Getting Started With Condor
Contents

- Getting Started
  - Collecting Web Content
  - OneDegreeCollector
- Building your own Startlists
- Collecting your E-Mail
- Collecting Facebook Data
- Collecting Wikipedia Data
- Collecting CoolPeople
- Coolhunting Blueprint
Getting Data into Condor

- IMAP E-Mail (Mailcollector) Eudora mailboxes
- Web/Blog/News/Scholar (WebCollector)
- Wikipedia (WikiFactFetcher)
- Snippets (OneDegreeCollector)
- Twitter (TwitterCollector)
- FlatFiles (FileLoader)
- PeopleNetworks (CoolPeople)
- Facebook
Temporal Visualization by a Sliding Time Frame

With history:

No history:

Time Frame size $n$

$1 \rightarrow n$

$2 \rightarrow n+1$

$3 \rightarrow n+2$

$4 \rightarrow n+3$

$5 \rightarrow n+4$
With and without history
Preparation

- Install MySQL
- Install Java (only Windows)
- Install Java 3D (only Windows)
- Start Java (if it does not run yet)
Contents

- Getting Started
- Collecting Web Content
- OneDegreeCollector
- Building your own Startlists
- Collecting your E-Mail
- Collecting Facebook Data
- Collecting Wikipedia Data
- Collecting CoolPeople
- Coolhunting Blueprint
Collect Web Content

Select relevant results for query barack obama (20 results in total):

- Collective Web Content
- One Degree Collector...
- Mail Collector...
- WikiFactFetcher...
- Twitter Collector...
- FacebookFriends...
- File Loader...
- Cool People...

Query
- Text Query
  - For exact phrase search, enclose queries with "..." (e.g. "coca cola", "virgin cola")
  - Separate multiple queries with commas (e.g.: coca cola, virgin cola)

URL Query
- Enter the websites you want to search (e.g.: http://www.cocacola.com, virgincola.com, pepsi.com/news.html)
  - Separate multiple queries with commas


Options
- Results of Top: 20
- Degree of Separation: 2
- Collect content (this takes more time)
- Preview results

No database connection, use the menu of this form to open or create one before starting the collector.
Communication View
Term view index

Term processing

- Use existing index (uncheck to create a new index)

Terms Selection Strategy
- Preview terms
- Upload list of terms

Indexing
- Include numbers (whether numbers should be indexed)
- Use Porter stemming (whether words should be reduced to their stem)
- Set constraints on document date

Stopwords
Stopwords are highly frequent words like "the" or 'for'. Stopwords are skipped during...
- use minimal stopwords list
- use default stopwords list
- use extended stopwords list
- use blog stopwords list
- use custom stopwords list

(one word or phrase per line, no commas, lowercase words)
Term view index - 2

This dialog lets you select which terms will be used for the term view.
Click the words or phrases from the left column to add it to the right selection.

Add Words
word (norm. frequency) (click to add)
- v6.00.2800.1165 (56.0)
- mimeoio (56.0)
- priority (56.0)
- microsoft (56.0)
- msmall (56.0)
- charset (32.0)
- iso-8859-1 (32.0)
- http (29.0)

Add 10 from the top

Add Phrases
phrase (norm. frequency) (click to add)
- microsoft before (56.0)
- microsoft very (56.0)
- msmall before (56.0)
- msmall very (56.0)
- charset very (32.0)
- charset before (32.0)
- http before (29.0)
- http very (29.0)

Add 10 from the top

Selected Terms
(click item to remove)
- produced (56.0)
- survey (36.0)
- thomas (31.0)
- mobile (30.0)
- deloitte (28.0)
- nico (28.0)
- clm (28.0)
- consulting (28.0)
- sent (26.0)
- schmalberger (26.0)
- web (26.0)
- please (26.0)
- sent before (26.0)
- sent very (26.0)
- web before (26.0)
- web very (26.0)
- before mailing (26.0)
- very mailing (26.0)
- 2002 (25.0)
- thurgauerstrasse (24.0)
- converted (24.0)
- pgloor (24.0)
- original (24.0)
- zurich (24.0)
- users (24.0)
- boundary (24.0)
- people (23.0)

- sort terms by norm. frequency
- sort terms by name
- Factor actor's weight in when computing communication weight
- Index dates (necessary for dynamic termview)
- Keep top % of cooc. 0.5

All your previous terms selections are automatically stored.
Use previous selections

Only Export selected terms

OK, collect communications ...
Contents

- Getting Started
- Collecting Web Content
- OneDegreeCollector
- Building your own Startlists
- Collecting your E-Mail
- Collecting Facebook Data
- Collecting Wikipedia Data
- Collecting CoolPeople
- Coolhunting Blueprint
One-Degree-Collector

- Complementary to the Blog Collector
- Fetches only one degree
- Retrieved websites are not aggregate
One-Degree-Collector - UI

- GUI resembles Blog Collector
One-Degree-Collector - result

- typical result of one-degree search
Contents

- Getting Started
- Collecting Web Content
- OneDegreeCollector
- Building your own Startlists
- Collecting your E-Mail
- Collecting Facebook Data
- Collecting Wikipedia Data
- Collecting CoolPeople
- Coolhunting Blueprint
Creating Term View Without OneDegreeCollector Start
List: Create Stoplist First

Add Words
word (norm. frequency) (click to add)
- deutschland (44.0)
- deutschen (41.0)
- allgemein (36.0)
- unternehmen (36.0)
- veröffentlicht (35.0)
- soll (35.0)
- euro (35.0)
- geld (25.0)

