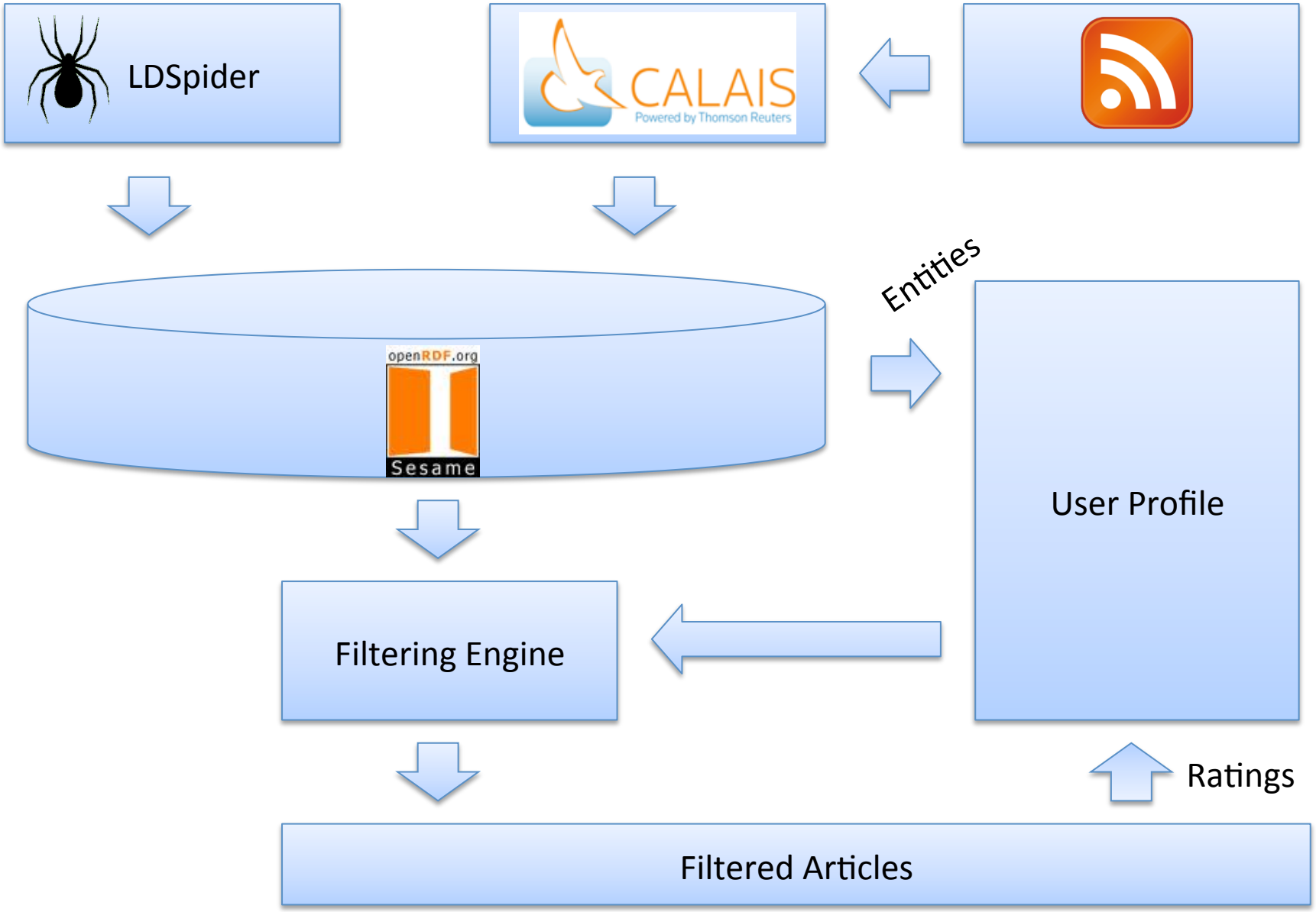
Named Entity Recognition for Enhanced Annotation

Ivo Lašek
Faculty of Information Technology
Czech Technical University in Prague
lasekivo@fit.cvut.cz

14.6.2012 – Knowledge Engineering Group Seminar

Reminder


LDSpider
CALAIS
Powered by Thomson Reuters
openRDF.org
Sesame
Entities
User Profile
Filtering Engine
Ratings
Filtered Articles

Crawling Problems

- Many irrelevant data sources (hundreds of thousands of foaf profiles of completely unimportant people)
- Focus on selected good managed data sources

Storage

- Sesame is extremely slow for certain queries
- Virtuoso performs better
- A custom storage for predefined type of data (e.g. Lucene index for fulltext search, relational databases, graph databases)

