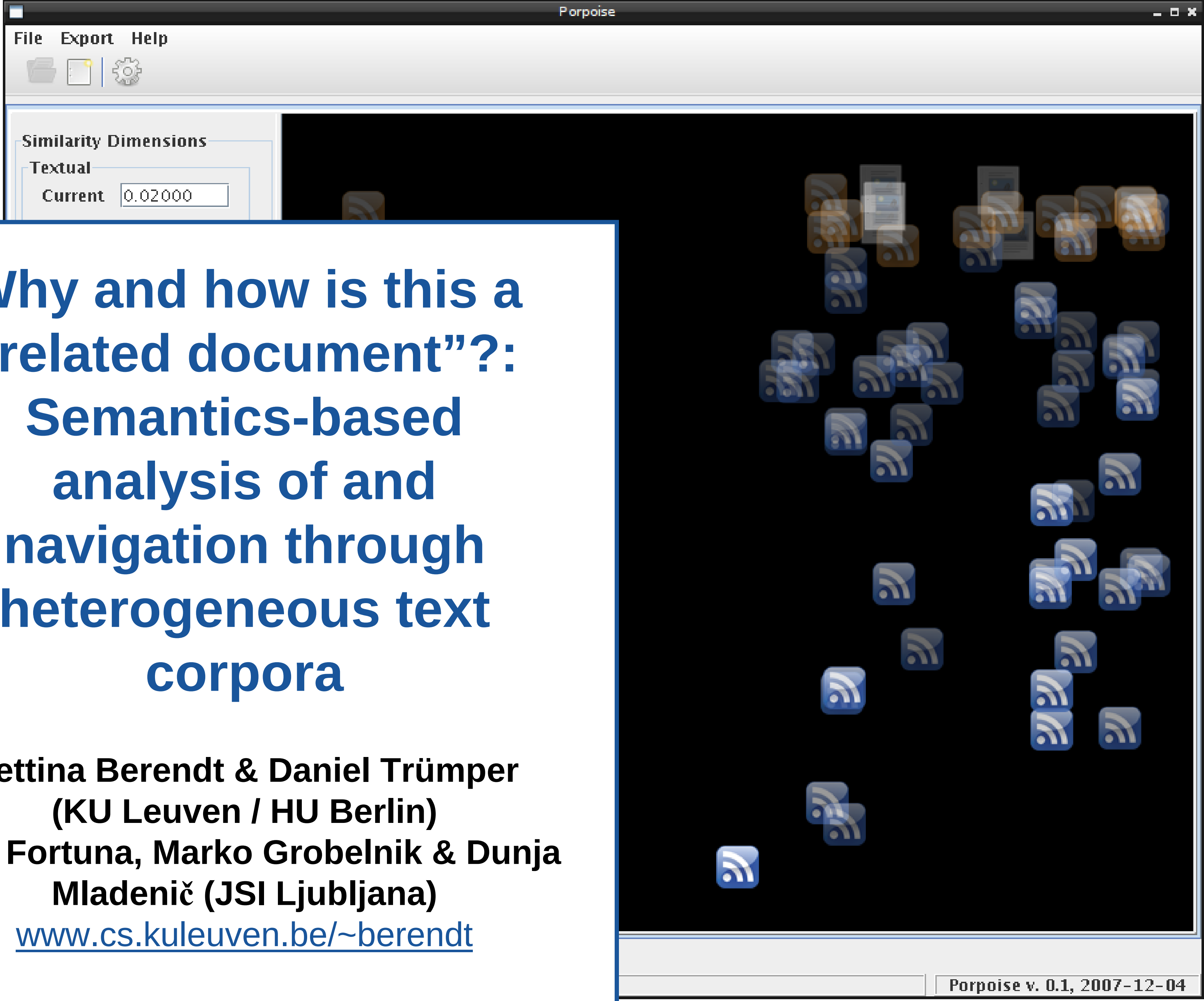
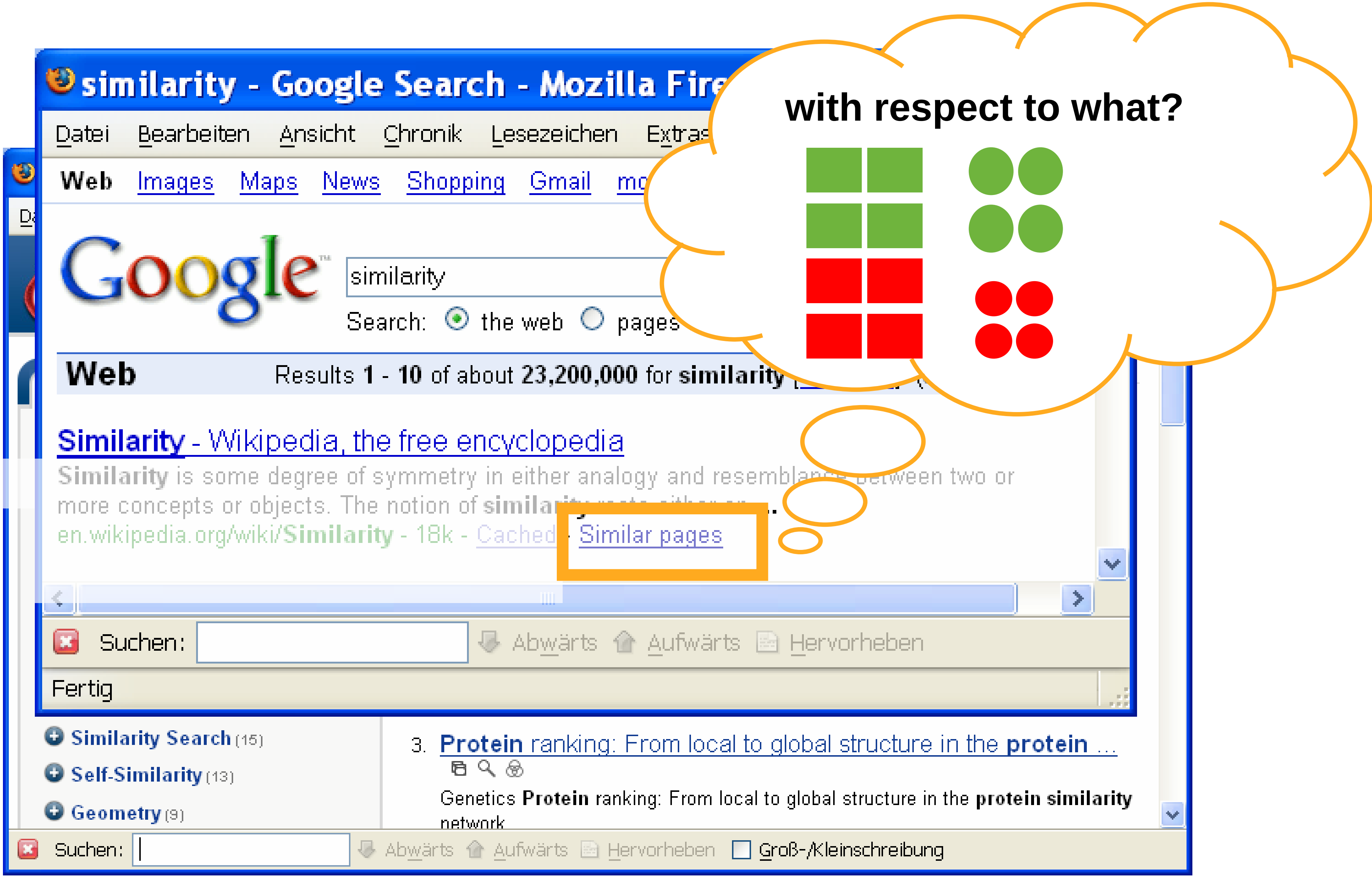
Why and how is this a "related document"?: Semantics-based analysis of and navigation through heterogeneous text corpora

Bettina Berendt & Daniel Trümper
(KU Leuven / HU Berlin)
Blaž Fortuna, Marko Grobelnik & Dunja Mladenič (JSI Ljubljana)
www.cs.kuleuven.be/~berendt



Current 0.02000

ICT Motivation:
Global+local interaction; beyond “similar documents“



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** - 18k - Cached - Similar pages

Similarity Search (15)
Self-Similarity (13)
Geometry (9)

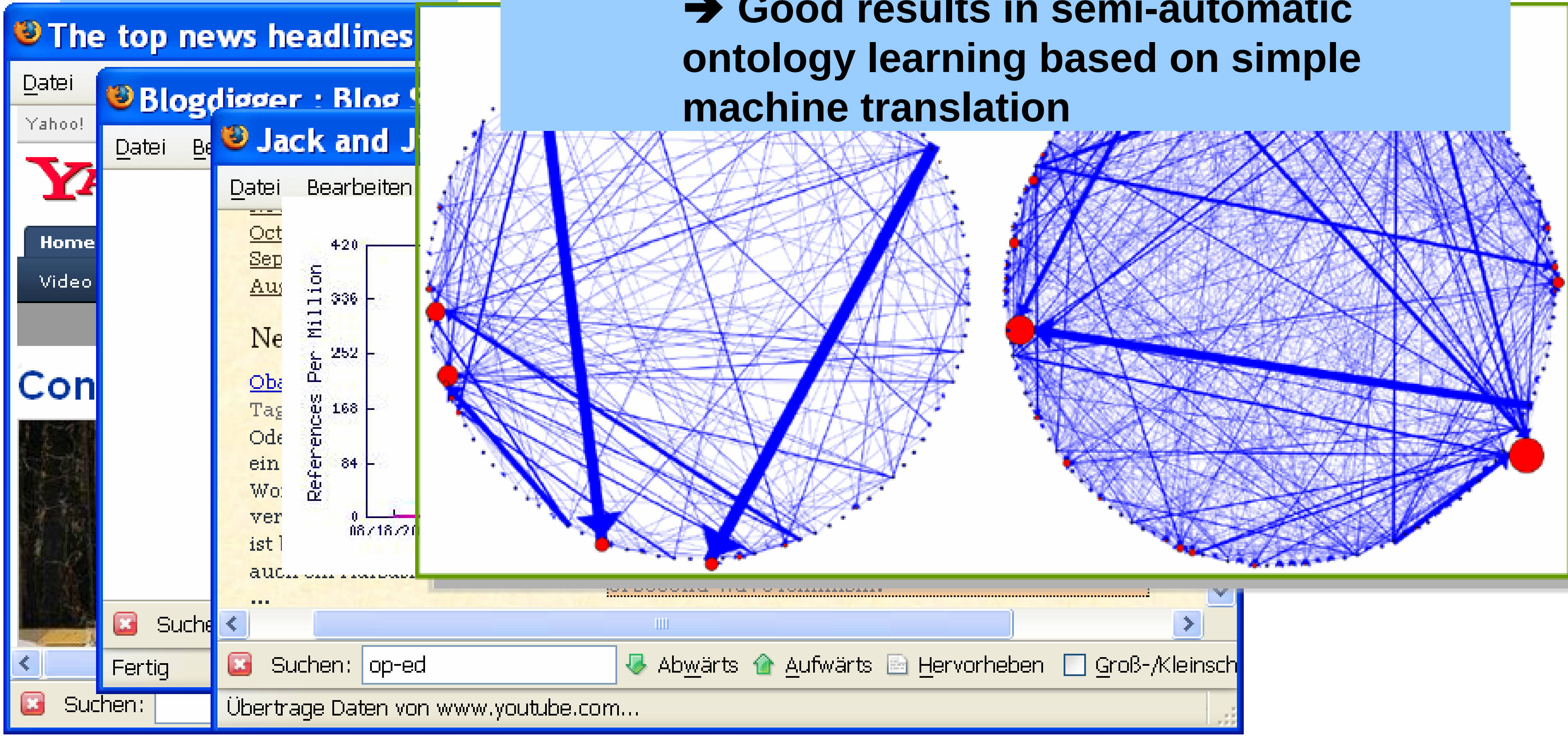
3. **Protein** ranking: From local to global structure in the **protein** ...
Genetics **Protein** ranking: From local to global structure in the **protein similarity** network

