

# 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](http://www.cs.kuleuven.be/~berendt)

# ICT Motivation:

## Global+local interaction; beyond “similar documents”

The image shows a screenshot of a Google search results page for the query "similarity". The browser is Mozilla Firefox. The search results show the Wikipedia entry for "Similarity". A thought bubble with the text "with respect to what?" is overlaid on the page, containing two 2x2 grids of colored squares (green and red). Below the thought bubble, a link "Similar pages" is highlighted with an orange box. The bottom of the page shows a sidebar with related searches: "Similarity Search", "Self-Similarity", and "Geometry".

similarity - Google Search - Mozilla Firefox

Web Images Maps News Shopping Gmail

Google similarity

Search: ☒ the web ☐ pages

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 exists either on...

[en.wikipedia.org/wiki/Similarity](#) - 18k - [Cached](#) - [Similar pages](#)

Suchen:  Abwärts Aufwärts Hervorheben

Fertig

Similarity Search (15)

Self-Similarity (13)

Geometry (9)

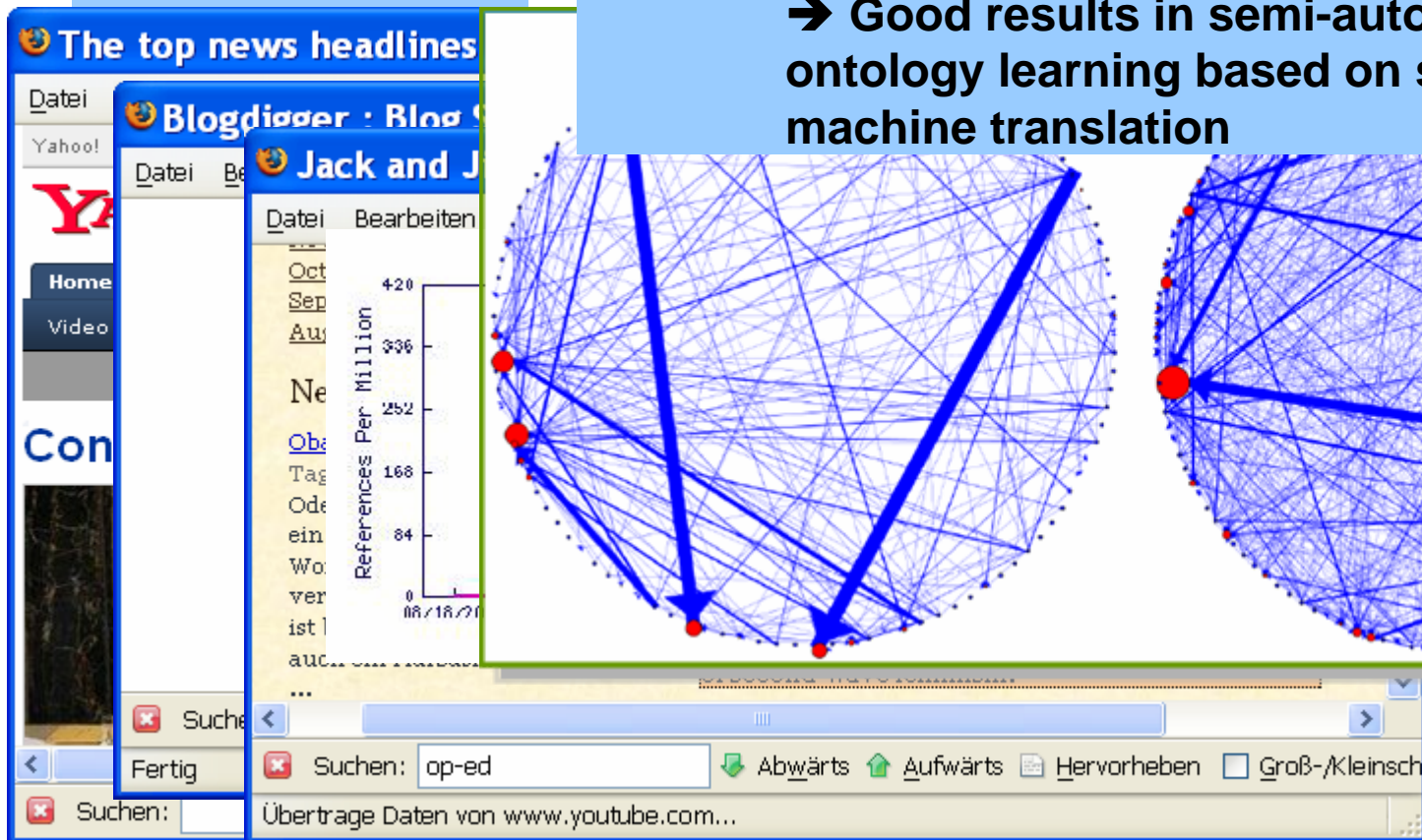
3. [Protein ranking: From local to global structure in the protein ...](#)