Add 10 from the top

Add Phrases
phrase (norm. frequency) (click to add)
- mit dem (38.0)
- und der (37.0)
- mit einem (34.0)
- auf dem (30.0)
- mit einer (29.0)
- nicht mehr (29.0)
- der deutschen (27.0)
- sie sich (26.0)

Add 10 from the top

Selected Terms
(click item to remove)
- jahrel (41.0)
- abornieren (41.0)
- neuer (41.0)
- ohm (41.0)
- heut (40.0)
- diesen (40.0)
- februar (40.0)
- medien (39.0)
- kontakt (39.0)
- august (38.0)
- denn (38.0)
- mich (38.0)
- mal (38.0)
- etwa (38.0)
- juli (37.0)
- diesem (37.0)
- sowi (37.0)
- hat (36.0)
- mail (36.0)
- die (36.0)
- sehr (36.0)
- kategorien (36.0)
- googl (36.0)
- dezemb (36.0)
- ent (36.0)
- machen (36.0)
- recht (36.0)

Only Export selected terms

Factor actor's weight in when computing communication w
Index dates
(necessary for dynamic termview)
Keep top % of cooc. 0.5

All your previous terms selections are automatically stored.
Use previous selections
OK, collect communications ...
... then use this stop list for the term view
Creating Term View With Start AND Stop List

- Use existing index (uncheck to create a new index)
- Terms Selection Strategy:
  - Preview term
  - Upload list of terms
  - elcat/ArtRelCat-2.08/bankingstratlist

- Indexing:
  - Include numbers (whether numbers should be indexed)
  - Use Porter stemming (whether words should be reduced to their stem)
  - Set constraints on document date

- Stopwords:
  - Stopwords are highly frequent words like "the" or "for". Stopwords are skipped.
  - Use minimal stopwords list
  - Use default stopwords list
  - Use extended stopwords list
  - Use blog stopwords list
  - elcat/ArtRelCat-2.08/bankingstoplist
    - Custom stopword list

- Edges:
  - Wikipedia
  - Anleihen
  - HSBC
  - Anwaltskanzlei
  - Spanien
  - EUR
  - Bank
  - Genf
  - Genossenschaft
  - Kunden
  - Asset Management
  - Privatbank
Contents

- Getting Started
- Collecting Web Content
- OneDegreeCollector
- Building your own Startlists
- Collecting your E-Mail
- Collecting Facebook Data
- Collecting Wikipedia Data
- Collecting CoolPeople
- Coolhunting Blueprint
Collect E-Mail

- `java -Xmx2048M -jar condor-2.1.jar`

![Image of GALAXYADVISORS login interface]

- **Condor Key**
- **MySQL password** (default: no password)

Please contact pgloor@mit.edu for key requests.
Tools to collect data
For username, host, port, and ssl check with your email provider (for gmail, see next slide)

**Left side:** enter here the specification of the mailbox

**Right side:** database related data, eg. username: root, no password

**Content:** yes will download the whole emails, w/o content only the sender, recipients and the subject line are downloaded

**Anonymize** will replace email addresses with random identifiers

**Here you can choose specific folders to download**

**Delete the present data in the database?**

---

**Galaxy Mail Collector**

Mailbox

- **Username:**
- **Password:**
- **Host:**
- **Folder:** ALL
- **Port:** 993

**Database**

- **Username:**
- **Password:**
- **Host:** localhost
- **Database Name:** DBMail_2010_06_30212723
- **DataSet Name:** myDs
- **Anonymize:**
- **Format:**
- **Clear database:**

---

[galaxyadvisors]
Settings for gmail

Username: yourname@gmail.com
Password: Your gmail password
Host: imap.gmail.com
Folder: ALL
Port: 993
Content: no
SSL: yes
Protocol: imap

Don't forget the access information for your mysql database on the right, then press start. It might take a while (esp. with huge mailboxes) before you see a progress bar.

Database
Username: 
Password: 
Host: localhost
Database Name: DBMail_2010_06_30212723
DataSet Name: myDs
Anonymize: no
Format: Condor
Clear database: no
Visualize Mail-Data

1. Open Database...
2. Enter Database Name: `information_schema`
3. Connected to database: ptermailDec09@localhost. Communication type is: For Emails.
4. Create Communication Views...
Visualize E-Mail Data (3)

Input

keep actors whose communication frequency >= your input. The range is [ 1 – 9100 ]?

OK

Cancel

Select top N nodes by betweenness

total nr of nodes: 311

group density of graph: 0.01550669

Select top N nodes by BC

top N nodes: 311 (min. 0, max. 311)

Extrapolate removed nodes over 2 degrees

Weight extrapolated edges with factor: 0.25

Extrapolate removed nodes over 3 degrees

Weight extrapolated edges with factor: 0.1
Dynamic View of Communication
Visualize E-Mail Contents
Visualize E-Mail Contents (2)

1. Dynamic Control
   - Time Unit: DAY
   - Time Window Size [1, 61]: 1
   - Options: With History, Fast Decay

2. Select top N nodes by betweenness
   - Total nr of nodes: 36
   - Group density of graph: 9.0
   - Top N nodes: 36 (min. 0, max. 36)
Dynamic View of Terms
Contents

- Getting Started
- Collecting Web Content
- OneDegreeCollector
- Building your own Startlists
- Collecting your E-Mail
- Collecting Facebook Data
- Collecting Wikipedia Data
- Collecting CoolPeople
- Coolhunting Blueprint
15.599 Workshop in IT: Collaborative Innovation Networks
Fall 2011

For information about citing these materials or our Terms of Use, visit: http://ocw.mit.edu/terms.