Application motivation: Beyond dedicated search engines

1. News and blogs

2. Multilingual sources

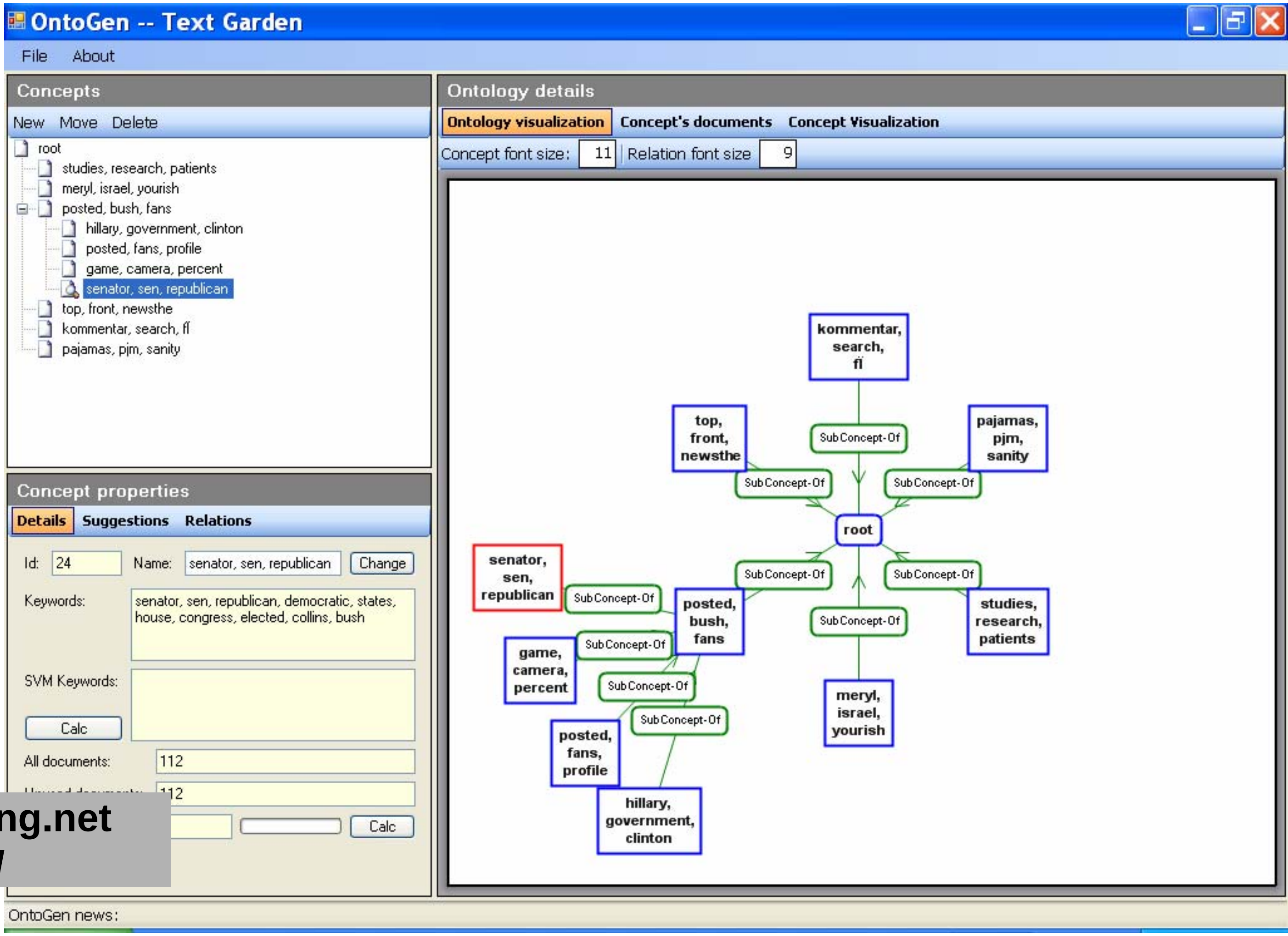
➔ Good results in semi-automatic ontology learning based on simple machine translation



(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



- root
  - studies, research, patients
  - meryl, israel, yourish
  - posted, bush, fans
    - hillary, government, clinton
    - posted, fans, profile
    - game, camera, percent
    - senator, sen, republican
  - top, front, newsthe
  - kommentar, search, fl
  - pajamas, pjm, sanity

Concept font size: 11 Relation font size 9

Id: 24 Name: senator, sen, republican

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

SVM Keywords:

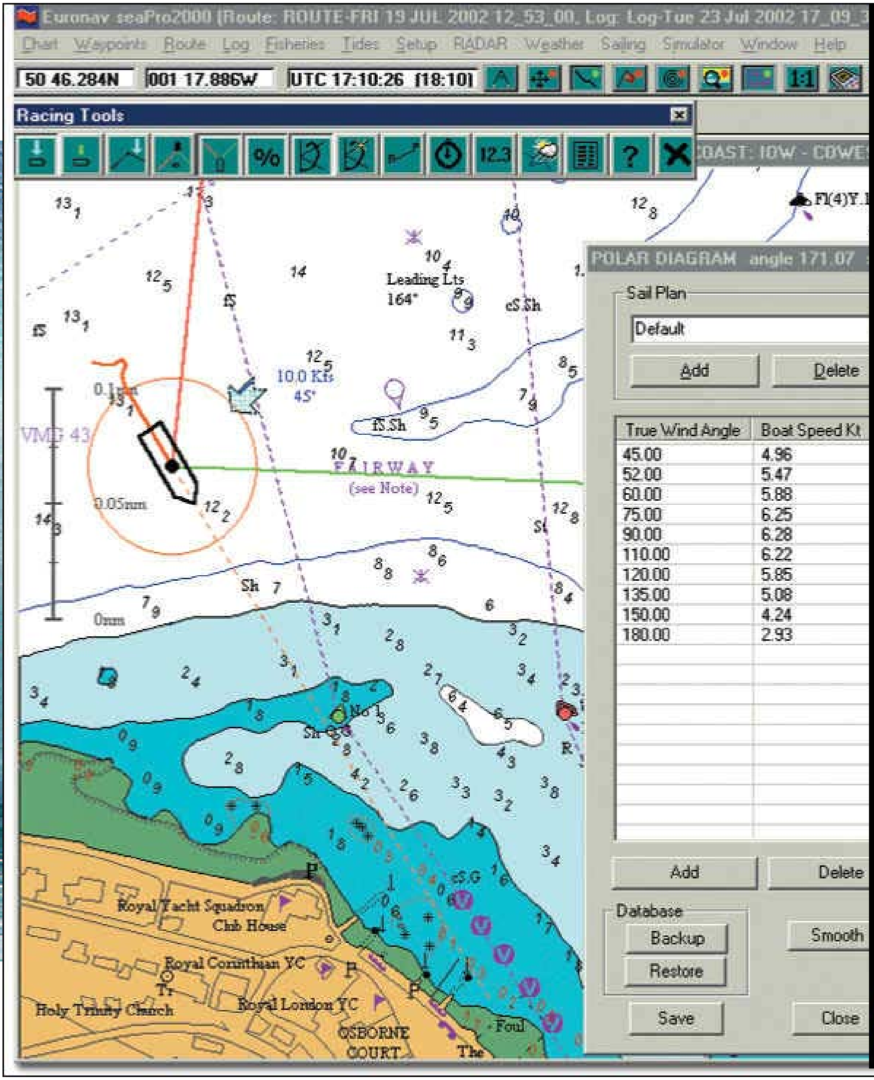
All documents: 112

112

http://www.textmining.net
http://ontogen.ijs.si/

Solution vision:
PORPOISE – Sailing the Internet

Search
Global Analysis
Local analysis

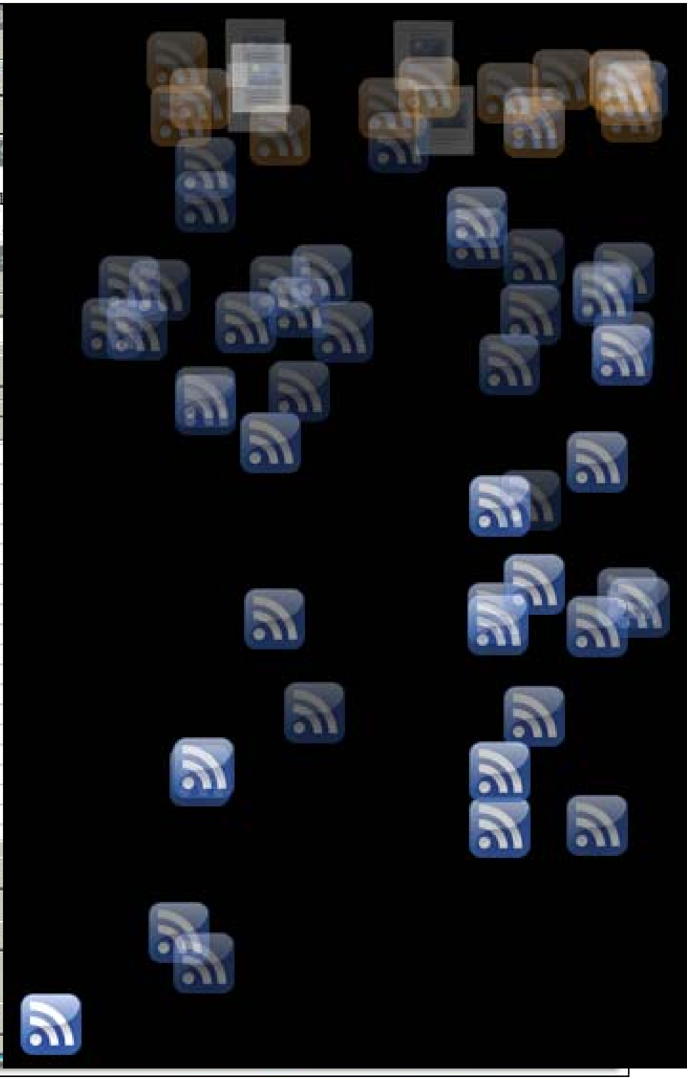

Euronav seaPro2000 (Route: ROUTE-FRI 19 JUL 2002 12_53_00, Log: Log-Tue 23 Jul 2002 17_09_3

50 46.284N | 001 17.886W | UTC 17:10:26 (18:10)

POLAR DIAGRAM angle 171.07

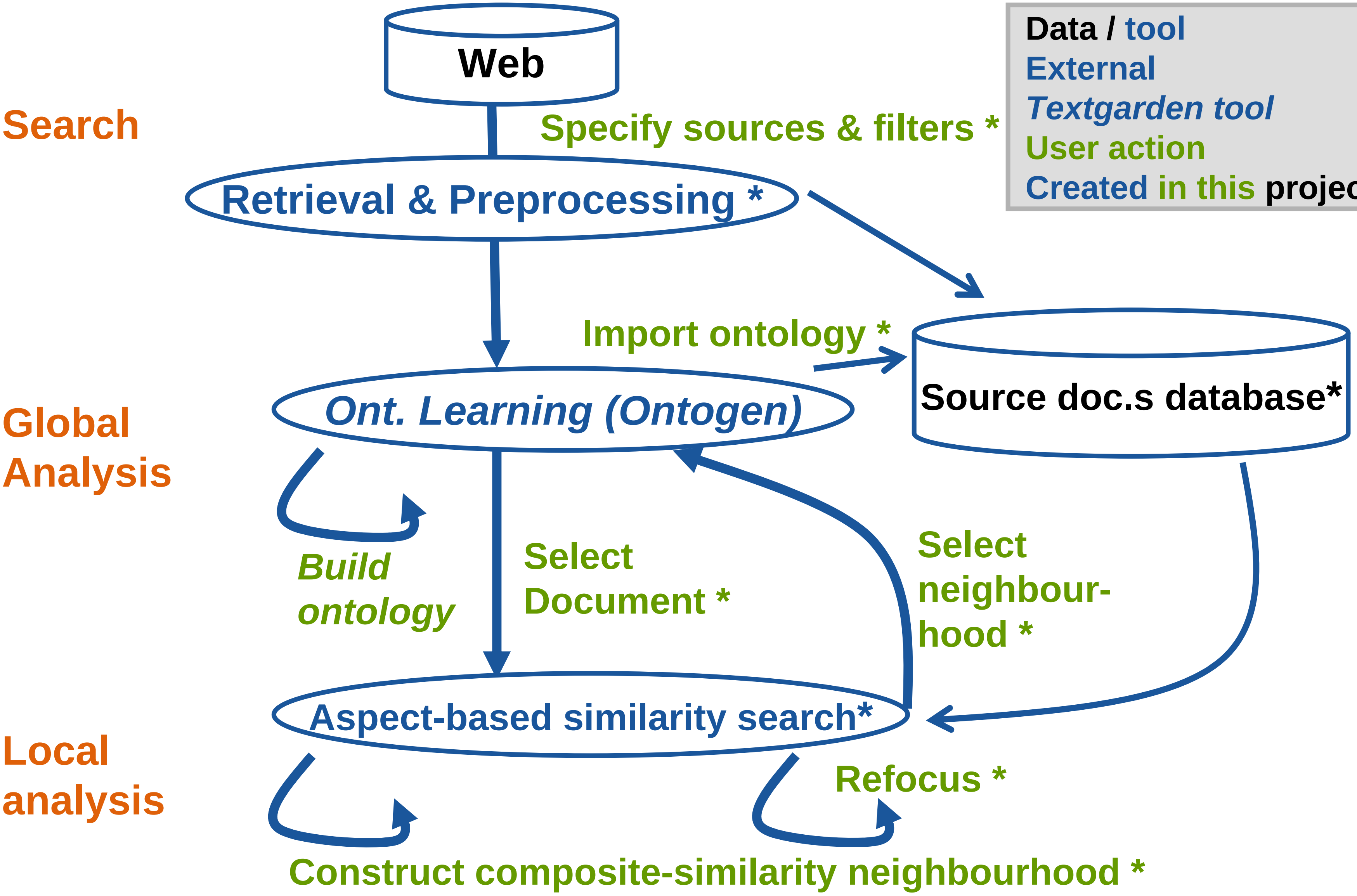
Sail Plan: Default

True Wind Angle	Boat Speed Kt
45.00	4.96
52.00	5.47
60.00	5.88
75.00	6.25
90.00	6.28
110.00	6.22
120.00	5.85
135.00	5.08
150.00	4.24
180.00	2.93