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

Suchen:  Abwärts Aufwärts Hervorheben ☐ Groß-/Kleinschreibung

# 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

The screenshot displays the OntoGen -- Text Garden application window. The interface is divided into several sections:

- Concepts:** A tree view on the left showing a hierarchy starting from 'root'. The selected concept is 'senator, sen, republican'.
- Concept properties:** A panel below the Concepts list showing details for the selected concept. It includes fields for Id (24), Name (senator, sen, republican), Keywords (senator, sen, republican, democratic, states, house, congress, elected, collins, bush), SVM Keywords, and document counts (All documents: 112, Unused documents: 112).
- Ontology details:** A panel on the right showing the 'Ontology visualization' tab. It includes sliders for 'Concept font size' (11) and 'Relation font size' (9).
- Ontology visualization:** A large central area displaying a hierarchical tree diagram. The root node is 'root'. It has several children, including 'kommentar, search, fi', 'top, front, newsthe', 'pajamas, pj, sanity', 'studies, research, patients', 'meryl, israel, yourish', 'posted, bush, fans', 'game, camera, percent', 'posted, fans, profile', and 'hillary, government, clinton'. The 'senator, sen, republican' node is highlighted in red. Green arrows labeled 'Sub Concept-Of' indicate the relationships between the nodes.

