

Extracting Semantics from Crowds

International Summer School on Semantic Computing (SSSC'2011)
August 8 - 12, 2011, Berkeley, California

Markus Strohmaier

Assistant Professor

Graz University of Technology, Austria

Visiting Scientist

(XEROX) PARC, USA

with **collaborators** D. Helic, C. Körner, J. Pöschko, R. Kern, C. Trattner, H.P. Grahsl C. Wagner (Graz University of Technology, Austria), D. Benz, G. Stumme (U. of Kassel, Germany), A. Hotho (U. of Würzburg, Germany), S. Hellmann, J. Lehmann, C. Stadler, J. Unbehauen (U. of Leipzig, Germany), L. Hong (Parc, USA) and A. Zubiaga (UNED, Spain)

Egypt 2011 Twitter Trends

27.Jän	28.Jän	29.Jän	30.Jän	31.Jän	01.Feb	02.Feb	03.Feb	04.Feb
#jan25	#jan25	#jan25	#jan25	#jan25	#jan25	pro-mubarak	#tahrir	#tahrir
████████	████████	████████	tahrir	tahrir	tahrir	#jan25	#jan25	#jan25
████████	████████	████████	████████	protesters	#tahrir	#tahrir	tahrir	tahrir
biden	is ██████████	looters	jazeera	████████	square	tahrir	journalists	square
is ██████████	fine	looting	████████	army	#feb1	thugs	pro-mubarak	protesters
fine	curfew	protesters	square	isp	protester	on	thugs	departure
protests	████████	suleiman	protesters	jazeera	re-electio	supporters	abc	pipeline
blackout	████████	jazeera	clinton	noor	arabiya	square	suleiman	loading
internet	protesters	vp	fox	████████	obama	protesters	square	pro-mubarak
shuts	toto	tanks	curfew	#tahrir	transition	gunfire	protesters	reporters
#censorship	jazeera	army	army	offline	pro-mubarak	molotov	anderson	thugs
████████	#jan28	curfew	shuts	journalists	#mubarak	demonstrators	cooper	sq
████████	protests	#jan28	#aljazeera	protests	████████	punched	vp	#scariestwordsev
sms	army	fine	fighter	revolution	speech	clashes	kenneth	kenneth
blocked	tear	is ██████████	jets	military	crowds	army	detained	reporter
#jan28	protestors	protests	planes	evacuate	sq	pro	supporters	vp
numerous	biden	thugs	transition	voicemail	revolution	sq	cole	transition
access	gibbs	████████	looters	provider	jazeera	attacked	army	#prayforjustin
#suez	internet	tahrir	protests	unrest	step	cocktails	reporter	revolution
protesters	#censorship	vice	#tahrir	square	conan	roughed	attacked	anderson
shut	blackout	████████	map	internet	looters			

place

event

person


Work with Lichan Hong (then at Parc, now at Google)

Semantic Structures: The DMOZ Project


dmoz open directory project
 In partnership with
AOL search

[about dmoz](#) | [dmoz blog](#) | [suggest URL](#) | [help](#) | [link](#) | [editor login](#)

Search
[advanced](#)

<p><u>Arts</u> Movies, Television, Music...</p>	<p><u>Business</u> Jobs, Real Estate, Investing...</p>	<p><u>Computers</u> Internet, Software, Hardware...</p>
<p><u>Games</u> Video Games, RPGs, Gambling...</p>	<p><u>Health</u> Fitness, Medicine, Alternative...</p>	<p><u>Home</u> Family, Consumers, Cooking...</p>
<p><u>Kids and Teens</u> Arts, School Time, Teen Life...</p>	<p><u>News</u> Media, Newspapers, Weather...</p>	<p><u>Recreation</u> Travel, Food, Outdoors, Humor...</p>
<p><u>Reference</u> Maps, Education, Libraries...</p>	<p><u>Regional</u> US, Canada, UK, Europe...</p>	<p><u>Science</u> Biology, Psychology, Physics...</p>
<p><u>Shopping</u> Clothing, Food, Gifts...</p>	<p><u>Society</u> People, Religion, Issues...</p>	<p><u>Sports</u> Baseball, Soccer, Basketball...</p>
<p><u>World</u> Català, Dansk, Deutsch, Español, Français, Italiano,  Nederlands, Polski, Русский, Svenska...</p>		

Become an Editor
Help build the largest human-edited directory of the web



Copyright © 1998-2008 Netscape

4,607,963 sites - 81,339 editors - over 590,000 categories

Classification Systems in Information and Library Sciences

Overview of the D
The ten main classes

- 000 Com
- 100 Philo
- 200 Relig
- 300 Social sciences
- 400 Language
- 500 Science
- 600 Technology
- 700 Arts & recreation
- 800 Literature
- 900 History & g

Usually produced and maintained by few
(e.g. dozens of) domain experts.

but: used by many (potentially millions).

System (1998)

- [A.1](#) INTRODUCTORY AND SURVEY
- [A.2](#) REFERENCE (e.g., dictionaries, encyclopedias, glossaries)
- [A.m](#) MISCELLANEOUS
- [B.](#) Hardware
- [B.0](#) GENERAL

Can a very large group (a crowd) of *users*
contribute to ontology engineering efforts?