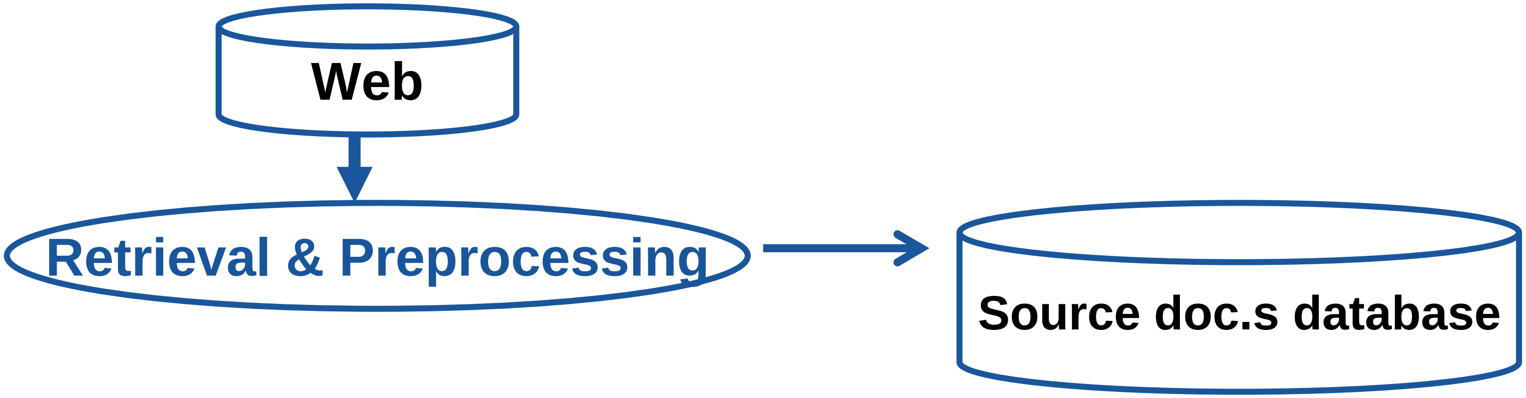


Solution approach: Architecture & states overview

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

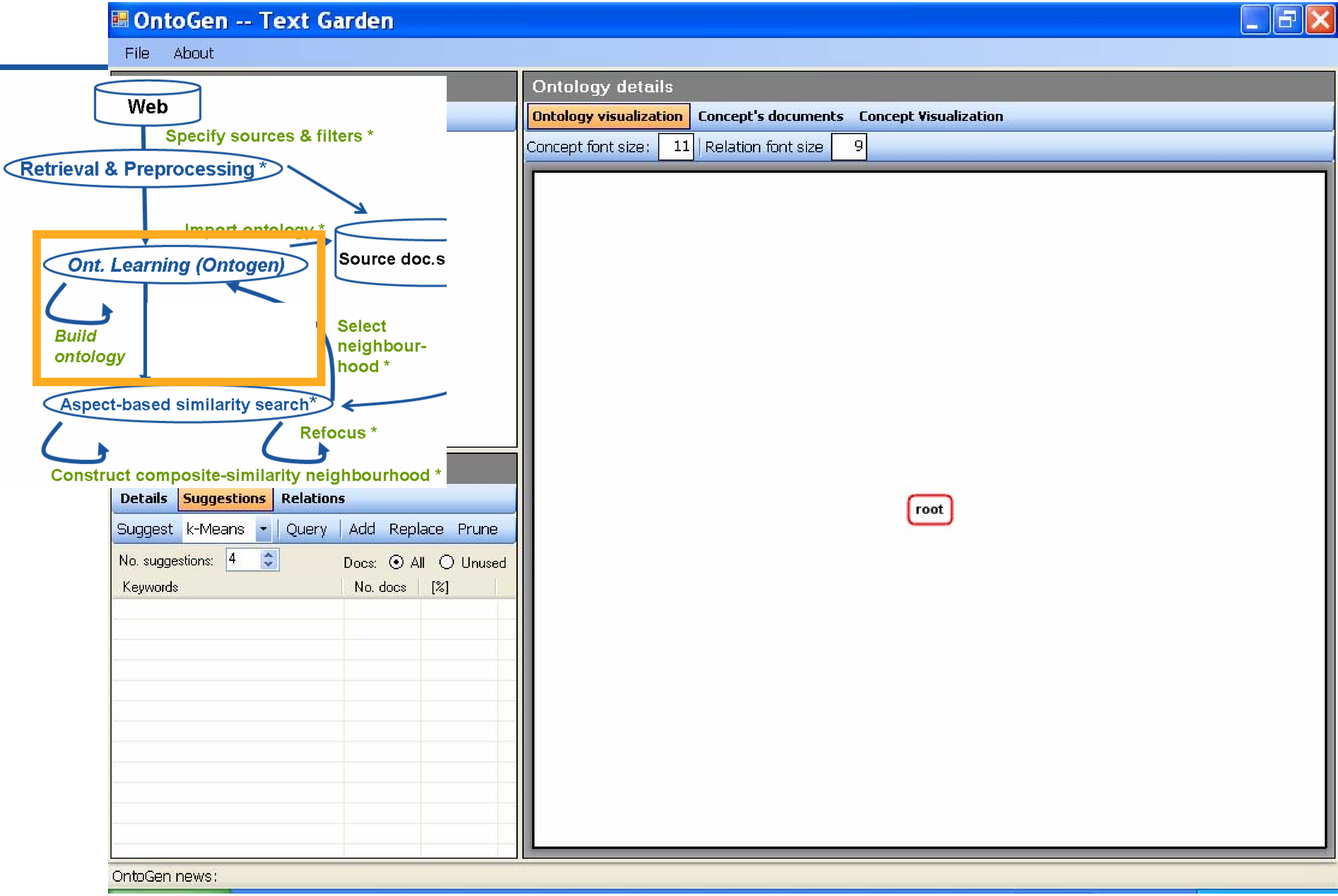


Retrieval and preprocessing



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

Ontology learning (1)

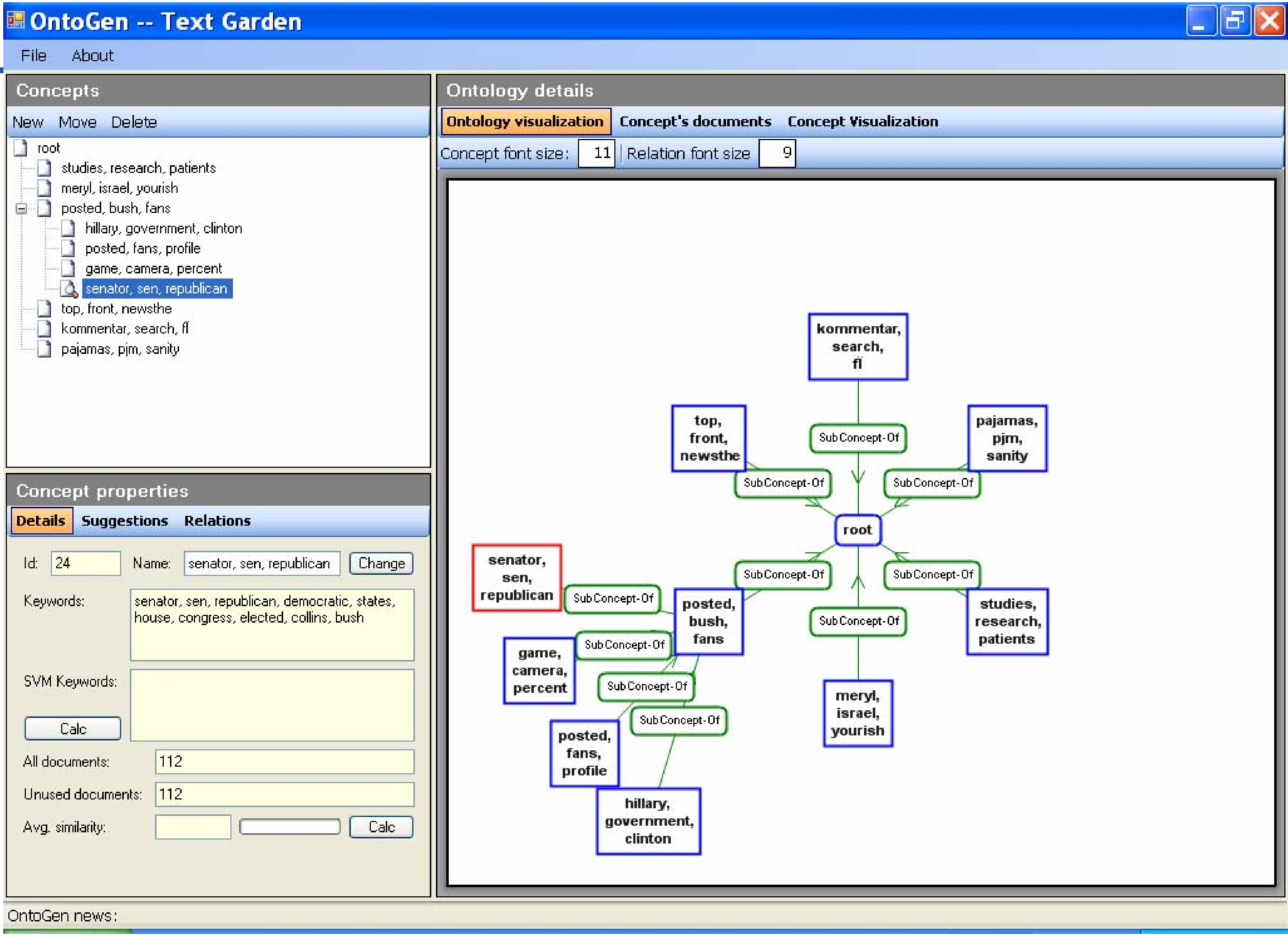


Concept font size: 11 | Relation font size 9

No. suggestions: 4

Docs: ● All ○ Unused

Ontology learning (2)



- root
  - studies, research, patients
  - meryl, israel, yourish
  - posted, bush, fans
    - hillary, government, clinton
    - posted, fans, profile
    - game, camera, percent
    - senator, sen, republican
  - top, front, newsthe
  - kommentar, search, fi
  - pajamas, pjm, sanity

Concept font size: 11 Relation font size 9

Id: 24

Name: senator, sen, republican

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

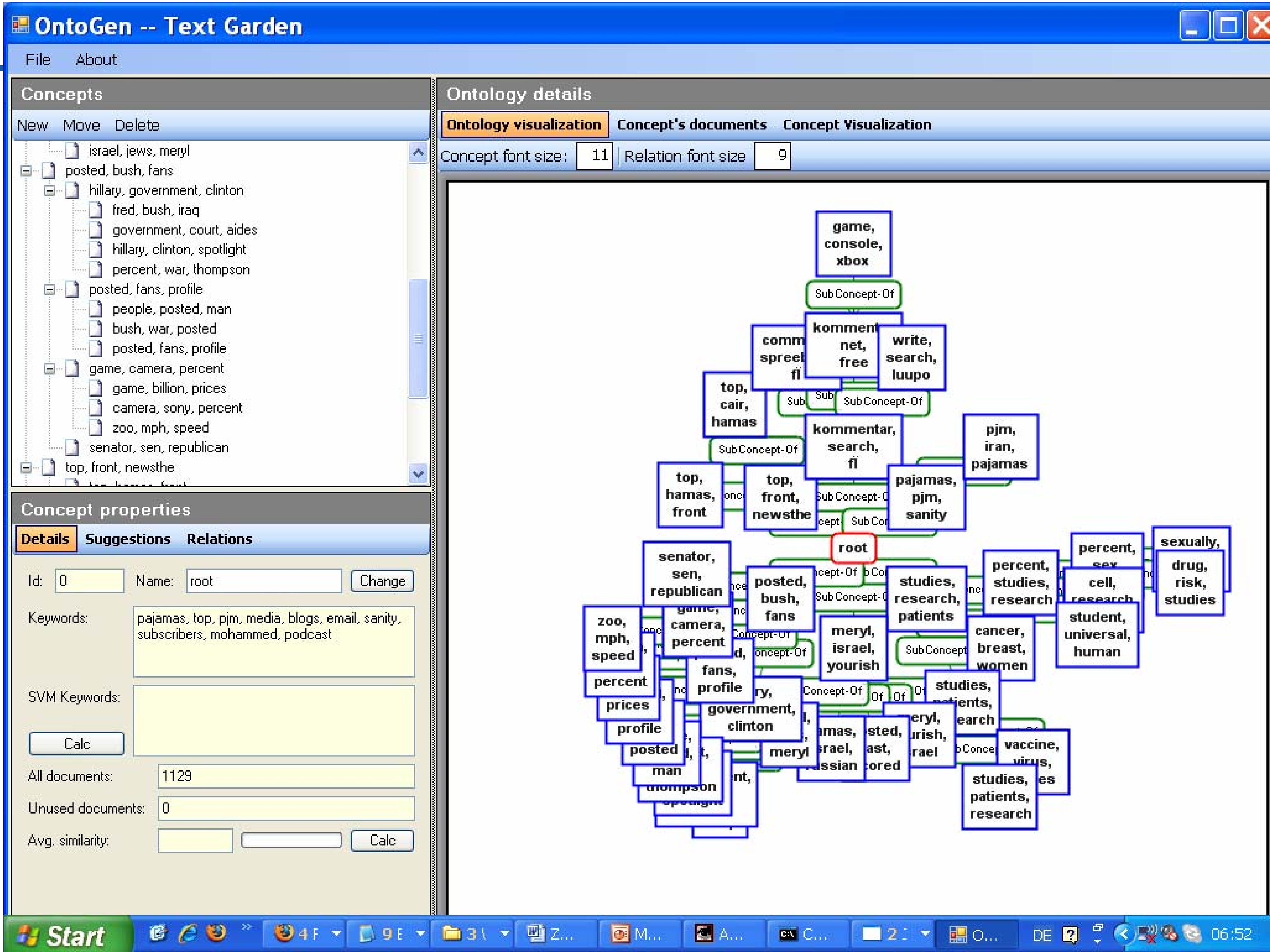
SVM Keywords:

All documents: 112

Unused documents: 112

Avg. similarity:

Ontology learning (3)



- israel, jews, meryl
- posted, bush, fans
  - hillary, government, clinton
    - fred, bush, iraq
    - government, court, aides
    - hillary, clinton, spotlight
    - percent, war, thompson
  - posted, fans, profile
    - people, posted, man
    - bush, war, posted
    - posted, fans, profile
  - game, camera, percent
    - game, billion, prices
    - camera, sony, percent
    - zoo, mph, speed
  - senator, sen, republican
- top, front, newsthe

Concept font size: 11 Relation font size 9

Id: 0

Name: root

Keywords: pajamas, top, pjm, media, blogs, email, sanity, subscribers, mohammed, podcast

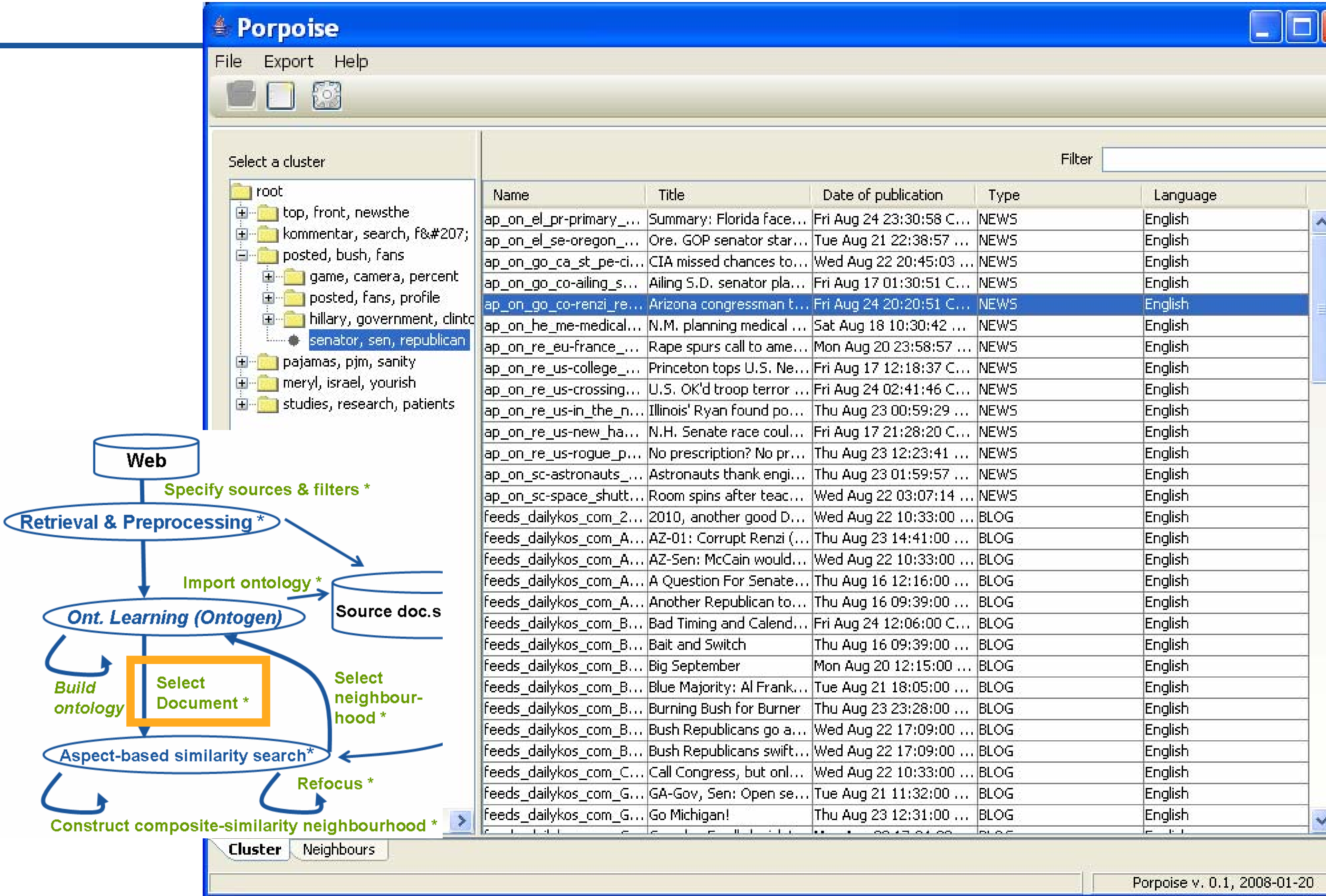
SVM Keywords:

All documents: 1129

Unused documents: 0

Avg. similarity:

Inspection of ontology and instances



Web

Specify sources & filters *

Retrieval & Preprocessing *

Import ontology *

Source doc.s

Ont. Learning (Ontogen)

Build ontology

Select Document *

Select neighbour-hood *

Aspect-based similarity search*

Refocus *

Construct composite-similarity neighbourhood *

Select a cluster

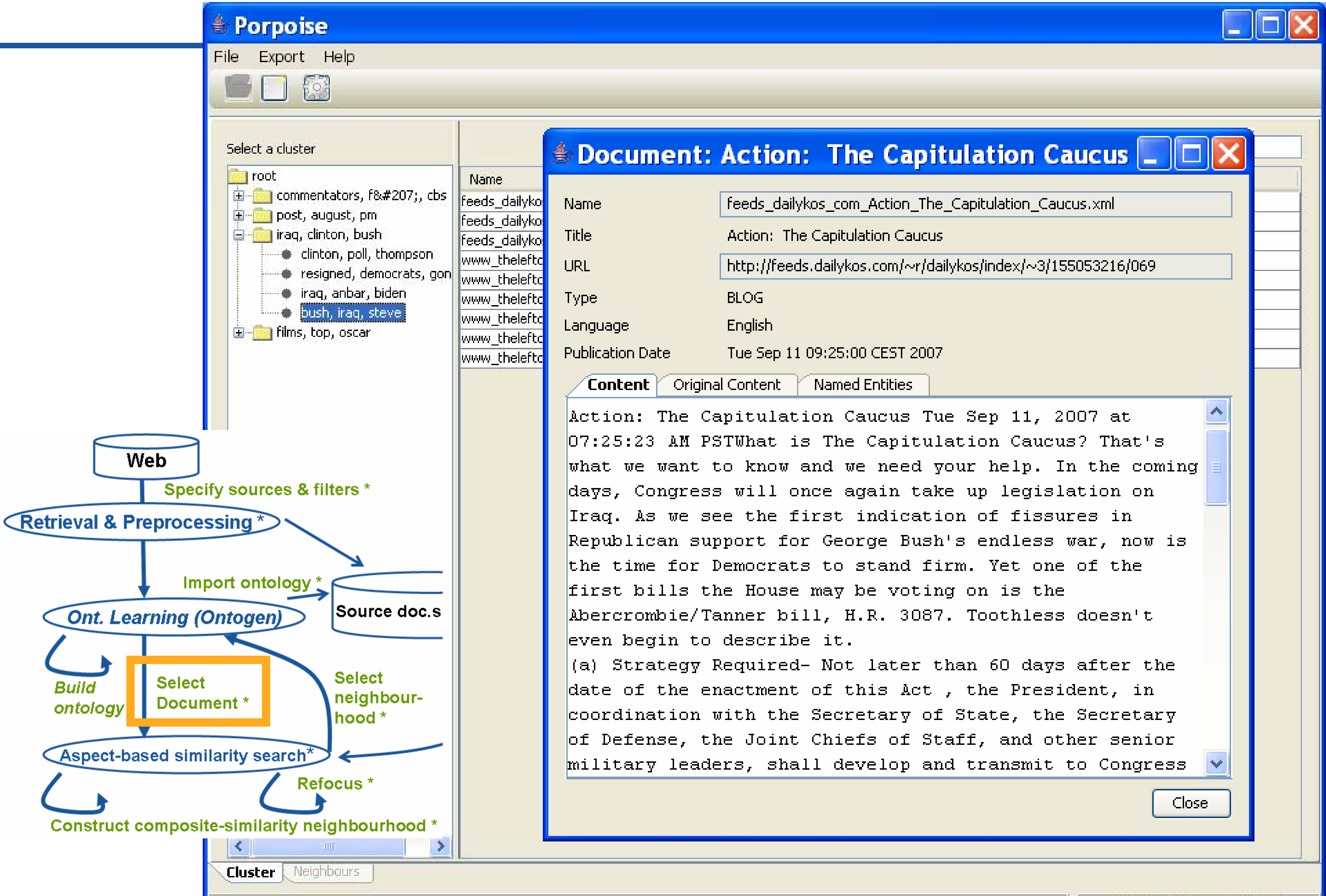
- root
  - top, front, newsthe
  - kommentar, search, f&#207;
  - posted, bush, fans
    - game, camera, percent
    - posted, fans, profile
    - hillary, government, clintc
    - senator, sen, republican
  - pajamas, pjm, sanity
  - meryl, israel, yourish
  - studies, research, patients