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

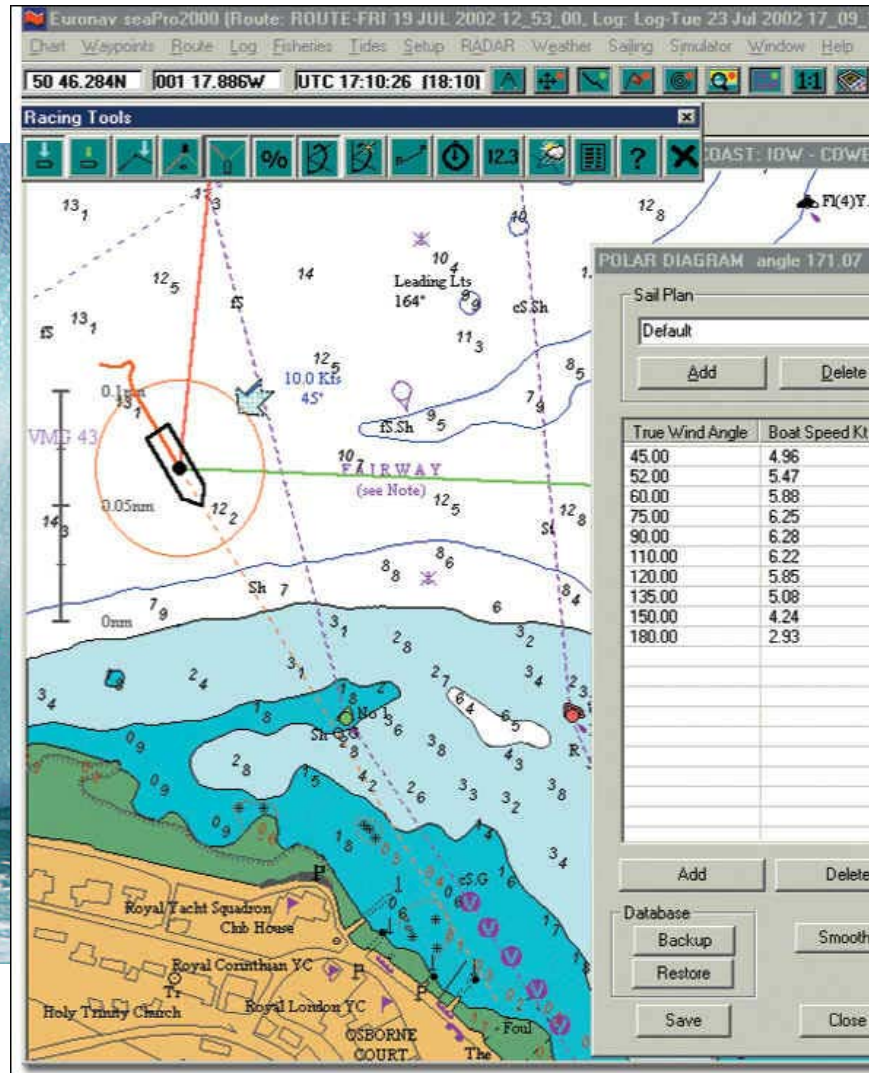


# Solution vision: *PORPOISE – Sailing the Internet*

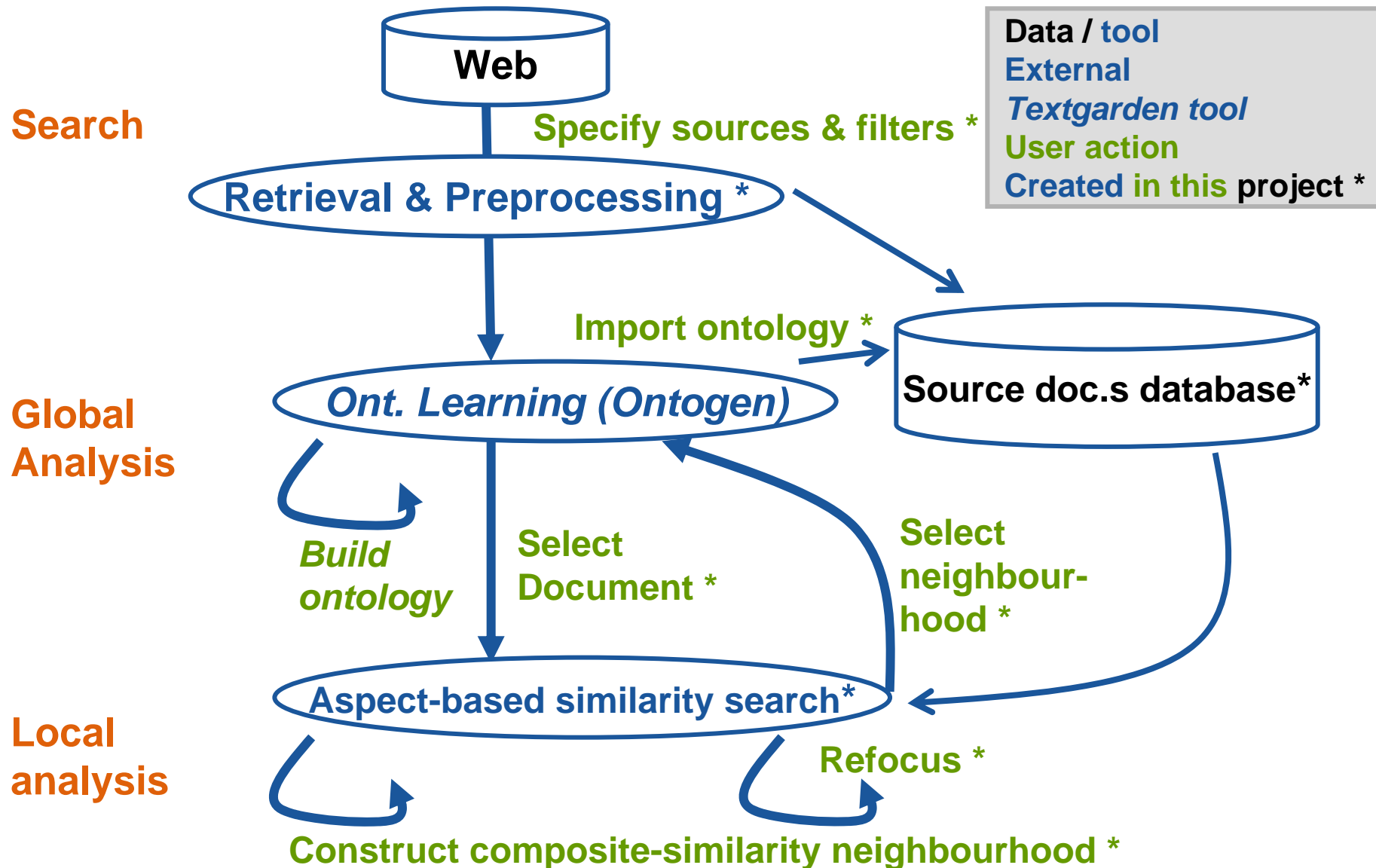
Search

Global  
Analysis

Local  
analysis



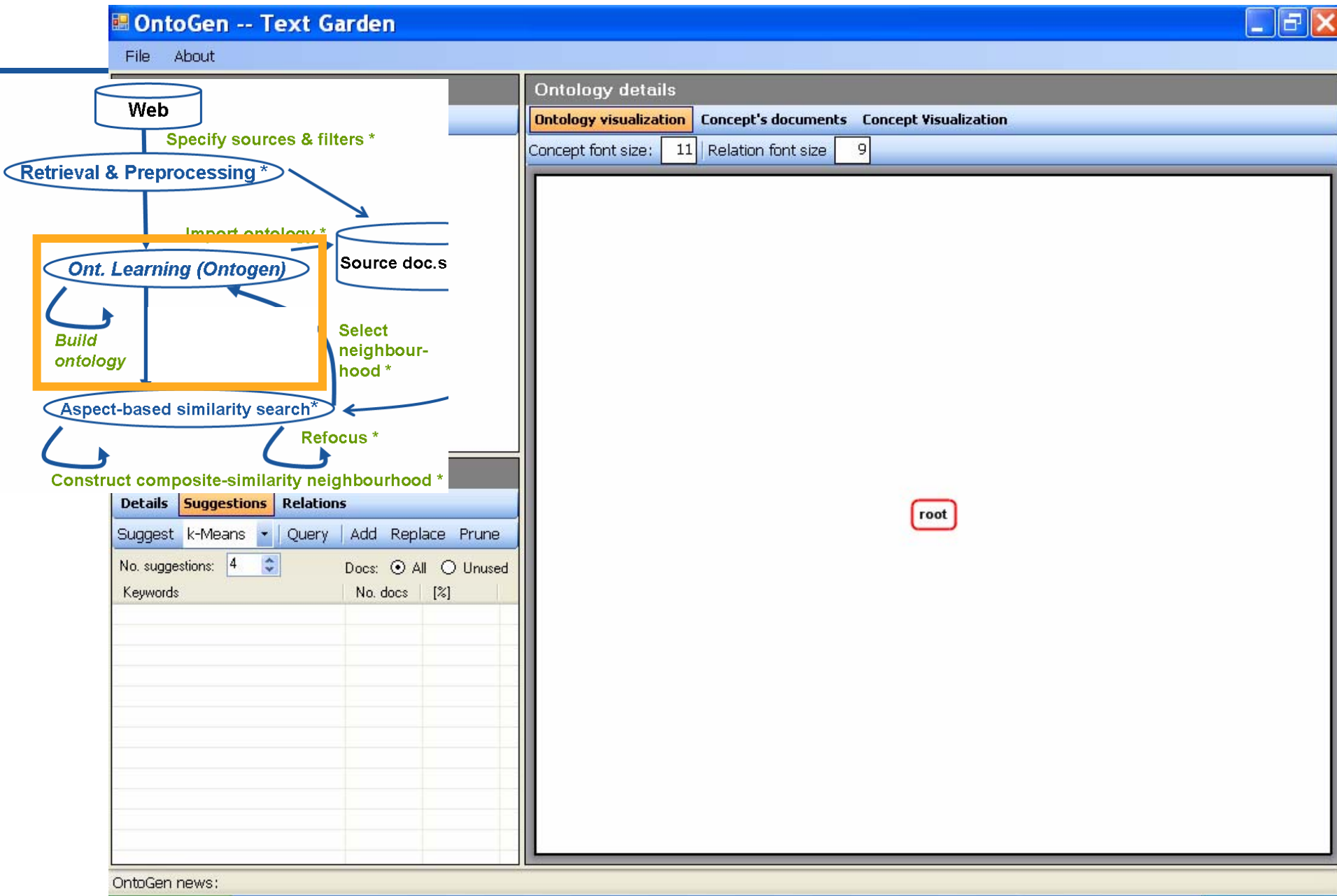
# Solution approach: Architecture & states overview



# Retrieval and preprocessing



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





**OntoGen -- Text Garden**

File About

### Concepts

New Move Delete

- 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, ff
  - pajamas, pjn, sanity

### Concept properties

**Details** Suggestions Relations

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 details

**Ontology visualization** Concept's documents Concept Visualization

Concept font size: 11 Relation font size: 9

```
graph TD
    root((root))
    kommentar[kommentar, search, ff] -- Sub Concept-Of --> root
    top[top, front, newsthe] -- Sub Concept-Of --> root
    pajamas[pajamas, pjn, sanity] -- Sub Concept-Of --> root
    senator[senator, sen, republican] -- Sub Concept-Of --> root
    posted_bush[posted, bush, fans] -- Sub Concept-Of --> root
    studies[studies, research, patients] -- Sub Concept-Of --> root
    meryl[meryl, israel, yourish] -- Sub Concept-Of --> root
    game[game, camera, percent] -- Sub Concept-Of --> posted_bush
    posted_profile[posted, fans, profile] -- Sub Concept-Of --> posted_bush
    hillary[hillary, government, clinton] -- Sub Concept-Of --> posted_profile
```

OntoGen news:



# Inspection of ontology and instances

**Porpoise**

File Export Help

Select a cluster

- root
  - top, front, newsthe
  - kommentar, search, f&#207;
  - posted, bush, fans
    - game, camera, percent
    - posted, fans, profile
    - hillary, government, clinto
    - senator, sen, republican
  - pajamas, pjw, 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

Web

Specify sources & filters \*

Retrieval & Preprocessing \*

Import ontology \*

Ont. Learning (Ontogen)

Source doc.s

Select Document \*

Select neighbourhood \*

Refocus \*

Construct composite-similarity neighbourhood \*

Cluster Neighbours

Porpoise v. 0.1, 2008-01-20



# Inspection of documents

The screenshot displays the Porpoise software interface, which is used for document inspection. The main window shows a file tree on the left and a document details pane on the right. A modal window titled "Document: Action: The Capitulation Caucus" is open, showing the document's metadata and content.

