WIRTSCHAFTS
UNIVERSITÄT
WIEN VIENNA
UNIVERSITY OF
ECONOMICS
AND BUSINESS

Ontology Learning based on Text Mining and Social Evidence Sources

Albert Weichselbraun

February 9, 2011

Agenda

- Background and Motivation
- Extracting Evidences from Social Sources
- Evidence Integration
  - System Diagram
  - Example Data
  - Spreading Activation - Weights
- Evaluation
  - Setting
  - Informal Evaluation
  - Formal Evaluation
- Outlook and Conclusions

Background and Motivation

- starting point: ontology learning framework (lightweight ontologies [Hendler, 2009, Alani et al., 2008])
- based on a seed ontology and domain documents
  - extract relevant terms
  - integrate them into the ontology
- benefits of integrating social sources
  - potential of providing background knowledge
  - contain the latest terminology [Angeletou et al., 2007] (evolve at much a higher pace as domain documents)

Ontology Learning Cycle

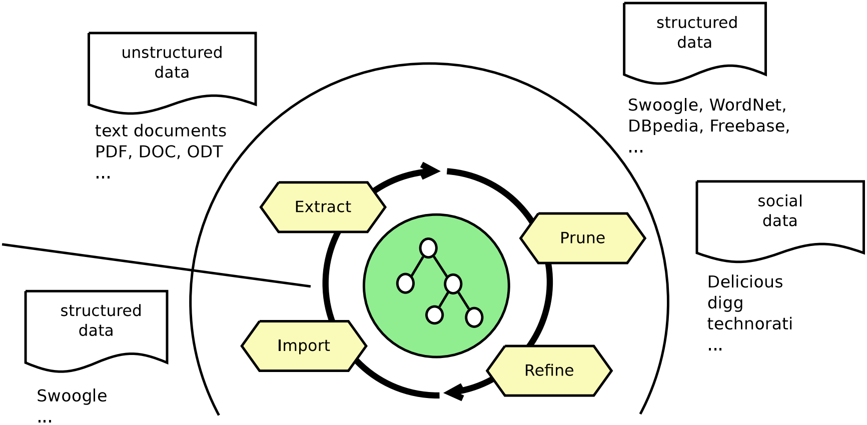


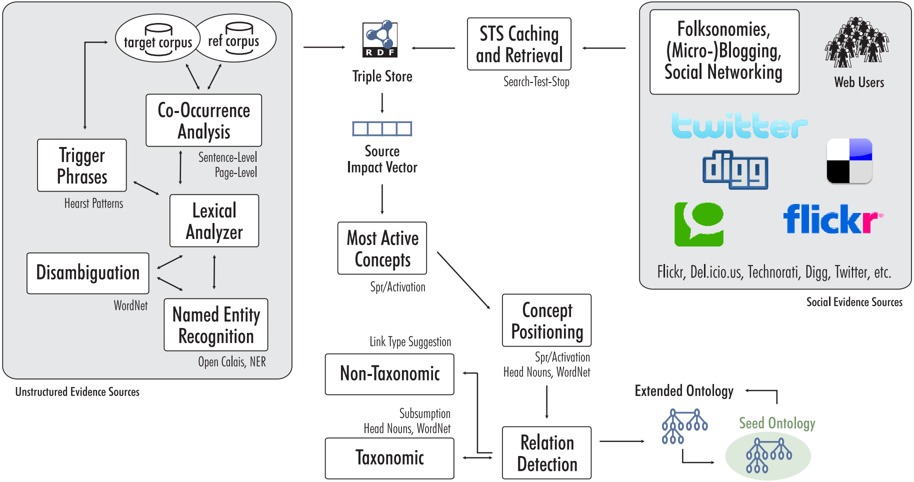
Figure: adapted from [Maedche and Staab, 2004]