Czech DBpedia

- Machine readable data extracted form Wikipedia articles
- Available in RDF
- SPARQL endpoint to query data
- Framework for extraction provided by a community
- Set up and run to extract data from Czech Wikipedia

About: Jan Svěrák

An Entity of Type : Agent, from Named Graph : http://cs.dbpedia.org, within Data Space : cs.dbpedia.org

Property	Value
http://cs.dbpedia.org/property/id	1462 (xsd:integer) 21987 (xsd:integer) 841232 (xsd:integer)
http://cs.dbpedia.org/property/jm%C3%A9no	Jan Jan Svěrák
http://cs.dbpedia.org/property/m%C3%ADstoNarozen%C3%AD	Žatec, Československo
http://cs.dbpedia.org/property/obr%C3%A1zek	Jan Sverak.jpg
http://cs.dbpedia.org/property/p%C5%99%C3%ADjmen%C3%AD	Svěrák
http://cs.dbpedia.org/property/popisek	Jan Svěrák
http://cs.dbpedia.org/property/popisekObr%C3%A1zku	Jan Svěrák
http://cs.dbpedia.org/property/wikiPageUsesTemplate	dbpedia:Šablona:Imdb_osoba dbpedia:Šablona:Čsfd_osoba dbpedia:Šablona:Čfn_osoba dbpedia:Šablona:Infobox_Biografie
dbpedia:Šablona:Čfn_osoba	jméno id
dbpedia:Šablona:Čsfd_osoba	jméno id
dbpedia:Šablona:Imdb_osoba	jméno id
dbpedia:Šablona:Infobox_Biografie	jméno příjmení obrázek místo narození popisek obrázku popisek

Enhanced Annotation

Goldman Sachs Rises as Investors Bet on Comeback

Goldman Sachs Group Inc. (GS) rose 5.5 percent in New York trad as investors looked past a third-q g revenue a riting and ta

dbpedia-owl:industry	dbpedia:Financial_services
dbpprop:locationCity	New York City
fb:organization.organization.date_founded	1869

yovisto
Academic video search

Sign Up!

Username:
Password:
Remember me. login
Forgot your login
Sign up a new account

Home | Search | MyYovisto | Universities | Lectures | Videos | Speakers | Upload | FAQ | Imprint | Blog | Feedback

standard search | advanced search

1 15. Freud on Sexuality and Civilization

Foundations of Modern Social Theory

Ivan Szelenyi

[video:19745] 0 Views Fall 2009

(Score:4.5684586) Duration: 00:53:28

2 03. Foundations: Freud

Introduction to Psychology

Paul Bloom

siw

[video:10183] 13 Views Spring, 2007

(Score:4.542417) Duration: 00:56:30

3 Spiel, Freude, Eierkuchen?

Time: 419ms
Results: 18

SEARCH

Freud | Odeslat

FILTER:

Category: -- not set --

Organisation: -- not set --

Language: -- not set --

Tag: clear

Popular tags in results

art (1)
birds (1)
changing (1)
competition (1)
cow (1)
create (1)
cross (1)
culture (1)
deli (1)
detail (1)
drawing (1)
dust (1)
education (1)
exam (1)
face (1)
fenster (1)
folded (1)
footprint (1)
foundation (1)
frame (1)
free (1)
freud (1)
god (1)
gravity (1)
in (1)
india (1)
insecurity (1)
king (1)
lacan (1)
le (1)

GEO information

ARCHIVAL content

WEB content

Named Entity Disambiguation

Context Representation

- Using a knowledge base
  - Very often Wikipedia
- Bag of words representation
- Structural representation (i.e. link analysis)

Bag of Words Representation

- Identify candidate entities
  - Spotlight – Aho-Corasick Algorithm
  - Some ready made NER tool (e.g. Stanford parser)
- **Context representation:** Collect paragraphs from Wikipedia, where a given entity is mentioned
- Compare the context with the target text
- Cosine similarity
- TF-ICF measure for weighting terms

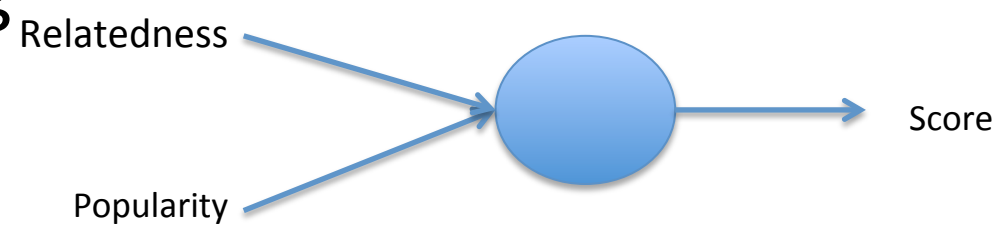
Structural Representation

- Entity is represented by a corresponding Wikipedia article
- **Context representation:** Compare an entity to other (disambiguated) entities in the same article
- Select the most similar (related) entity to non ambiguous entities in the text