**Porpoise Interface:**

- File Tree (Left):** Shows a hierarchy starting from "root", with folders like "commentators, f&#207;;, cbs", "post, august, pm", "iraq, clinton, bush", "clinton, poll, thompson", "resigned, democrats, gon", "iraq, anbar, biden", "bush, iraq, steve", and "films, top, oscar".
- Document Details (Right):**
  - 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

**Document Content Window:**

The "Content" tab is selected, showing the following text:

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

**Workflow Diagram (Bottom Left):**

```

graph TD
    Web((Web)) -- "Specify sources & filters *" --> RP([Retrieval & Preprocessing *])
    RP -- "Import ontology *" --> OL([Ont. Learning (Ontogen)])
    OL -- "Build ontology" --> AS([Aspect-based similarity search *])
    AS -- "Construct composite-similarity neighbourhood *" --> AS
    AS -- "Refocus *" --> AS
    AS -- "Select neighbourhood *" --> SD([Source doc.s])
    SD -- "Select Document *" --> OL
    OL -- "Select Document *" --> SD
  
```

The diagram illustrates the workflow for document inspection, starting from the Web, through Retrieval & Preprocessing, Ont. Learning (Ontogen), and Aspect-based similarity search, leading to the selection of a document and the construction of a composite-similarity neighbourhood.

# More on documents

**Porpoise**

File Export Help

Select a cluster

- root
  - commentators, f8&#207;, cbs
  - post, august, pm
  - iraq, clinton, bush
    - clinton, poll, thompson
    - resigned, democrats, gon
    - iraq, anbar, biden
    - bush, iraq, steve**
  - films, top, oscar

Name

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

Content Original Content **Named Entities**

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

Close

Web

Specify sources & filters \*

Retrieval & Preprocessing \*

Import ontology \*

Ont. Learning (Ontogen)

Source doc.s

Build ontology

Select Document \*

Select neighbour-hood \*

Aspect-based similarity search\*

Refocus \*

Construct composite-similarity neighbourhood \*

Cluster Neighbours

# The neighbourhood of a document

**Porpoise**

File Export Help

Similarity Thresholds

Textual

Current: 0.1000

Adjust threshold

Named Entities

Current: 0.02000

Adjust threshold

Lower Textual similarity

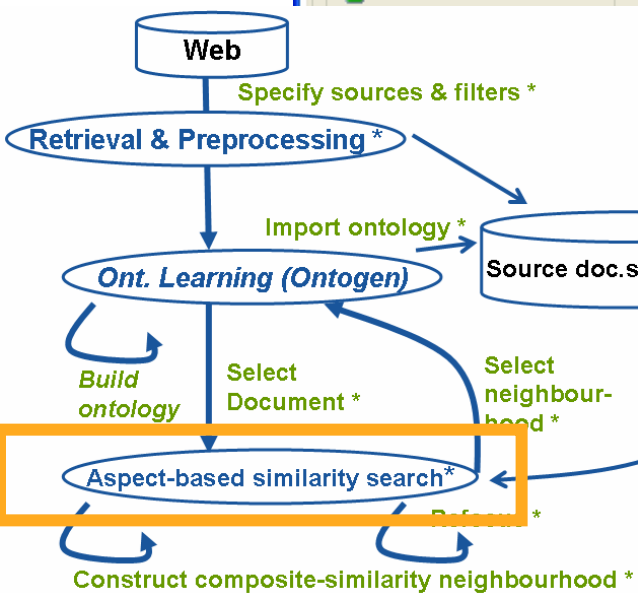
Published Before

Published After

Cluster Neighbours

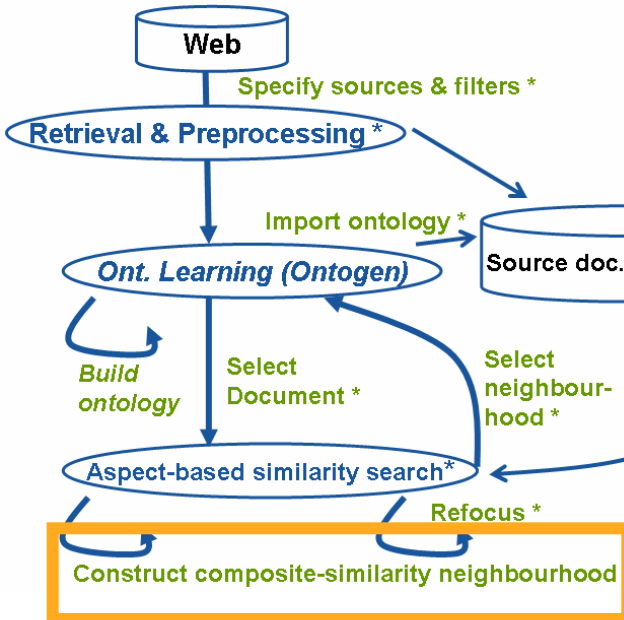
Porpoise v. 0.1, 2008-01-20

The screenshot shows the Porpoise software interface. On the left, there are two panels for adjusting similarity thresholds. The 'Textual' panel has a 'Current' value of 0.1000 and an 'Adjust threshold' slider. The 'Named Entities' panel has a 'Current' value of 0.02000 and an 'Adjust threshold' slider. The main area on the right is a dark visualization of document neighborhoods, with several RSS feed icons scattered across it. A vertical label 'Lower Textual similarity' is on the left side of this area. At the bottom, there are tabs for 'Cluster' and 'Neighbours', and a version string 'Porpoise v. 0.1, 2008-01-20'.