CROPPROGRAMMING ([D.3.2](#))

- *Hardwired control* [***]
- *Microprogrammed logic arrays* [***]
- *Writable control store* [***]
- [B.1.2](#) Control Structure Performance Analysis and Design Aids

Objectives

Provide (some) answers to the following questions:

- **What** is the difference between extracting semantics from text vs. extracting semantics from crowds?
- **Why** should we study crowds and crowd behavior from a semantic computing perspective?
- **How** can semantics be extracted from online crowd behavior, such as
 - ...from Social Labeling
 - ...from Social Tagging
 - ...from Social Navigation
- **What** are the implications for semantic computing research?

Extracting Semantics from ...

Motivation

Social Labeling

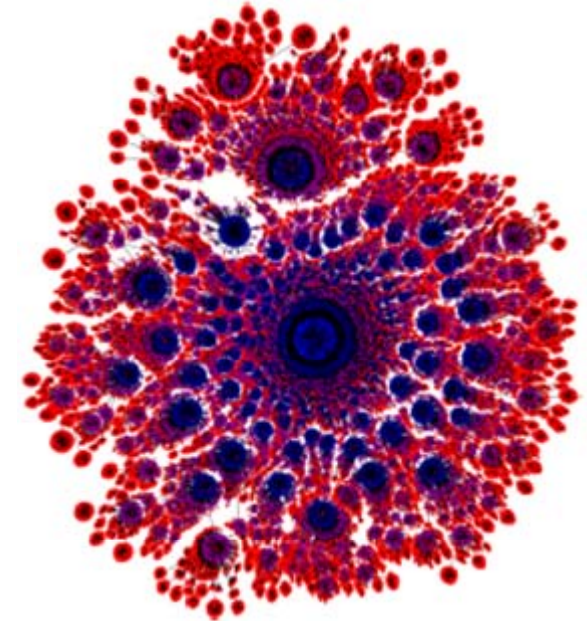
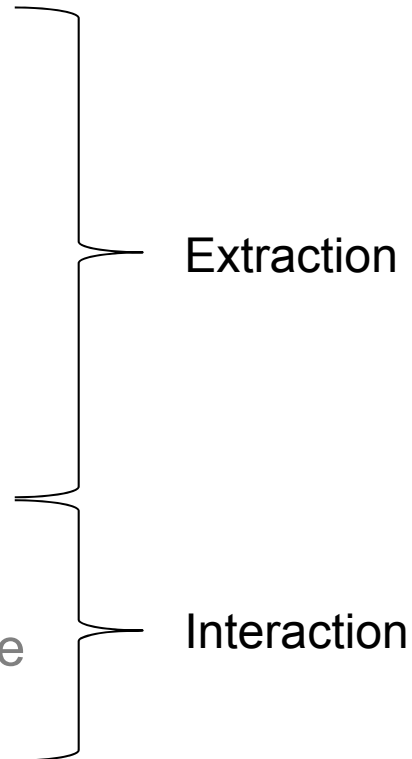
- Hashtag Semantics

Social Tagging

- Tag Relatedness
- Tag Generality
- Tag Hierarchies

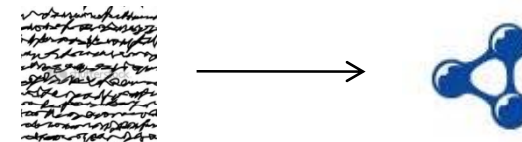
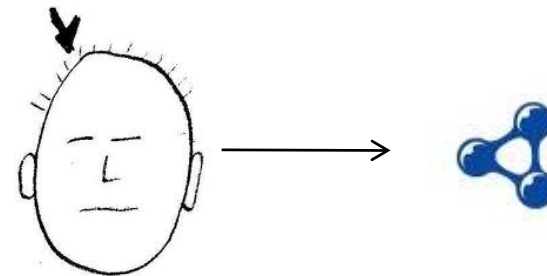
Social Navigation

- Navigational Knowledge Engineering



Prepositional vs. Distributional Semantics

- **Ontology Engineering**
 - Expert-driven, small-scale
 - Knowledge and concept identification
 - Preliminary informal representation
 - Formalization (RDF, OWL, etc)
 - Evaluation and Maintenance
- **Ontology learning**
 - Data-driven, large-scale
 - Source selection and data sampling
 - Data exploration and probing
 - Concept and link learning
 - Evaluation and Updating



Distributional Semantics

[Hovy 2011]

- In recent years, people in NLP and IR have started using a different representation for all kinds of problems: word distributions

$$\{ T_k, (w_{k1}, s_{k1}), (w_{k2}, s_{k2}), \dots, (w_{kn}, s_{kn}) \}$$

$$bank_1 = \{(bank\ 0.9), (thrift\ 0.11), (banking\ 0.4), (loan\ 0.4), (deposit\ 0.1), (money\ 0.7)\dots\}$$