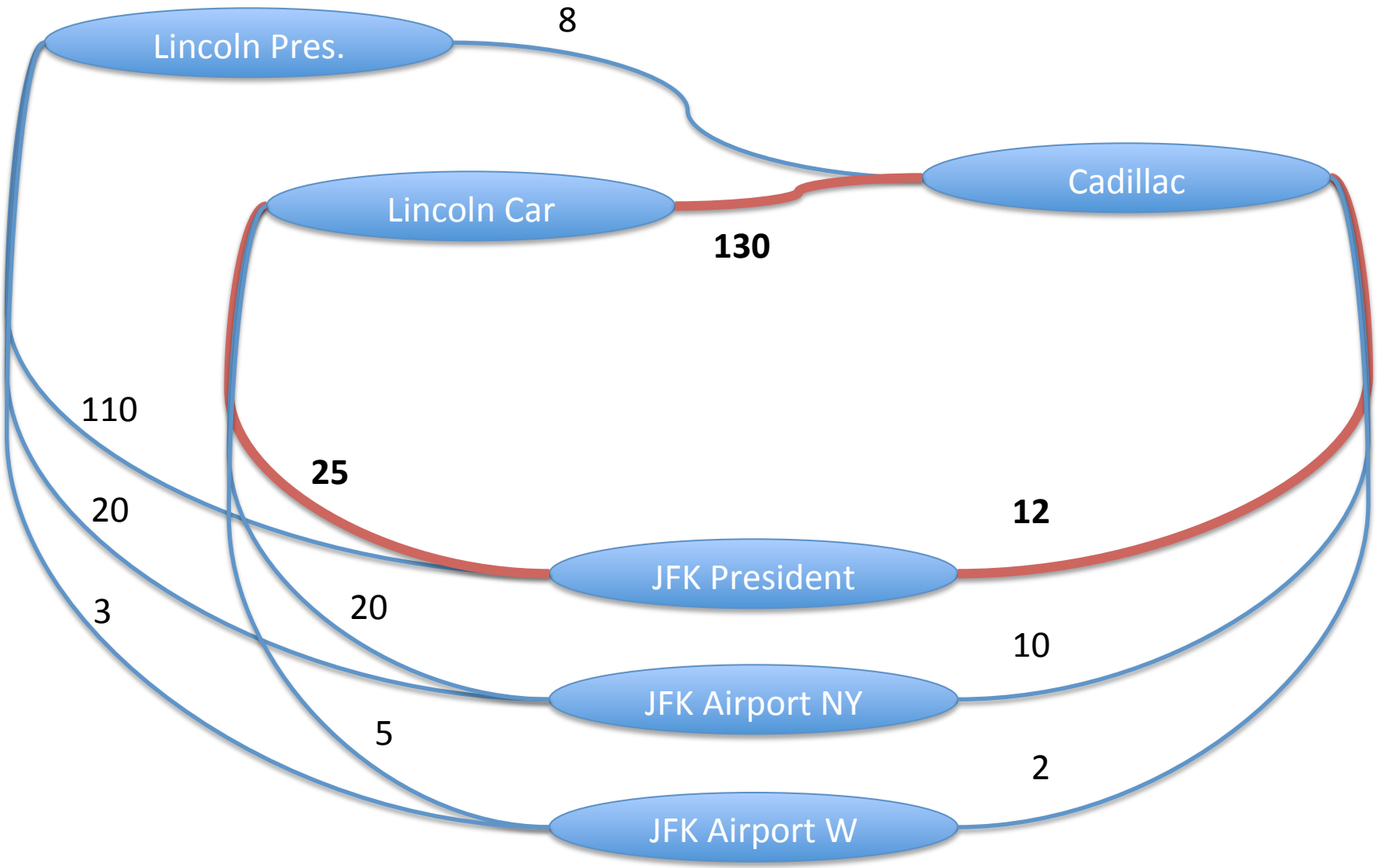
$$relatedness(a,b) = \frac{\log(\max(|A|,|B|)) - \log(|A \cap B|)}{\log(|W|) - \log(\min(|A|,|B|))}$$

Overall Entity Popularity

- An important measure is overall entity popularity
- On Wikipedia, this is proportional to number of links linking to a corresponding Article
- **Context representation:** No context needed – apriori measure
- Milne and Witten: Proportion between relatedness and popularity is tuned up using machine learning



Our Approach – Co-Occurrence Measure



Future Work

Future Work

- Combine described approaches
- Optimize performance
- Patterns identification for recognition of unknown entities and their properties