# Constructing the similarity measure & neighbourhood (I)<sup>15</sup>



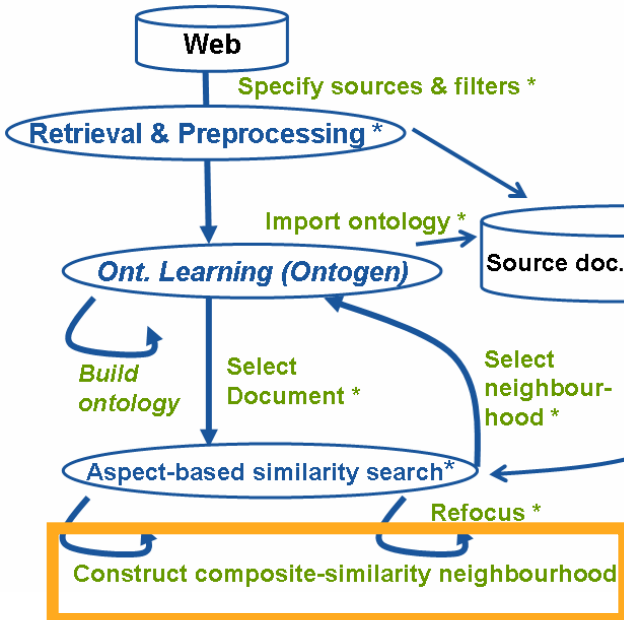
The screenshot shows the Porpoise software interface. The main window is titled 'Porpoise' and has a menu bar with 'File', 'Export', and 'Help'. Below the menu bar are icons for file operations. The interface is divided into two main sections: a left sidebar with settings and a main display area on the right.

The left sidebar contains three sections:

- Similarity Thresholds**:
  - Textual**: Current value is 0.1000. A red arrow points to the 'Adjust threshold' slider.
  - Named Entities**: Current value is 0.02000. It has an 'Adjust threshold' slider.
  - Date**: Current value is 0.5. It has an 'Adjust threshold' slider. Below the slider, it says 'Similarity 0.5 equals 4 days before and after the selected document'.

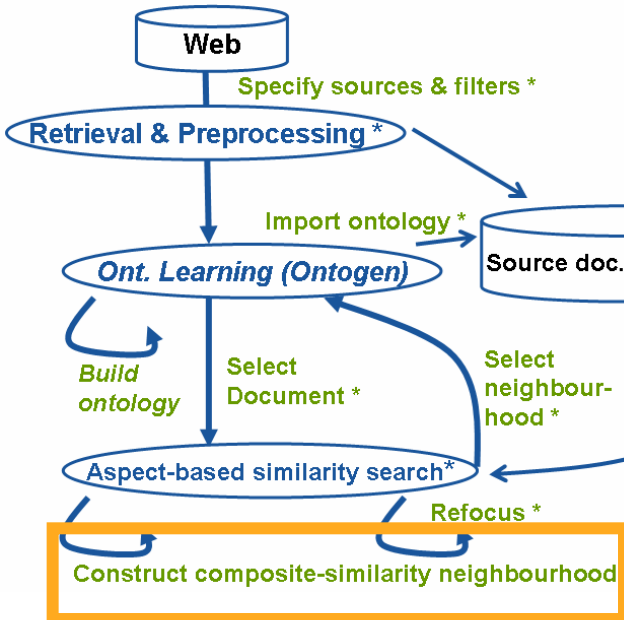
The main display area on the right is titled 'Lower Textual Similarity' and shows a network of nodes represented by RSS icons. At the bottom of the interface, there are tabs for 'Cluster' and 'Neighbours', with 'Neighbours' currently selected. The text 'Published Before' is visible at the bottom right.

# Constructing the similarity measure & neighbourhood (II)<sup>16</sup>



The screenshot shows the Porpoise software interface. The title bar reads 'Porpoise'. The menu bar includes 'File', 'Export', and 'Help'. The main window is divided into two panes. The left pane, titled 'Similarity Thresholds', contains three sections: 'Textual' with a 'Current' value of 0.02000 and an 'Adjust threshold' slider; 'Named Entities' with a 'Current' value of 0.02000 and an 'Adjust threshold' slider; and 'Date' with a 'Current' value of 0.5 and an 'Adjust threshold' slider. A red arrow points to the 'Named Entities' 'Current' value. The right pane, titled 'Lower Textual Similarity', displays a visualization of a network of nodes (represented by RSS icons) connected by lines. The bottom of the interface has a 'Cluster' button and a 'Neighbours' button.