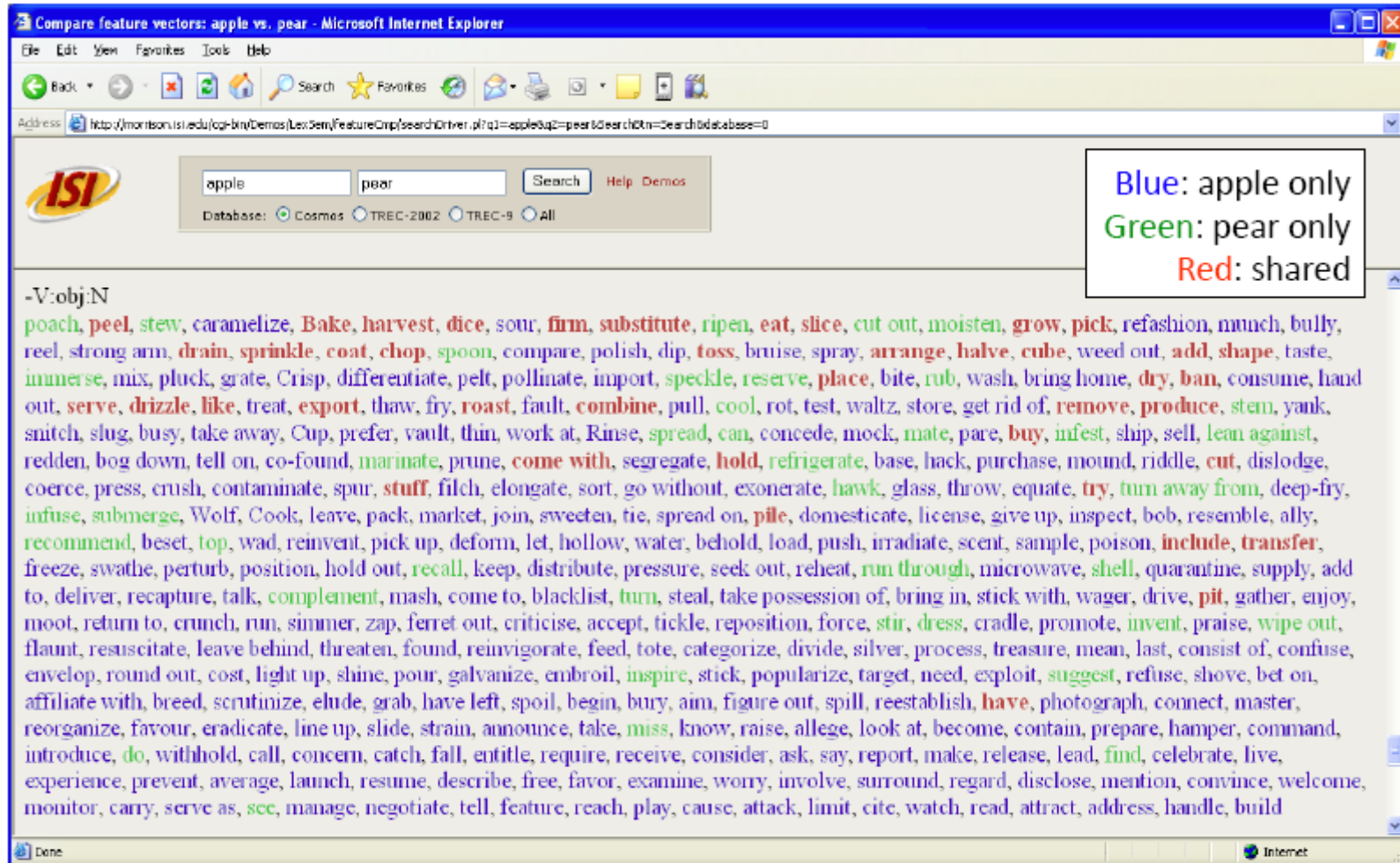
$$bank_2 = \{(bank\ 0.9), (turn\ 0.3), (veer\ 0.1), (lean\ 0.4)\dots\}$$

[Harris 1954]: “words that occur in the same contexts tend to have similar meanings“

- Statistical NLP operates at word level, frequency distributions are used as the (de facto) semantics of a word
- Not actual semantics, but captures something of contents
- Not compositional: how to ‘add’ two distributions?
- No explicit theory of semantics

Why apple is similar to pear

[based on slides by Hovy 2011 / Pantel 02]



The screenshot shows a web browser window titled "Compare feature vectors: apple vs. pear - Microsoft Internet Explorer". The address bar contains the URL: `http://montson.isi.edu/cgi-bin/Demos/Lex/Dem/FeatureCmp/SearchDriver.pl?q1=apple&q2=pear&searchDB=SearchDatabase=0`. The page features the ISI logo and a search interface with input fields for "apple" and "pear", a "Search" button, and a "Help Demos" link. Below the search fields, there are radio buttons for "Database: Cosmos", "TREC-2002", "TREC-9", and "All".