Filter

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



Document: Action: The Capitulation Caucus

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
```

Web

Specify sources & filters *

Retrieval & Preprocessing *

Import ontology *

Source doc.s

Ont. Learning (Ontogen)

Build ontology

Select Document *

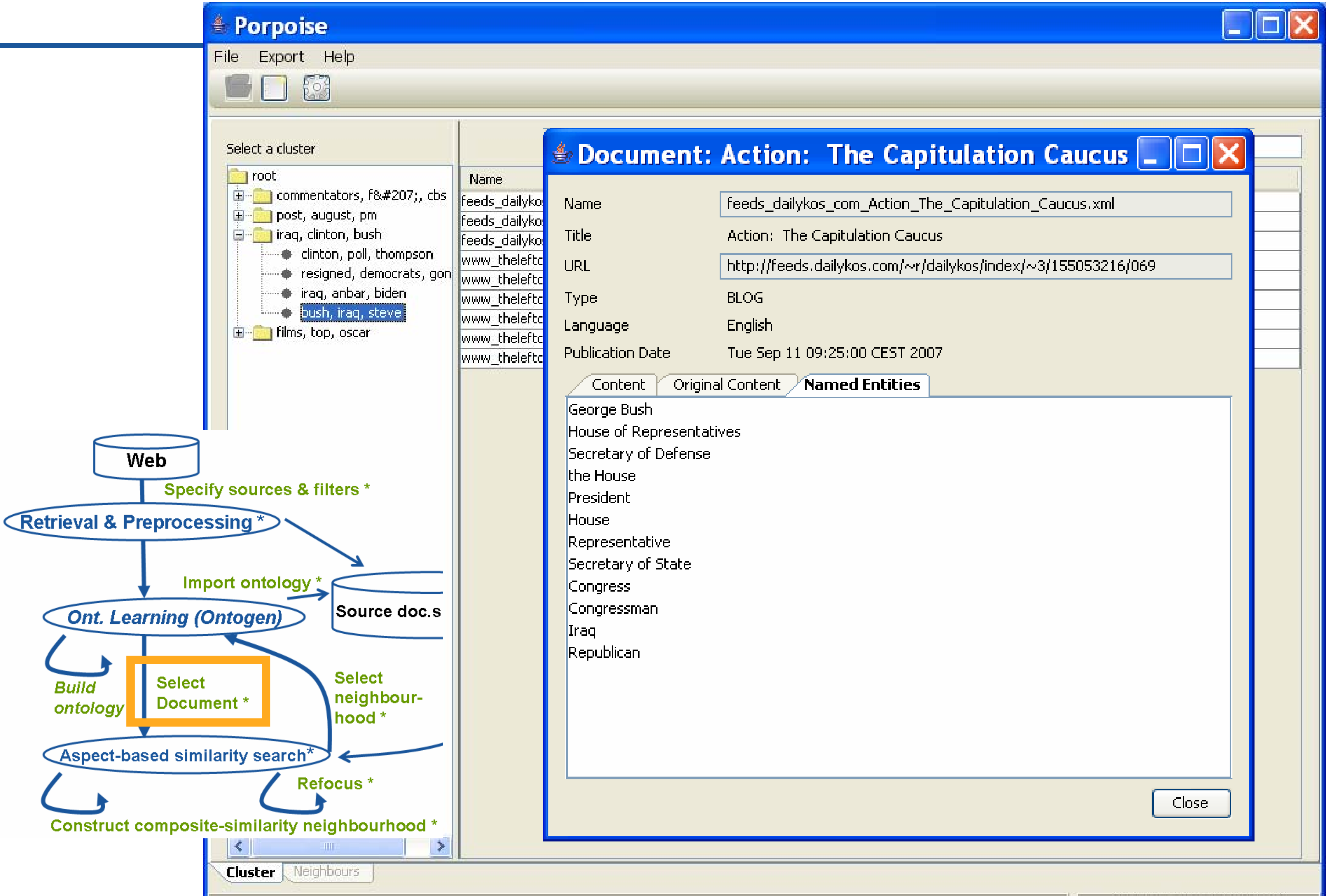
Select neighbourhood *

Aspect-based similarity search*

Refocus *

Construct composite-similarity neighbourhood *

More on documents



- root
  - commentators, f&#207;, cbs
  - post, august, pm
  - iraq, clinton, bush
    - clinton, poll, thompson
    - resigned, democrats, gon
    - iraq, anbar, biden
    - bush, iraq, steve
  - films, top, oscar

feeds_dailyko
feeds_dailyko
feeds_dailyko
www_theleftc
www_theleftc
www_theleftc
www_theleftc
www_theleftc
www_theleftc

Document: Action: The Capitulation Caucus

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

Named Entities:

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

Web

Specify sources & filters *

Retrieval & Preprocessing *

Import ontology *

Source doc.s

Ont. Learning (Ontogen)

Build ontology

Select Document *

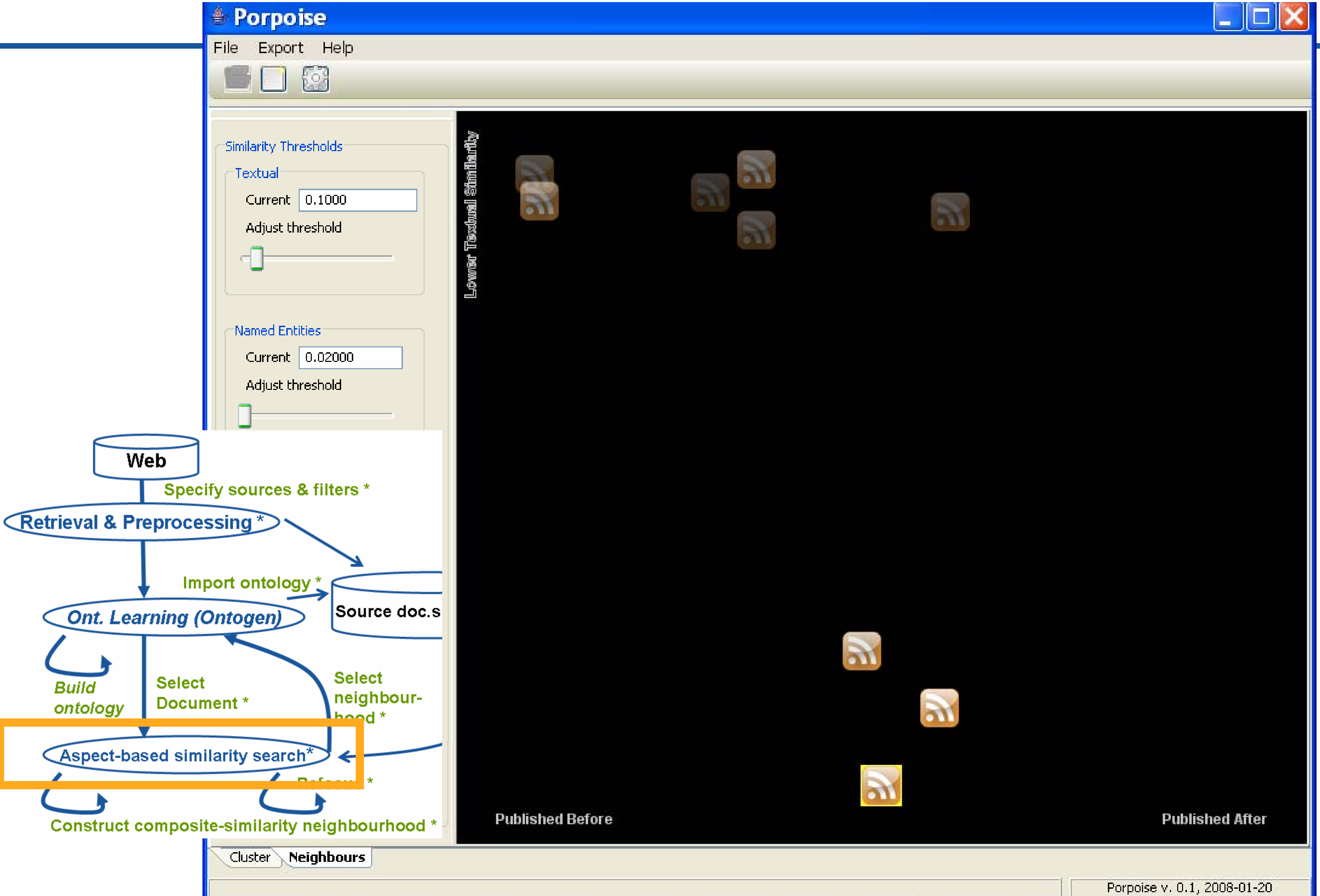
Select neighbourhood *

Aspect-based similarity search*

Refocus *

Construct composite-similarity neighbourhood *

The neighbourhood of a document



Similarity Thresholds

Textual

Current 0.1000

Adjust threshold

Named Entities

Current 0.02000

Adjust threshold

Web

Specify sources & filters *

Retrieval & Preprocessing *

Import ontology *

Source doc.s

Ont. Learning (Ontogen)

Build ontology

Select Document *

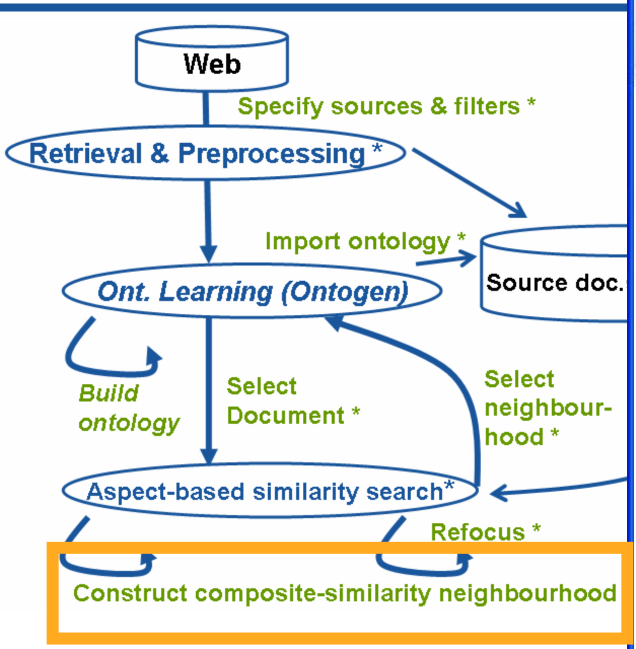
Select neighbourhood *

Aspect-based similarity search*

Refocus *

Construct composite-similarity neighbourhood *

Constructing the similarity measure & neighbourhood (I)



Web

Specify sources & filters *

Retrieval & Preprocessing *

Import ontology *

Source doc.

Ont. Learning (Ontogen)

Build ontology

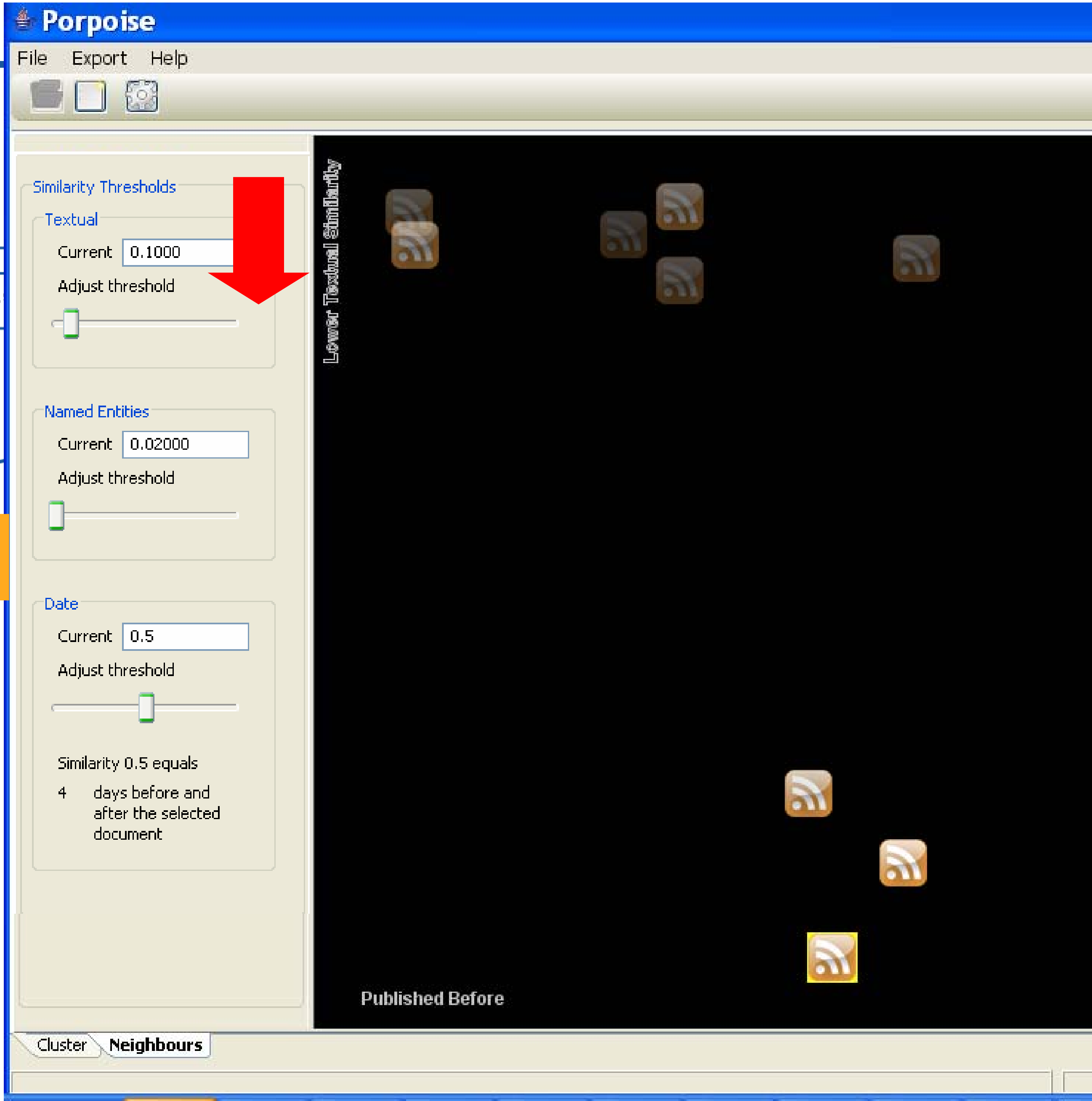
Select Document *

Select neighbour-hood *

Aspect-based similarity search*

Refocus *

Construct composite-similarity neighbourhood



Similarity Thresholds

Textual

Current 0.1000

Adjust threshold

Named Entities

Current 0.02000

Adjust threshold

Date

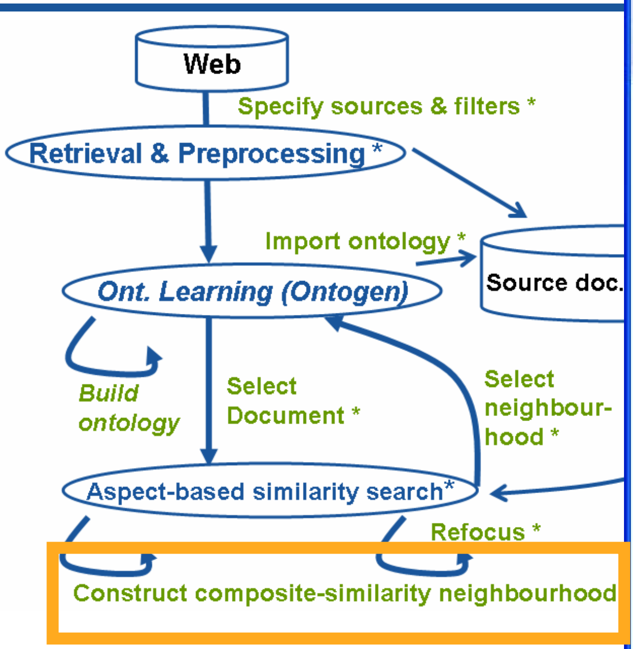
Current 0.5

Adjust threshold

Similarity 0.5 equals

4 days before and after the selected document

Constructing the similarity measure & neighbourhood (II)



Web

Specify sources & filters *

Retrieval & Preprocessing *

Import ontology *

Source doc.

Ont. Learning (Ontogen)

Build ontology

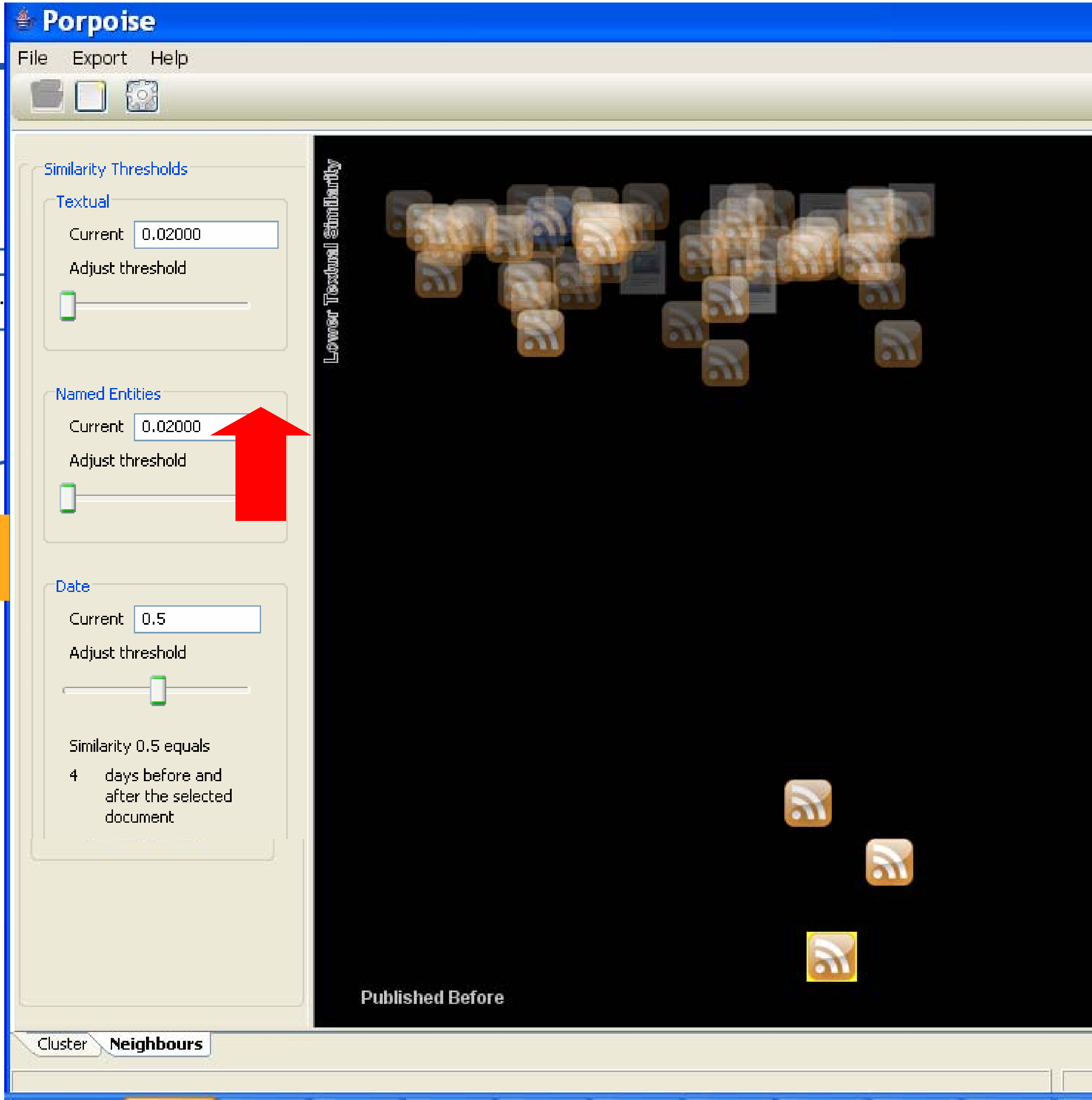
Select Document *

Select neighbourhood *

Aspect-based similarity search*

Refocus *

Construct composite-similarity neighbourhood



Similarity Thresholds

Textual

Current 0.02000

Adjust threshold

Named Entities

Current 0.02000

Adjust threshold

Date

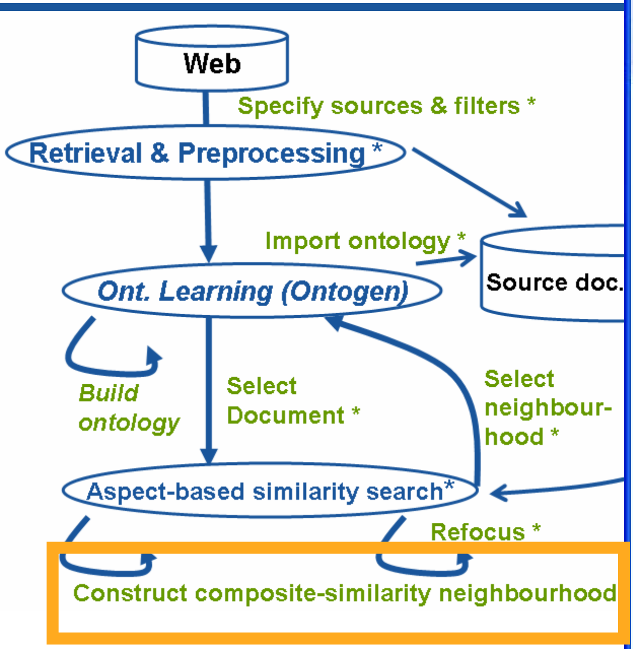
Current 0.5

Adjust threshold

Similarity 0.5 equals

4 days before and after the selected document

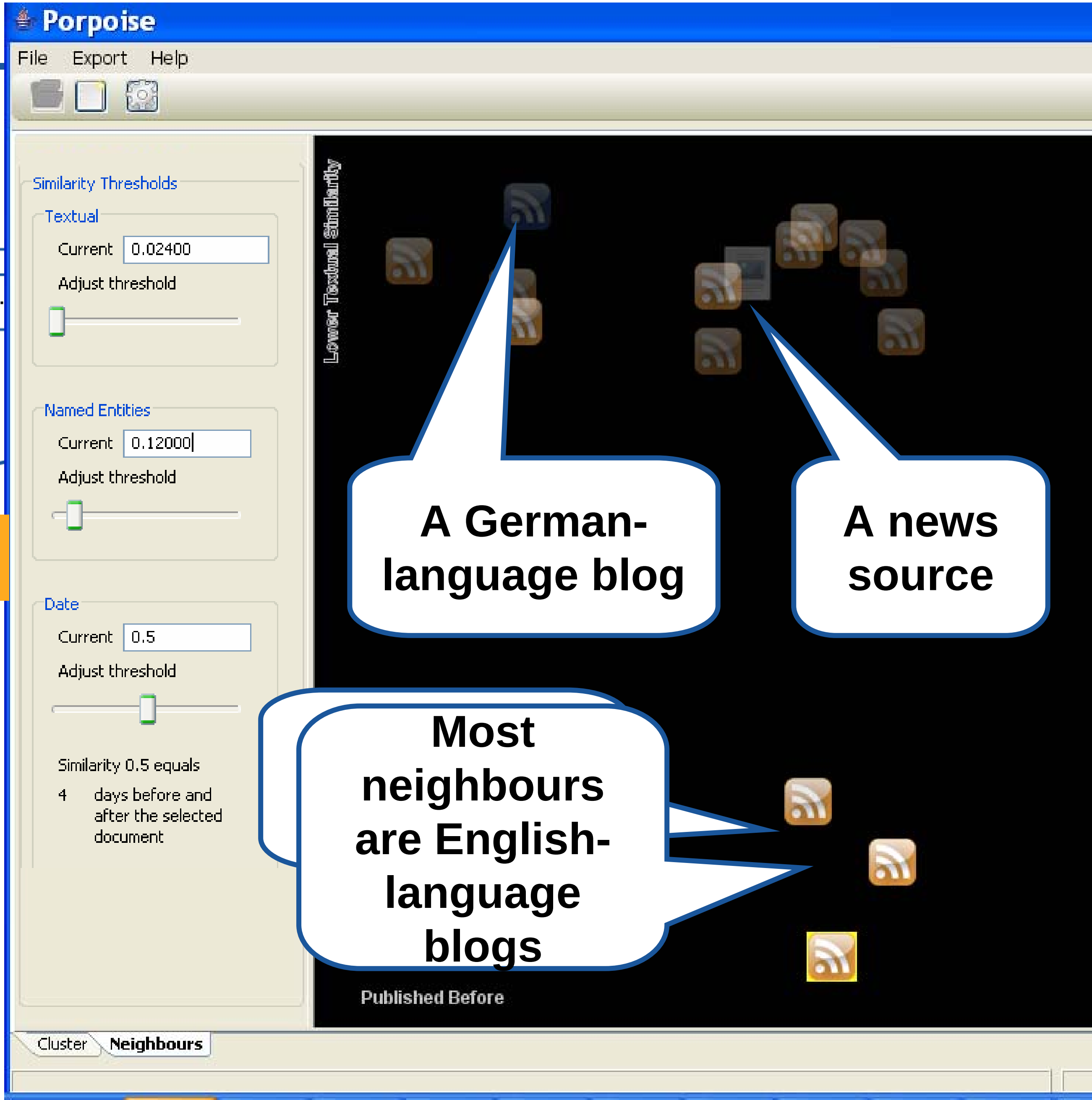
Constructing the similarity measure & neighbourhood (III)


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

English blog

German blog

English news



Similarity Thresholds

Textual
Current 0.02400
Adjust threshold

Named Entities
Current 0.12000
Adjust threshold

Date
Current 0.5
Adjust threshold

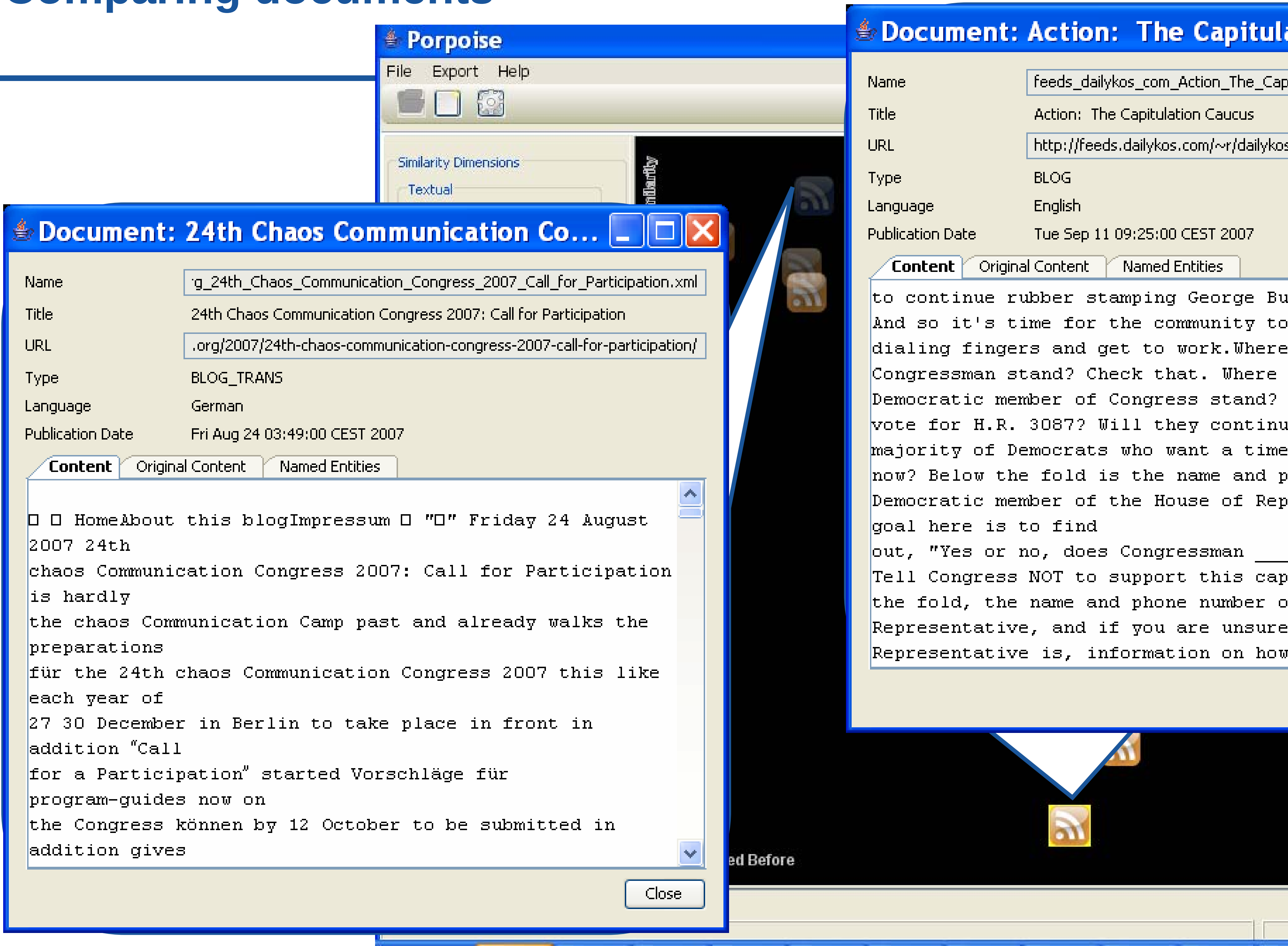
Similarity 0.5 equals
4 days before and after the selected document

A German-language blog

A news source

Most neighbours are English-language blogs

Comparing documents



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

Content | Original Content | Named Entities