# Constructing the similarity measure & neighbourhood (III)<sup>17</sup>



**Porpoise**  
File Export Help

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

Cluster Neighbours

Lower Textual Similarity

A German-language blog

A news source

Most neighbours are English-language blogs

Published Before

# Comparing documents

**Porpoise**  
File Export Help

Similarity Dimensions  
Textual

**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

□ □ Home&about 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

Close

**Document: Action: The Capitulation Caucus**

Name: feeds\_dailykos\_com\_Action\_The\_Capitulation\_Caucus  
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 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 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

**Porpoise**

File Export Help

Similarity Dimensions

Textual

**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

Home About 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

Close

**Document: Action: The Capitulation**

Name: feeds\_dailykos\_com\_Action\_The\_Capitulation

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 \_\_\_\_

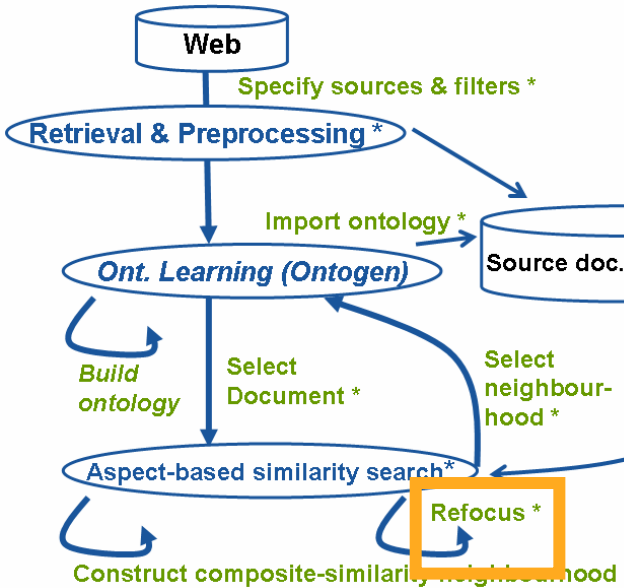
Tell Congress NOT to support this cap

the fold, the name and phone number o

Representative, and if you are unsure

Representative is, information on how

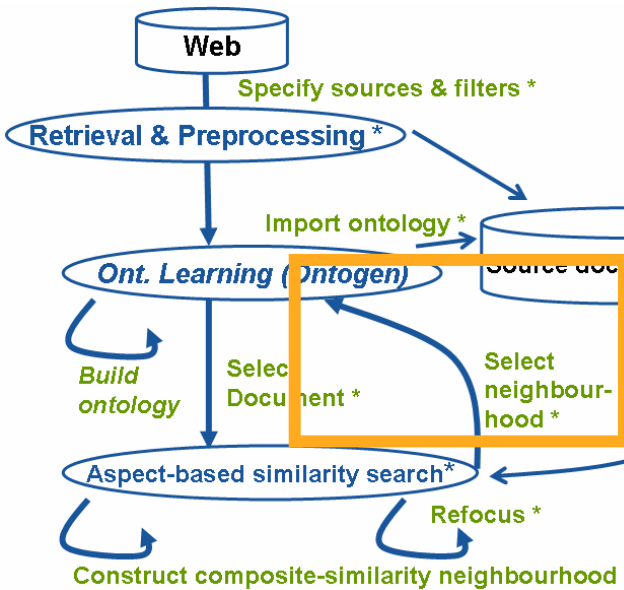
# Refocusing



The screenshot shows the Porpoise software interface. The title bar reads 'Porpoise'. The menu bar includes 'File', 'Export', and 'Help'. The main window is divided into two panes. The left pane, titled 'Similarity Dimensions', contains three sections: 'Textual', 'Named Entities', and 'Date'. Each section has a 'Current' value and 'Change similarity' buttons (-0.1, +0.1, -0.01, +0.01). The 'Textual' section has a current value of 0.02000. The 'Named Entities' section has a current value of 0.1200. The 'Date' section has a current value of 0.02000. The right pane, titled 'Lower Textual Similarity', displays a network graph with nodes represented by RSS icons. The bottom status bar shows 'Cluster' and 'Neighbours' tabs, with 'Neighbours' currently selected. The text 'Published Before' is visible at the bottom right of the main window.



# Structuring a neighbourhood



The screenshot shows the 'Porpois' software interface. The title bar includes 'File', 'Export', and 'Help' menus. The 'Export' menu is highlighted with a red rectangle. The main window is divided into two panes. The left pane, titled 'Similarity Dimensions', contains three sections: '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), and 'Date' (Current: 0.02000, Change similarity: -0.1, +0.1, -0.01, +0.01). The right pane, titled 'Lower Textual Similarity', displays a visualization of similarity data using blue RSS-like icons on a dark background. The bottom status bar shows 'Cluster' and 'Neighbours' tabs.

# Ex.: Finding a “story”

The screenshot displays the Porpoise software interface, which is used for analyzing document similarity and timelines. The main window is titled "Porpoise" and features a menu bar with "File", "Export", and "Help". Below the menu bar are three icons: a folder, a document, and a gear.

On the left side, there are three panels for adjusting similarity thresholds:

- Similarity Thresholds**
  - Textual**: Current threshold is 0.1000. An "Adjust threshold" slider is shown.
  - Named Entities**: Current threshold is 0.02000. An "Adjust threshold" slider is shown.
  - Date**: Current threshold is 0.5. An "Adjust threshold" slider is shown. Below the slider, it states: "Similarity 0.5 equals 4 days before and after the selected document".