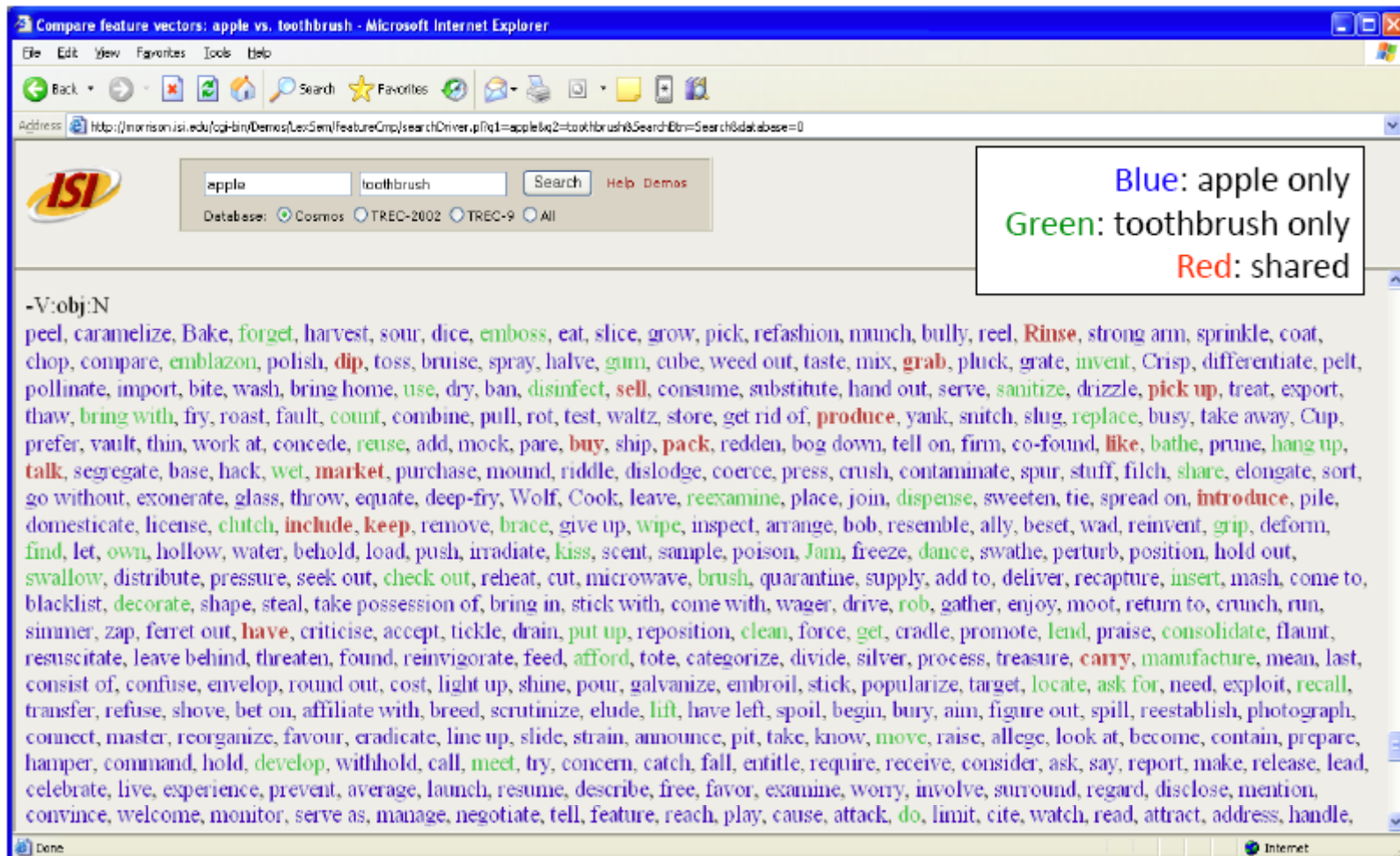
A legend box on the right side of the page defines the color coding for the word lists:

- Blue: apple only
- Green: pear only
- Red: shared

The main content area displays a list of words under the heading "-V:obj:N". The words are color-coded according to the legend: blue for words unique to 'apple', green for words unique to 'pear', and red for words shared between both. The list includes words such as poach, peel, stew, caramelize, Bake, harvest, dice, sour, firm, substitute, ripen, eat, slice, cut out, moisten, grow, pick, refashion, munch, bully, reel, strong arm, drain, sprinkle, coat, chop, spoon, compare, polish, dip, toss, bruise, spray, arrange, halve, cube, weed out, add, shape, taste, immerse, mix, pluck, grate, Crisp, differentiate, pelt, pollinate, import, speckle, reserve, place, bite, rub, wash, bring home, dry, ban, consume, hand out, serve, drizzle, like, treat, export, thaw, fry, roast, fault, combine, pull, cool, rot, test, waltz, store, get rid of, remove, produce, stem, yank, snitch, slug, busy, take away, Cup, prefer, vault, thin, work at, Rinse, spread, can, concede, mock, mate, pare, buy, infest, ship, sell, lean against, redden, bog down, tell on, co-found, marinate, prune, come with, segregate, hold, refrigerate, base, hack, purchase, mound, riddle, cut, dislodge, coerce, press, crush, contaminate, spur, stuff, filch, elongate, sort, go without, exonerate, hawk, glass, throw, equate, try, turn away from, deep-fry, infuse, submerge, Wolf, Cook, leave, pack, market, join, sweeten, tie, spread on, pile, domesticate, license, give up, inspect, bob, resemble, ally, recommend, beset, top, wad, reinvent, pick up, deform, let, hollow, water, behold, load, push, irradiate, scent, sample, poison, include, transfer, freeze, swathe, perturb, position, hold out, recall, keep, distribute, pressure, seek out, reheat, run through, microwave, shell, quarantine, supply, add to, deliver, recapture, talk, complement, mash, come to, blacklist, turn, steal, take possession of, bring in, stick with, wager, drive, pit, gather, enjoy, moot, return to, crunch, run, simmer, zap, ferret out, criticise, accept, tickle, reposition, force, stir, dress, cradle, promote, invent, praise, wipe out, flaunt, resuscitate, leave behind, threaten, found, reinvigorate, feed, tote, categorize, divide, silver, process, treasure, mean, last, consist of, confuse, envelop, round out, cost, light up, shine, pour, galvanize, embroil, inspire, stick, popularize, target, need, exploit, suggest, refuse, shove, bet on, affiliate with, breed, scrutinize, elude, grab, have left, spoil, begin, bury, aim, figure out, spill, reestablish, have, photograph, connect, master, reorganize, favour, eradicate, line up, slide, strain, announce, take, miss, know, raise, allege, look at, become, contain, prepare, hamper, command, introduce, do, withhold, call, concern, catch, fall, entitle, require, receive, consider, ask, say, report, make, release, lead, find, celebrate, live, experience, prevent, average, launch, resume, describe, free, favor, examine, worry, involve, surround, regard, disclose, mention, convince, welcome, monitor, carry, serve as, see, manage, negotiate, tell, feature, reach, play, cause, attack, limit, cite, watch, read, attract, address, handle, build.