```
□ □ 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
```

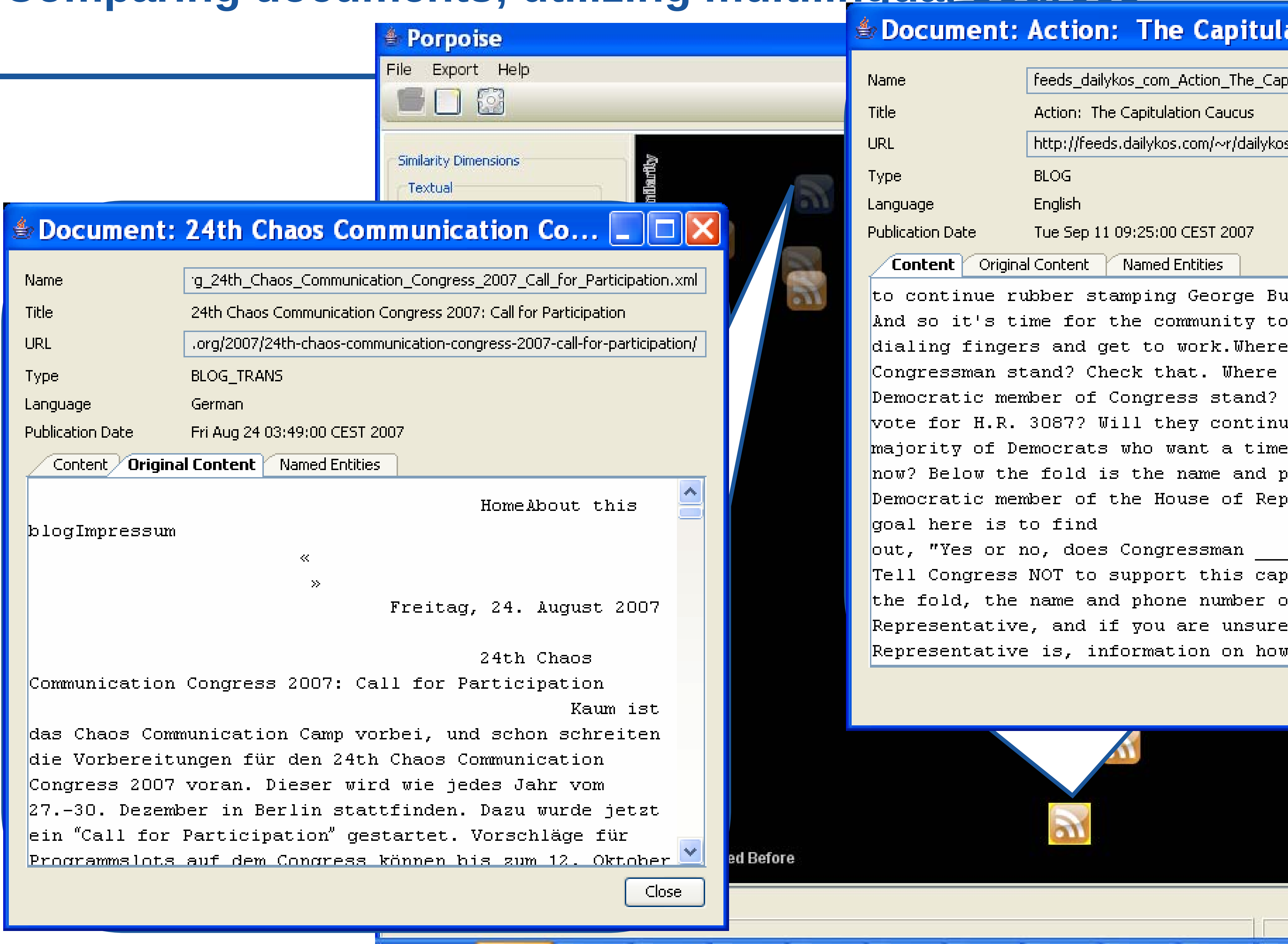
Document: Action: The Capitul

Name	feeds_dailykos_com_Action_The_Cap
Title	Action: The Capitulation Caucus
URL	http://feeds.dailykos.com/~r/dailykos
Type	BLOG
Language	English
Publication Date	Tue Sep 11 09:25:00 CEST 2007

Content | Original Content | Named Entities