The central area is a large black rectangle representing a timeline. It is labeled "Lower Textual Similarity" on the left. The timeline is divided into two sections: "Published Before" on the left and "Published After" on the right. Various document icons (RSS feeds and document thumbnails) are scattered across the timeline, representing documents found within the specified similarity and date ranges.

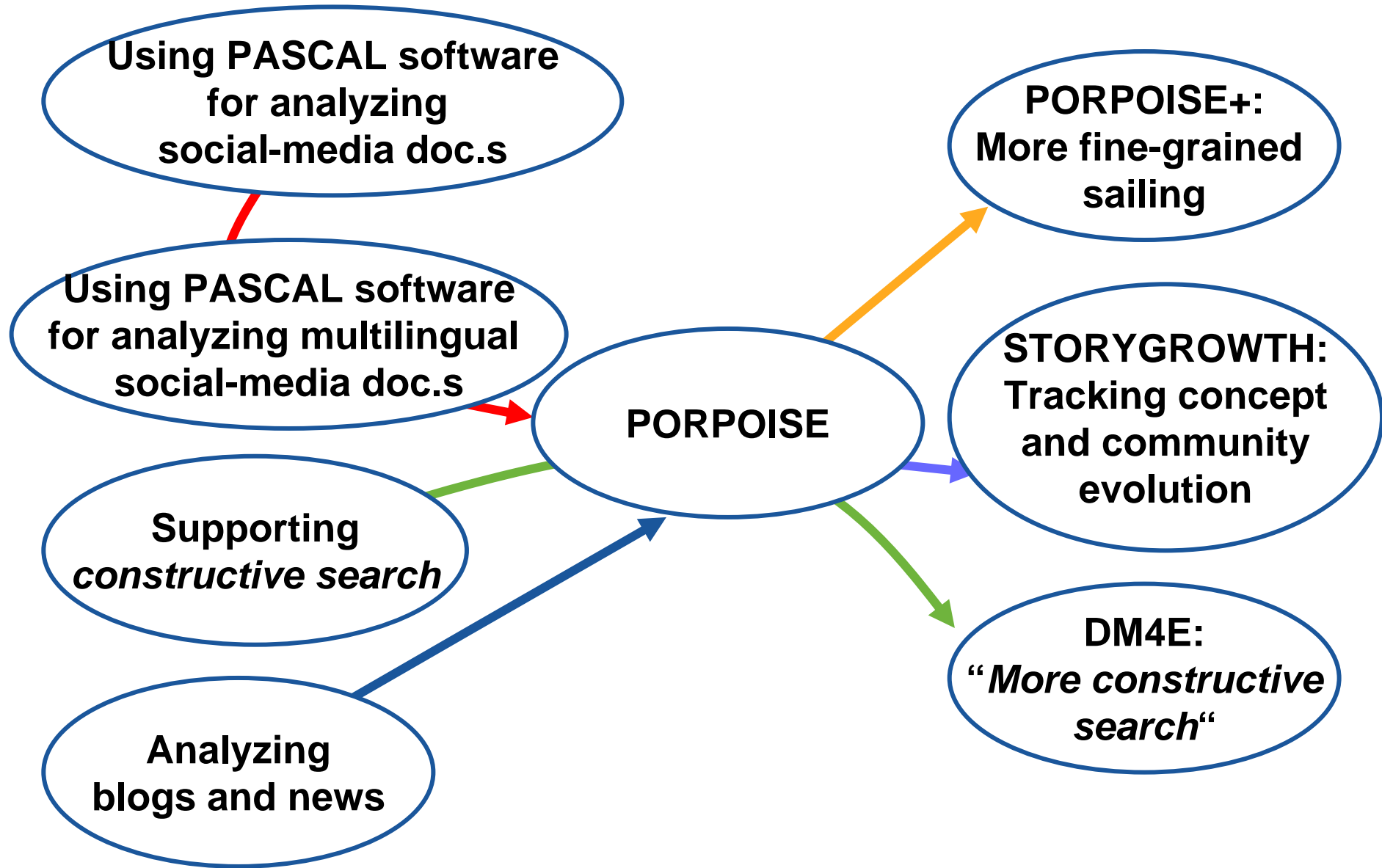
On the right side, there is a vertical list of text snippets, likely extracted from the documents, including words like "use", "ext", "be", "ves", "the", "enzi", "alls", and "nst". A "Close" button is located at the bottom right of this list.

At the bottom of the Porpoise window, there is a status bar with the text "Cluster Neighbours" and "Porpoise v. 0.1, 2008-01-20".

An orange box with white text is overlaid on the bottom right of the image, containing the text:

**Evaluation?  
User studies!**

## “Pump-priming”: PORPOISE as catalyst



# Finally ... could I express it better?