Why apple is not similar to toothbrush

[based on slides by Hovy 2011 / Pantel 02]



Compare feature vectors: apple vs. toothbrush - Microsoft Internet Explorer

Address: <http://morison.isi.edu/cgi-bin/Demos/Lev5em/FeatureCmp/searchDriver.pl?q1=apple&q2=toothbrush&SearchEbn=Search&database=0>

ISI

apple toothbrush Search Help Demos

Database: Cosmos TREC-2002 TREC-9 All

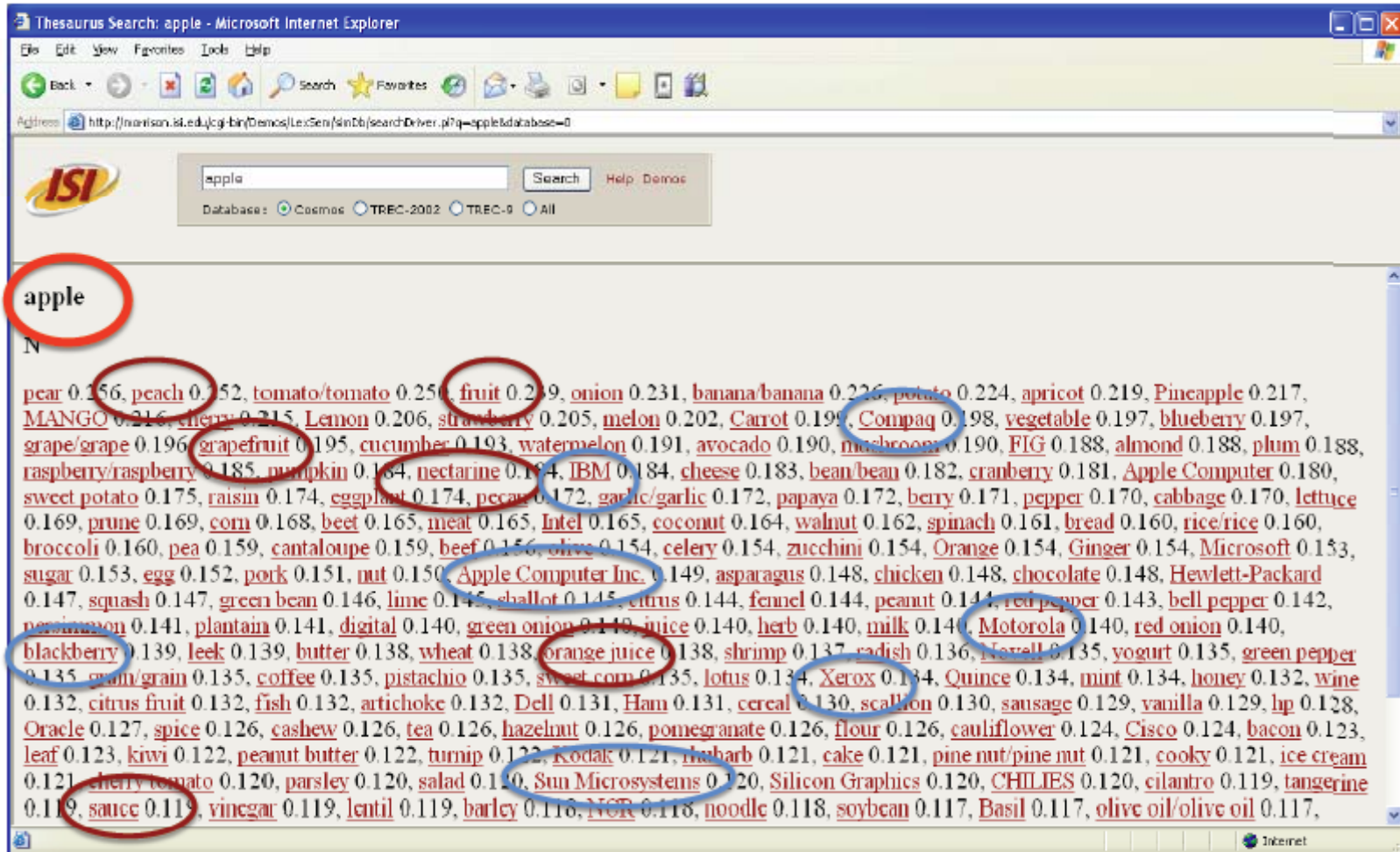
Blue: apple only
Green: toothbrush only
Red: shared