```
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 continu
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



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

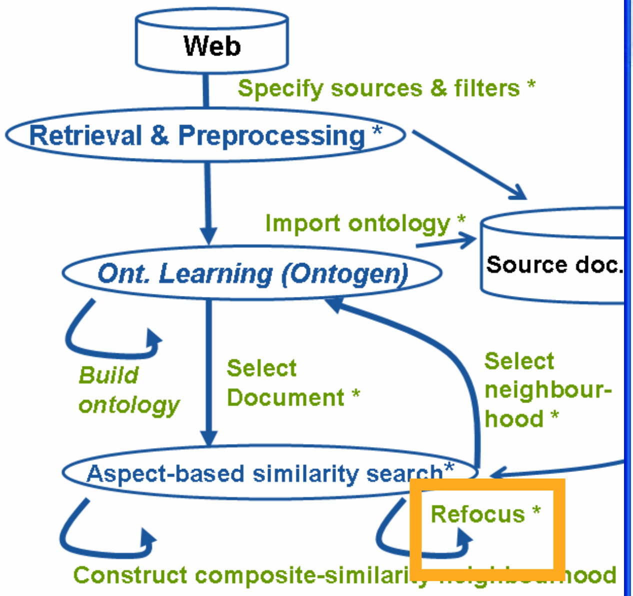
Kaum ist das Chaos Communication Camp vorbei, und schon schreiten die Vorbereitungen für den 24th Chaos Communication Congress 2007 voran. Dieser wird wie jedes Jahr vom 27.-30. Dezember in Berlin stattfinden. Dazu wurde jetzt ein "Call for Participation" gestartet. Vorschläge für Programmslots auf dem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

Document: Action: The Capitula

Name	feeds_dailykos_com_Action_The_Cap
Title	Action: The Capitulation Caucus
URL	http://feeds.dailykos.com/~r/dailykos
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 continu
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