Extracting Evidences from Social Sources

- based on the seed terms → transformation function (t) → source specific (e.g. monograms for Delicious)
- disambiguation: WordNet
- social evidence sources:
  - easy Web Retrieval Toolkit (www.semanticlab.net/eWRT)
  - TagInfoService
  - implemented for Delicious (social bookmarking), flickr (photo/video hosting), technorati (blogs) and twitter (micro blogging)
- suggested tags → relation weights based on the Dice coefficient

$$s_d(T_s, T_c) = \frac{2 \cdot n_{T_{sc}}}{n_{T_s} + n_{T_c}} \tag{1}$$

System Diagram



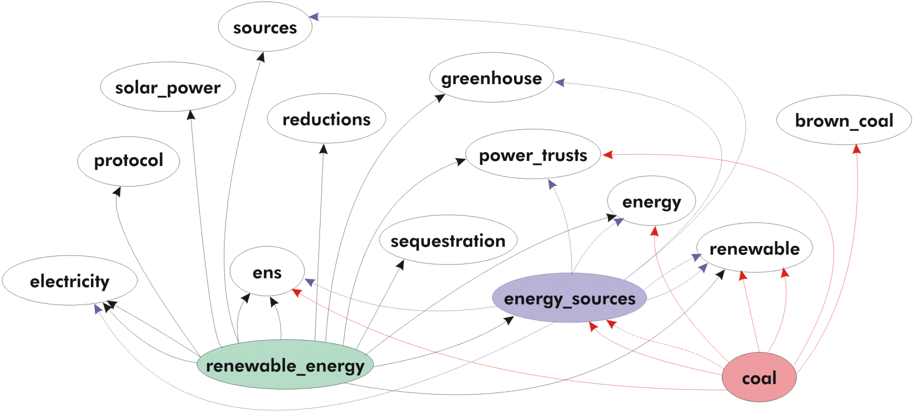
Suggested Terms

unstructured	**social**	
	delicious	flickr
targets	animalcare	architecture
building	architects	art
coal	atmosphere	auckland
levels	award	beach
climate change policy	britney	bicycle
pact	carbonfootprint	brian
reduce greenhouse gas	...	...
pollution		
firm		
carbon dioxide emissions	technorati	twitter
ets	agile	aces
its carbon	apple	afghan
	architecture	afghanistan
	art	africa
	automotive	al_gore
	...	...

Triple Store

Seed Concept(C_s)	Evidence Source (e)	Candidate Concept(C_v)
climate_change	oe:coOccurs	co2
_:1	rdf:subject	climate_change
_:1	rdf:predicate	oe:coOccurs
_:1	rdf:object	co2
_:1	rdf:type	rdf:Statement
_:1	oe:significance	”3.20”
climate_change	oe:twitterTag	co2
_:2	rdf:subject	climate_change
_:2	rdf:predicate	oe:twitterTag
_:2	oe:dice	”1.59”
	...	
climate_change	wn:hypernym	temperature_change

Spreading Activation



Spreading Activation

Goal: select the most promising candidate terms

Result from the previous process:

- evidence vector $\vec{r}$ → contains evidence sources e:

$$\vec{r}(C_s, C_c) = \begin{pmatrix} r_{e_1}(C_s, C_c) \\ ... \\ r_{e_n}(C_s, C_c) \end{pmatrix} \tag{2}$$

Transforming Evidences to Spreading Activation Weights

- Heuristic per-evidence-source translation rules s_e transform these relations using the source impact vector $\vec{S} = (s_{e_1}, s_{e_2}, ...s_{e_n})^T$ into a numerical weight

$$w(C_s, C_c) = |\vec{S}(\vec{r}(C_s, C_c))| \tag{3}$$

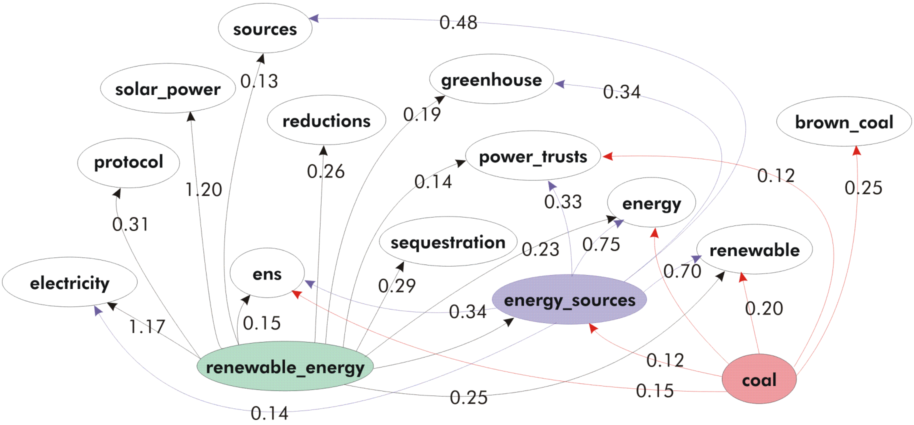
Spreading Activation - Example

$$\vec{r}(\text{cc,fuel}) = \begin{pmatrix} (oe : coOccurs, sign = 3.2) \\ (oe : deliciousTag, dice = 1.59) \\ (oe : triggerPhrase) \end{pmatrix}$$

$$\vec{S} = \begin{pmatrix} 0.1 + 0.5 \cdot sign \\ 0.2 \cdot dice \\ 0.3 \end{pmatrix}$$