-V:obj:N

peel, caramelize, Bake, forget, harvest, sour, dice, emboss, eat, slice, grow, pick, refashion, munch, bully, reel, Rinse, strong arm, sprinkle, coat, chop, compare, emblazon, polish, dip, toss, bruise, spray, halve, gum, cube, weed out, taste, mix, grab, pluck, grate, invent, Crisp, differentiate, pelt, pollinate, import, bite, wash, bring home, use, dry, ban, disinfect, sell, consume, substitute, hand out, serve, sanitize, drizzle, pick up, treat, export, thaw, bring with, fry, roast, fault, count, combine, pull, rot, test, waltz, store, get rid of, produce, yank, snitch, slug, replace, busy, take away, Cup, prefer, vault, thin, work at, concede, reuse, add, mock, pare, buy, ship, pack, redden, bog down, tell on, firm, co-found, like, bathe, prune, hang up, talk, segregate, base, hack, wet, market, purchase, mound, riddle, dislodge, coerce, press, crush, contaminate, spur, stuff, filch, share, elongate, sort, go without, exonerate, glass, throw, equate, deep-fry, Wolf, Cook, leave, reexamine, place, join, dispense, sweeten, tie, spread on, introduce, pile, domesticate, license, clutch, include, keep, remove, brace, give up, wipe, inspect, arrange, bob, resemble, ally, beset, wad, reinvent, grip, deform, find, let, own, hollow, water, behold, load, push, irradiate, kiss, scent, sample, poison, Jam, freeze, dance, swathe, perturb, position, hold out, swallow, distribute, pressure, seek out, check out, reheat, cut, microwave, brush, quarantine, supply, add to, deliver, recapture, insert, mash, come to, blacklist, decorate, shape, steal, take possession of, bring in, stick with, come with, wager, drive, rob, gather, enjoy, moot, return to, crunch, run, simmer, zap, ferret out, have, criticise, accept, tickle, drain, put up, reposition, clean, force, get, cradle, promote, lend, praise, consolidate, flaunt, resuscitate, leave behind, threaten, found, reinvigorate, feed, afford, tote, categorize, divide, silver, process, treasure, carry, manufacture, mean, last, consist of, confuse, envelop, round out, cost, light up, shine, pour, galvanize, embroil, stick, popularize, target, locate, ask for, need, exploit, recall, transfer, refuse, shove, bet on, affiliate with, breed, scrutinize, elude, lift, have left, spoil, begin, bury, aim, figure out, spill, reestablish, photograph, connect, master, reorganize, favour, eradicate, line up, slide, strain, announce, pit, take, know, move, raise, allege, look at, become, contain, prepare, hamper, command, hold, develop, withhold, call, meet, try, concern, catch, fall, entitle, require, receive, consider, ask, say, report, make, release, lead, celebrate, live, experience, prevent, average, launch, resume, describe, free, favor, examine, worry, involve, surround, regard, disclose, mention, convince, welcome, monitor, serve as, manage, negotiate, tell, feature, reach, play, cause, attack, do, limit, cite, watch, read, attract, address, handle.

Distributional Semantics

[based on slides by Hovy 2011 / Pantel 02]



Hearst Patterns

M. Hearst, Automatic Acquisition of Hyponyms from Large Text Corpora 1992

(S1) The bow lute, such as the Bambara ndang, is plucked and has an individual curved neck for each string.