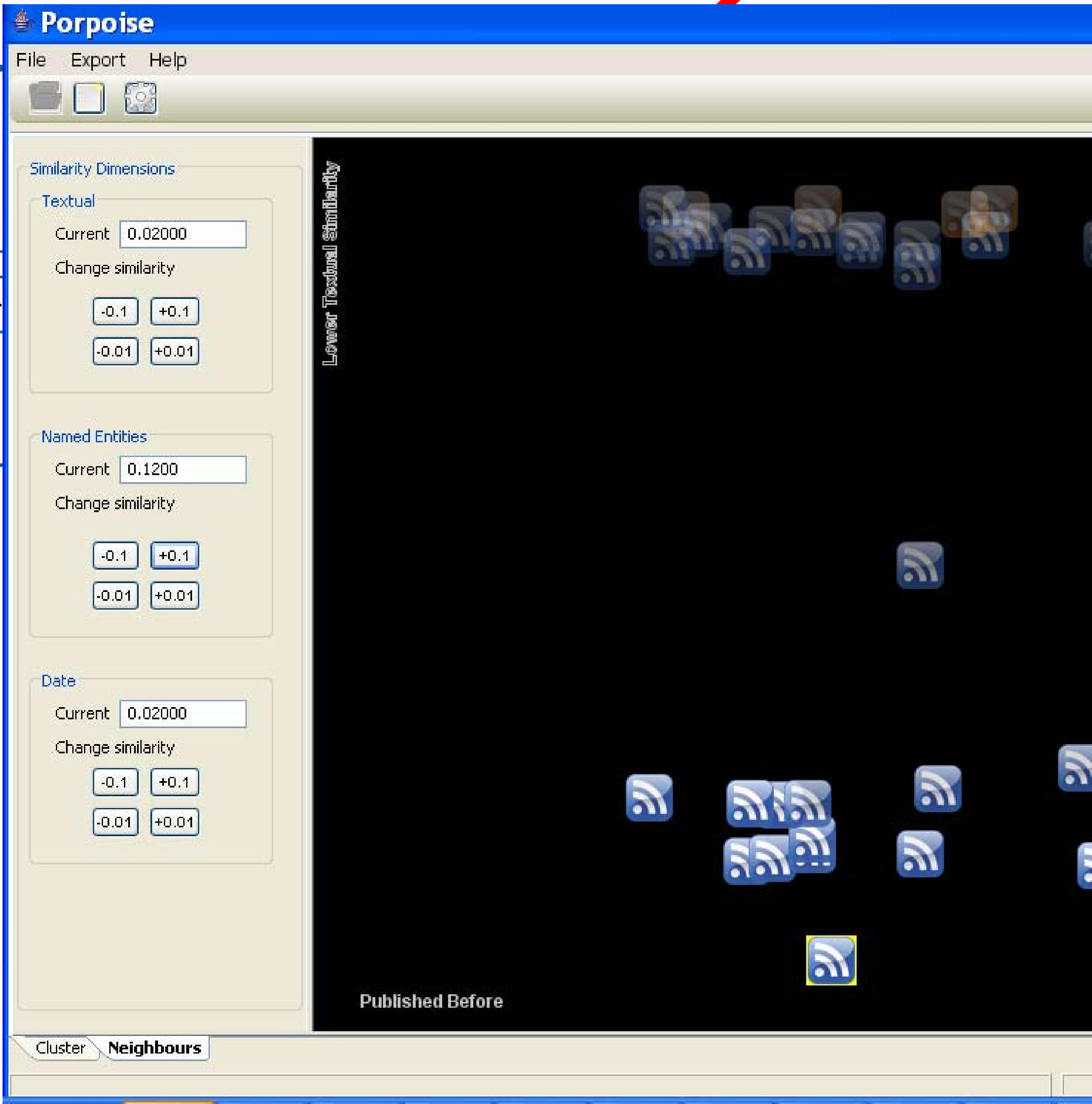
Select Document *

Select neighbour-hood *

Aspect-based similarity search*

Refocus *

Construct composite-similarity neighbourhood



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

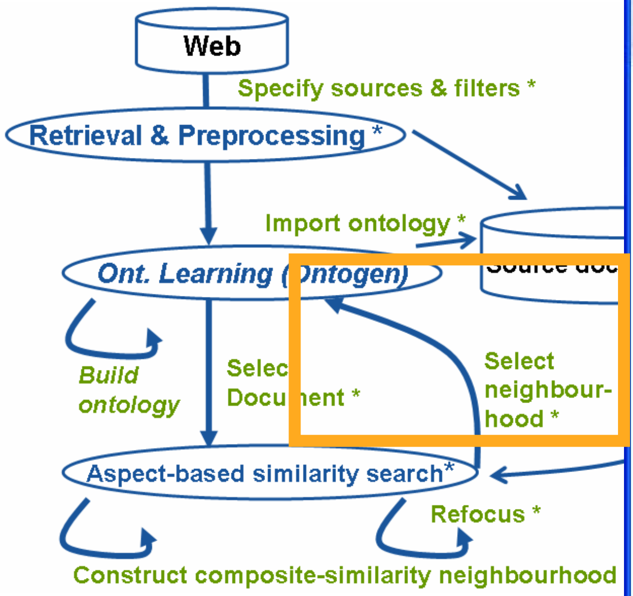
Date

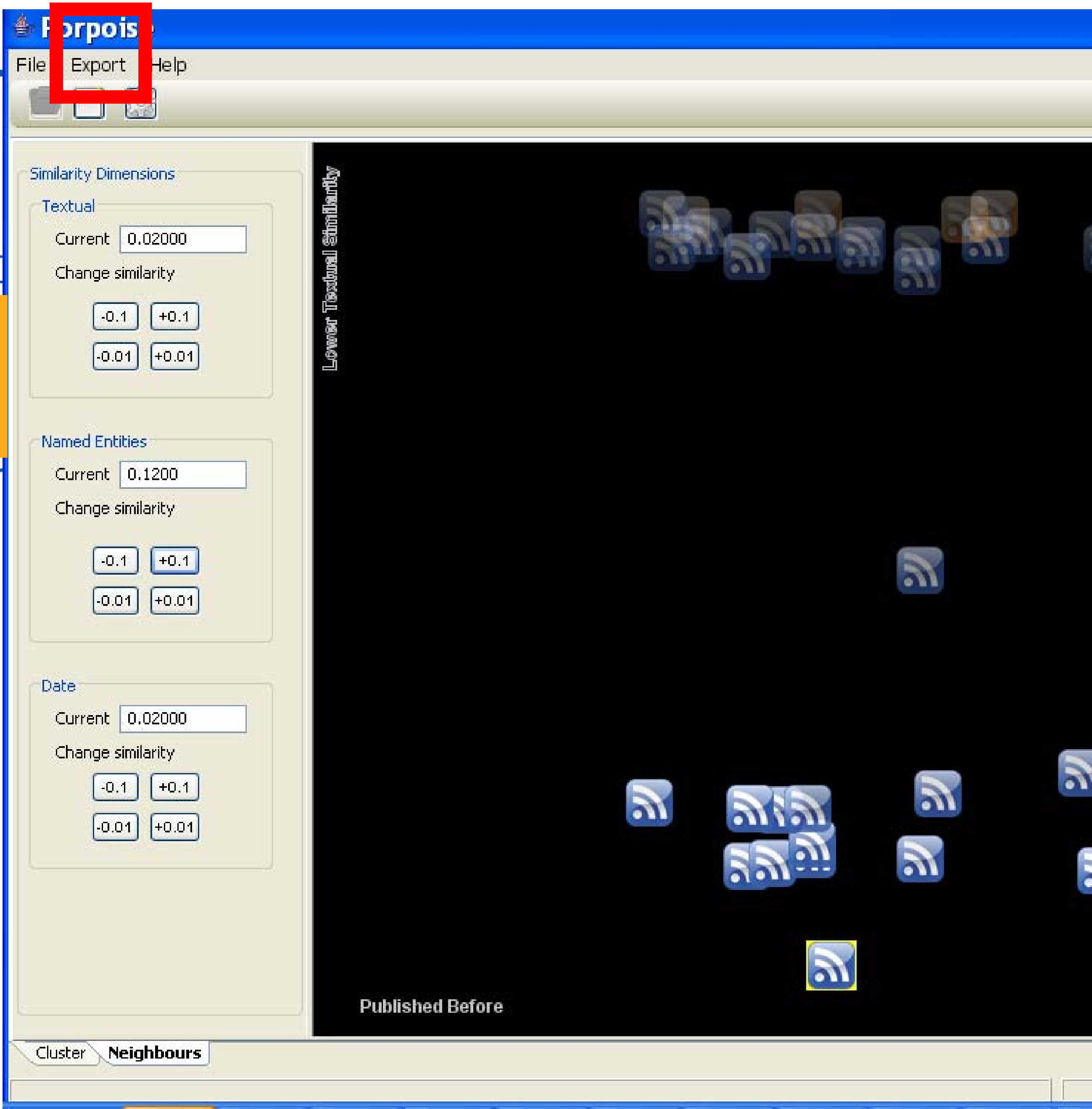
Current 0.02000

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



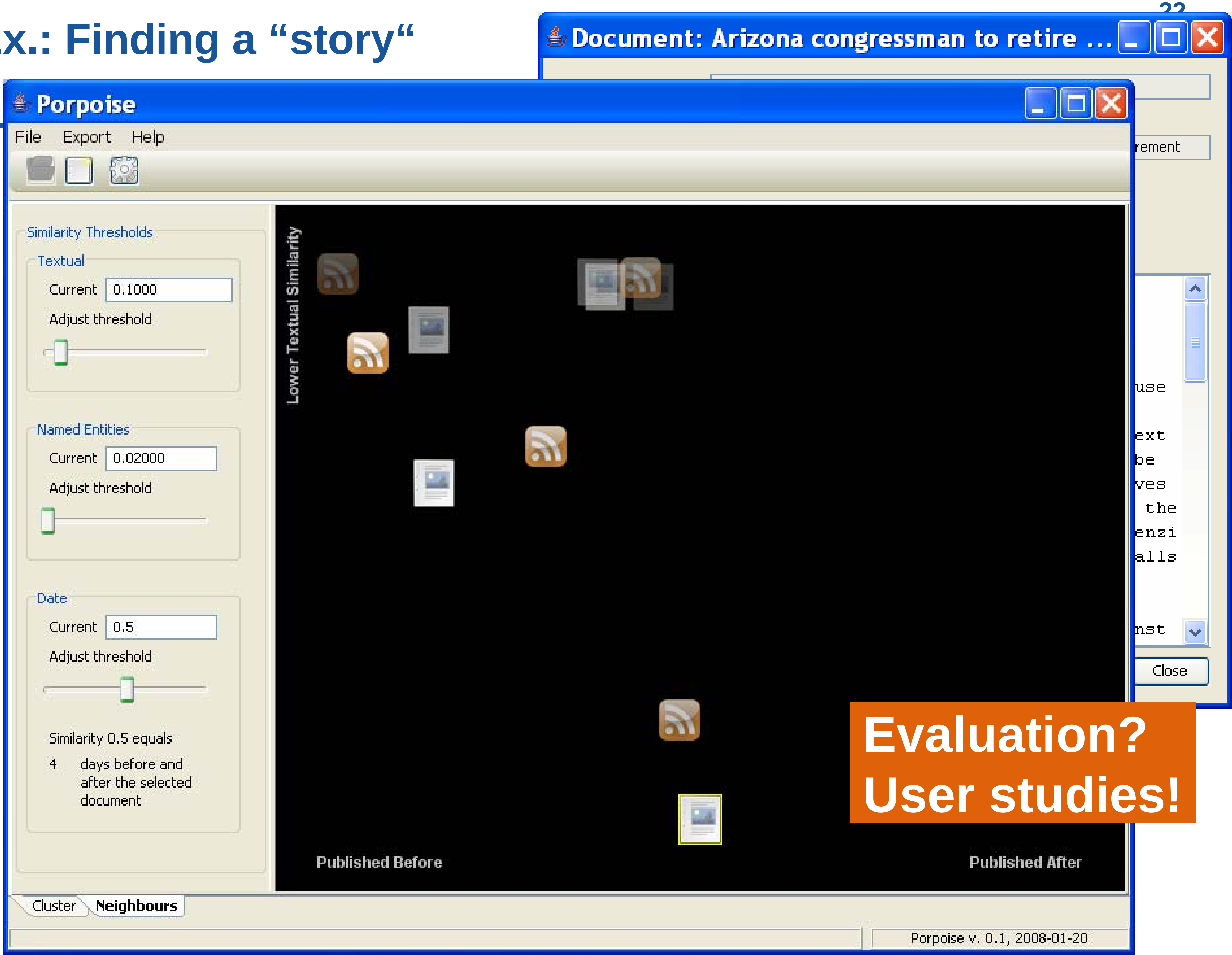
Similarity Dimensions

Textual
Current 0.02000
Change similarity

Named Entities
Current 0.1200
Change similarity

Date
Current 0.02000
Change similarity

Ex.: Finding a “story“



Document: Arizona congressman to retire ...

Porpoise

Similarity Thresholds

Textual

Current 0.1000

Adjust threshold

Named Entities

Current 0.02000

Adjust threshold

Date

Current 0.5

Adjust threshold

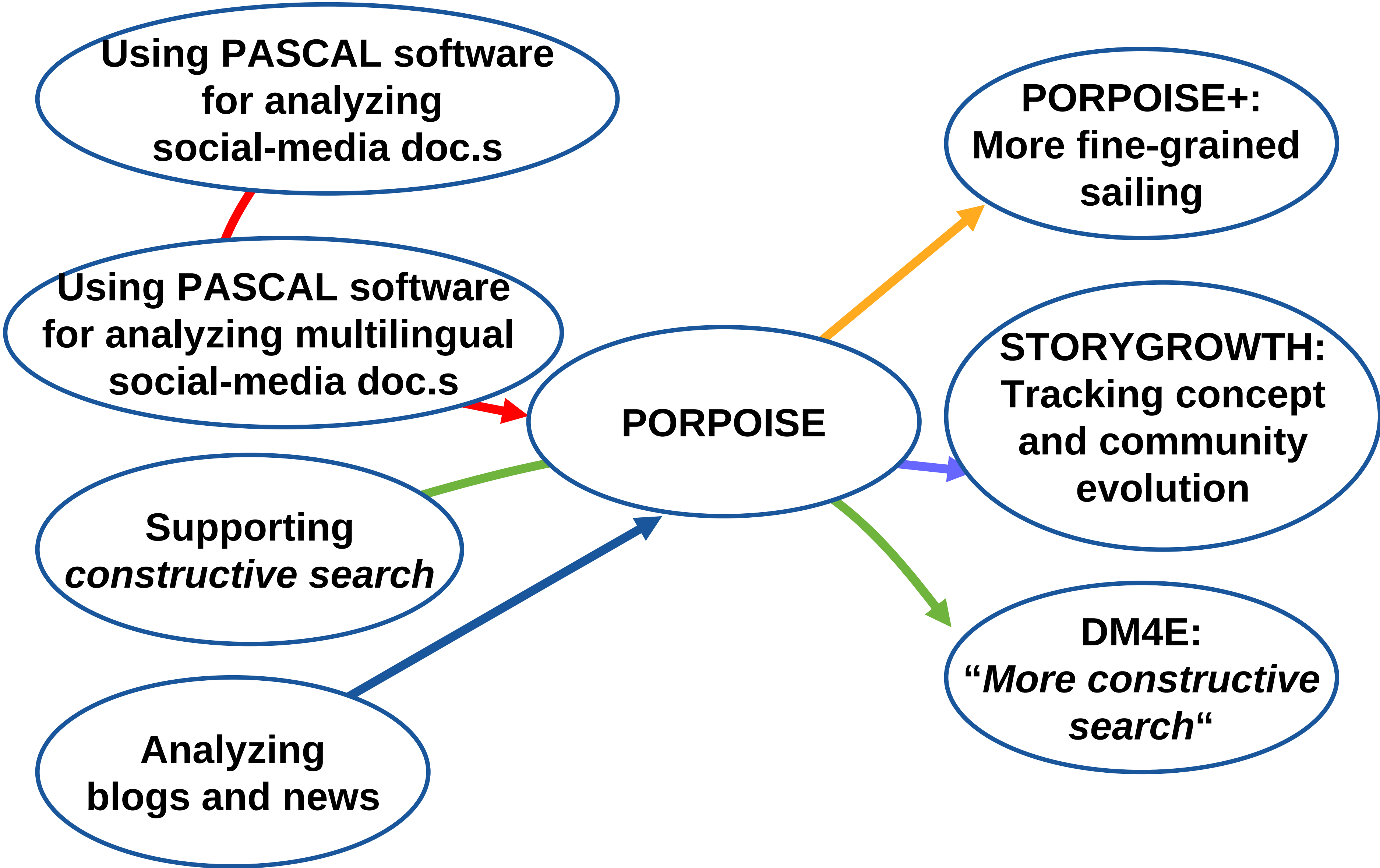
Similarity 0.5 equals

4 days before and after the selected document

Porpoise v. 0.1, 2008-01-20

Evaluation?
User studies!

“Pump-priming“: PORPOISE as catalyst



Using PASCAL software for analyzing social-media doc.s

Using PASCAL software for analyzing multilingual social-media doc.s

Supporting *constructive search*

Analyzing blogs and news

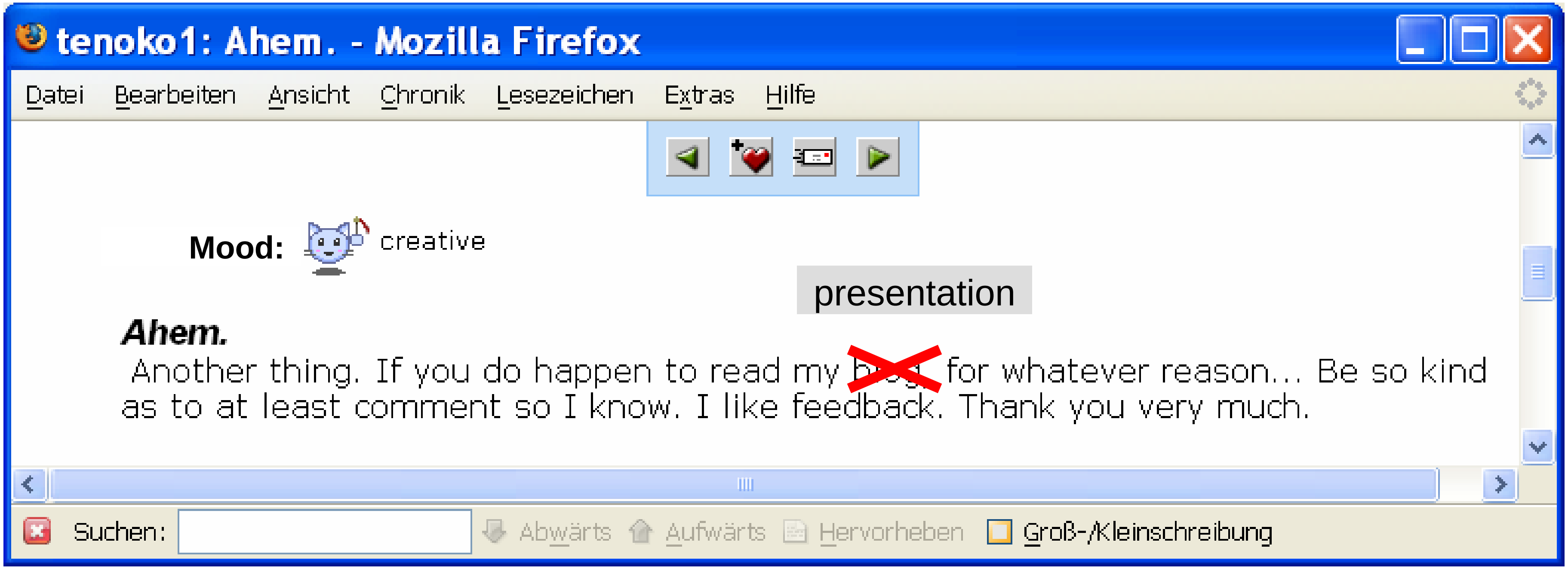
PORPOISE

PORPOISE+: More fine-grained sailing

STORYGROWTH: Tracking concept and community evolution

DM4E: “*More constructive search*“

Finally ... could I express it better?



Mood: creative

presentation

Ahem.

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