$\rightarrow$ weight $w(\text{cc, fuel}) = 2.318$.

Spreading Activation - Example



Evaluation - Setting

- seed “ontology”:
  - *fossil fuels* $\xrightarrow{relatedTo}$ *climate change* and
  - *fossil fuels* $\xrightarrow{relatedTo}$ *greenhouse gas(es)*
- domain corpora
  - 156 news media sites from the Newslink.org, Kidon.com and ABYZNewsLinks.com directories → 200,000 documents per week
  - six monthly corpora (April 2009 - August 2009)
  - domain detection based on regular expressions → climate change corpus containing 1250 documents / month
- social sources
  - Delicious, flickr, technorati, twitter
- two iterations → 24 new terms

Evaluation - Terms Removed and Added

terms removed	terms added
carbon dioxide emissions	agw
climate change policy	biomass
developing nations	cprs
kyoto protocol	cars
scientific assessments	epa
sulfur dioxide	ethanol
tom magliozzi	greenhouse-gas

Evaluation - Method

- pointwise mutual information (PMI)
  → how well are terms associated to each other
- four domain experts
  → relevance of the given relation
  → (0 .. irrelevant, 1 slightly relevant, 2 ...very relevant)

Evaluation - Pointwise Mutual Information

- Web metric (Yahoo! counts): seed tag counts (n_{T_s}), candidate tag (n_{T_c}) counts, common counts ($n_{T_{sc}}$)

$$
\begin{aligned}
n_z &= n_{T_{sc}} + n_{T_s} + n_{T_c} && (4)\\
f(i) &= \frac{n_i}{n_z} e^{-\frac{ni}{n_z}} && (5)\\
PMI(T_s, T_c) &= f(n_{T_{sc}})/f(n_{T_s}) \cdot f(n_{T_c}) && (6)
\end{aligned}
$$

Impact of social evidence sources

avg. PMI	corpus-based	corpus-based & social
April 2009	0.694 (16)	0.833 (17)
May 2009	0.753 (15)	0.921 (10)
June 2009	0.569 (16)	0.544 (15)
July 2009	0.625 (8)	0.862 (8)
August 2009	0.493 (5)	0.874 (9)
Sum	0.503 (60)	0.646 (59)

expert eval.	corpus-based	corpus-based & social
April 2009	0.875 (16)	1.353 (17)
May 2009	0.883 (15)	1.550 (10)
June 2009	1.000 (16)	1.283 (15)
July 2009	1.469 (8)	1.563 (8)
August 2009	1.150 (5)	1.167 (9)
Sum	1.013 (60)	1.369 (59)

Outlook and Conclusions

- including social sources provides significant improvements to the ontology extension process (99.9% for a Welch two sample t-test and for the Wilcoxon rank sum test)
- drawbacks and potential pitfalls:
  - many social sources yield only unigrams
  - balancing corpus-based and social sources
- Future work:
  - support for n-grams
  - optimize source impact vectors based on user feedback
  - optimize access to remote resources (optimal stopping)

Alani, H., Hall, W., O'Hara, K., Shadbolt, N., Chandler, P., and Szomszor, M. (2008).
Building a pragmatic semantic web.
IEEE Intelligent Systems, 23(3):61–68.

Angeletou, S., Sabou, M., Specia, L., and Motta, E. (2007).
Bridging the gap between folksonomies and the semantic web: An experience report.
In *Workshop: Bridging the Gap between Semantic Web and Web*, volume 2.

Hendler, J. (2009).
Web 3.0 emerging.
Computer, 42(1):111–113.

Maedche, A. and Staab, S. (2004).
Handbook on ontologies.
In Staab, S. and Studer, R., editors, *Handbook on Ontologies*, International Handbooks on Information Systems, chapter Ontology Learning, pages 173–190. Springer.