(1a) NP_0 such as $\{NP_1, NP_2 \dots, (and | or)\} NP_n$

are such that they imply

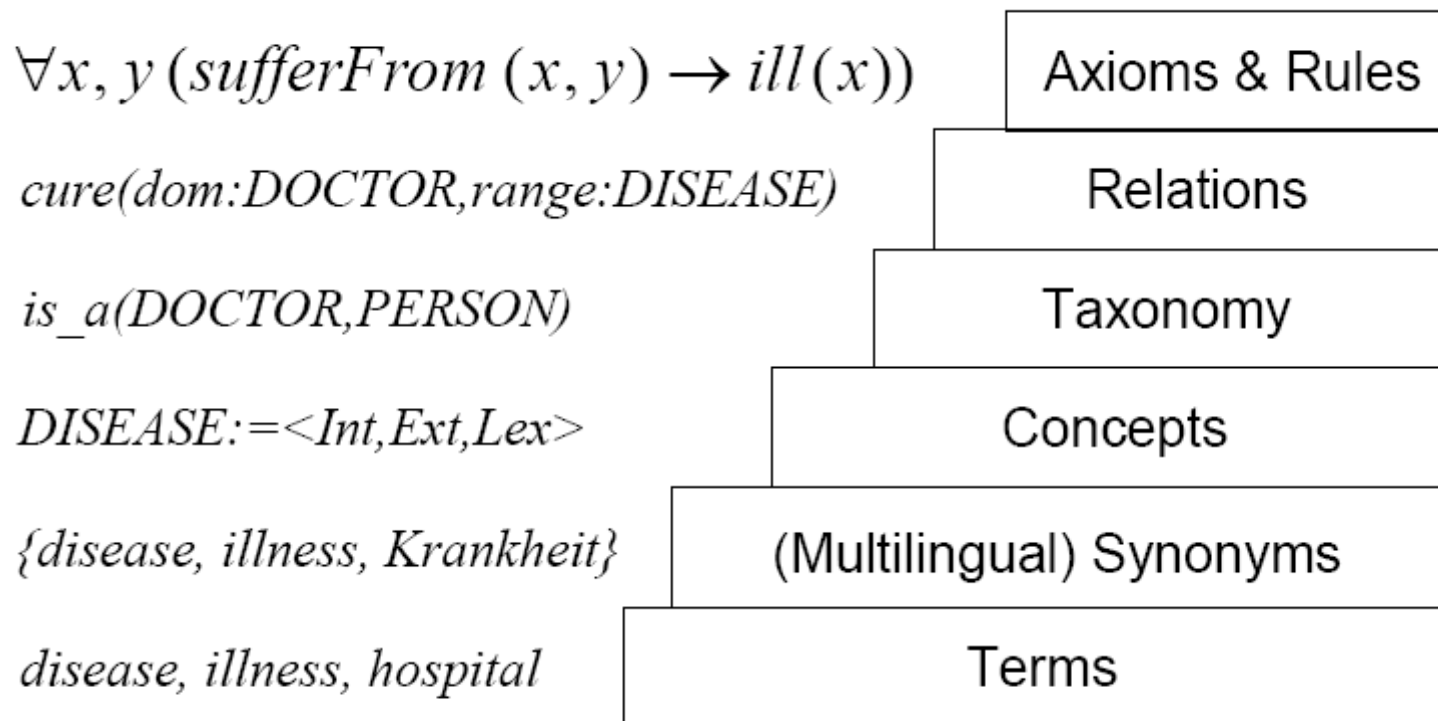
(1b) for all $NP_i, 1 \leq i \leq n, \text{hyponym}(NP_i, NP_0)$

Thus from sentence (S1) we conclude

$\text{hyponym}(\text{"Bambara ndang"}, \text{"bow lute"})$.

- (2) *such NP as* $\{NP, \}^* \{(or | and)\} NP$
 ... works by such authors as Herrick, Goldsmith, and Shakespeare.
 $\implies \text{hyponym}(\text{"author"}, \text{"Herrick"}),$
 $\text{hyponym}(\text{"author"}, \text{"Goldsmith"}),$
 $\text{hyponym}(\text{"author"}, \text{"Shakespeare"})$
- (3) *NP* $\{, NP\}^* \{, \}$ *or other NP*
 Bruises, wounds, broken bones or other injuries ...
 $\implies \text{hyponym}(\text{"bruise"}, \text{"injury"}),$
 $\text{hyponym}(\text{"wound"}, \text{"injury"}),$
 $\text{hyponym}(\text{"broken bone"}, \text{"injury"})$
- (4) *NP* $\{, NP\}^* \{, \}$ *and other NP*
 ... temples, treasuries, and other important civic buildings.
 $\implies \text{hyponym}(\text{"temple"}, \text{"civic building"}),$
 $\text{hyponym}(\text{"treasury"}, \text{"civic building"})$
- (5) *NP* $\{, \}$ *including* $\{NP, \}^* \{(or | and)\} NP$
 All common-law countries, including Canada and England ...
 $\implies \text{hyponym}(\text{"Canada"}, \text{"common-law country"}),$
 $\text{hyponym}(\text{"England"}, \text{"common-law country"})$
- (6) *NP* $\{, \}$ *especially* $\{NP, \}^* \{(or | and)\} NP$
 ... most European countries, especially France, England, and Spain.
 $\implies \text{hyponym}(\text{"France"}, \text{"European country"}),$
 $\text{hyponym}(\text{"England"}, \text{"European country"}),$
 $\text{hyponym}(\text{"Spain"}, \text{"European country"})$

Ontology Learning From Text



Introduced in: Philipp Cimiano, PhD Thesis University of Karlsruhe

Evaluation

Levels	Golden standard	Application-based	Data-driven	Assessment by humans
Lexical, vocabulary, data	X	X	X	X
Hierarchy, taxonomy	X	X	X	X
Other semantic relations	X	X	X	X
Context, application		X		X
Syntactic	X			X
Structure, architecture, design				X
Philosophical				X

A SURVEY OF ONTOLOGY EVALUATION TECHNIQUES, Janez Brank, Marko Grobelnik, Dunja Mladenić, Proceedings of the Conference on Data Mining and Data Warehouses (SiKDD 2005), 2005

Limitations of Knowledge Extraction from Text (or why we want to study semantics in crowds)

- **Delay**
 - Time it takes to write (high quality) text on a topic
- **Population bias and dependence**
 - Authors' demographics
 - study language of target groups (e.g. software developers)
- **Style**
 - Internal monologue vs. conversation with others
- **Topical bias**
 - General Fiction, Mystery, Science Fiction, Romance, Humor, ..
 - Books, newspapers, magazines [Brown Corpus]

Extracting Semantics from Crowds (i.e. Social Media)



coupled with behavior

semi-structured

near-real time

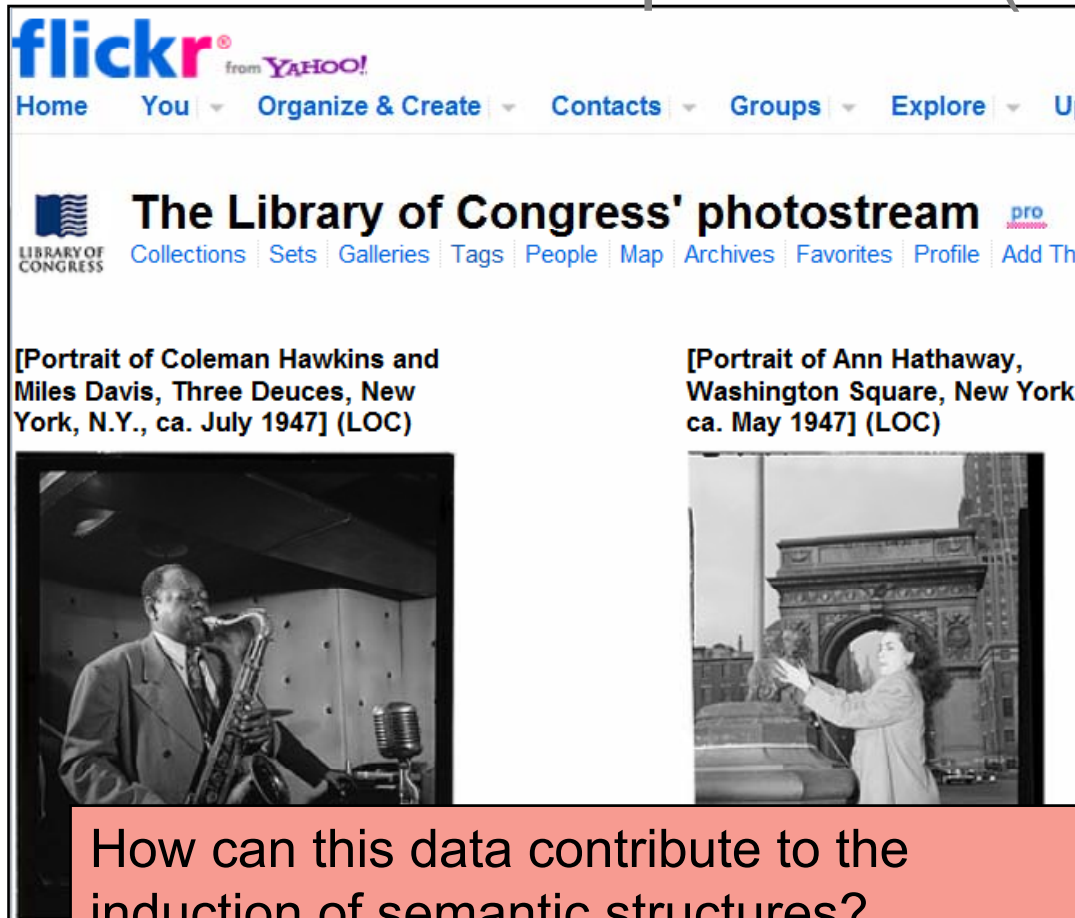
How can we tap into users interaction with data and with each other for the extraction of semantic structures?

Crowdsourcing

“Crowdsourcing represents the act of a company or institution taking a function once performed by employees and outsourcing it to an undefined (and generally large) network of people in the form of an open call.”

[J. Howe 2006]

Crowdsourcing Semantics: An Experiment (2008)



How can this data contribute to the induction of semantic structures?



LoC posted 3000 photos to flickr:

24 hours after launch:

- over 4,000 unique tags
- about 19,000 tags added

Output? noisy, unstructured, unqualified, weak semantics

Extracting Semantics from Crowds

Vision:

Utilizing online behavior of crowds for the construction, maintenance and enrichment of large-scale semantic structures.

Mission:

- to model behavior of large numbers (millions) of users online
- to develop techniques and algorithms that acquire semantic structures from **users' interaction with data and with each other**
- to influence user behavior and emerging semantics
- to evaluate results

Activities of Crowds Online

Users engage in...

Labeling

Tagging

Navigation

healthynewday Healthy New Day
 #Obesity: More Weight Equals Longer Hospital Stays...<http://tinyurl.com/cf6ghm>
 27 Mar 09

10 NOV 10 Collaborative Online Diagram Software - Try it Free | 3575
 Creately SAVE
 vtruffy design mindmapping infographies utilitaires
 Webapp

ENCHANTED LEARNING HOME PAGE 4998
 klaird elementary resources lessonplans

Berliner Fußball-Verband e. V.: Startseite 3
 Emrah Celik verband

COMPUTER LIB
 DREAM MACHINES

Computer Lib/Dream Machines by Ted Nelson

Tags
 anti-establishment computer computer science
 computing creativity freedom fundamenta
 history of technology hypermedia h
 liberation priesthood principles programming publishi
 technology and culture time travel va142 office virt
 xanadu

Can we tap into the outcome of these activities to extract and evaluate semantic structures?

Extracting Semantics from Crowds

Motivation

Social Labeling

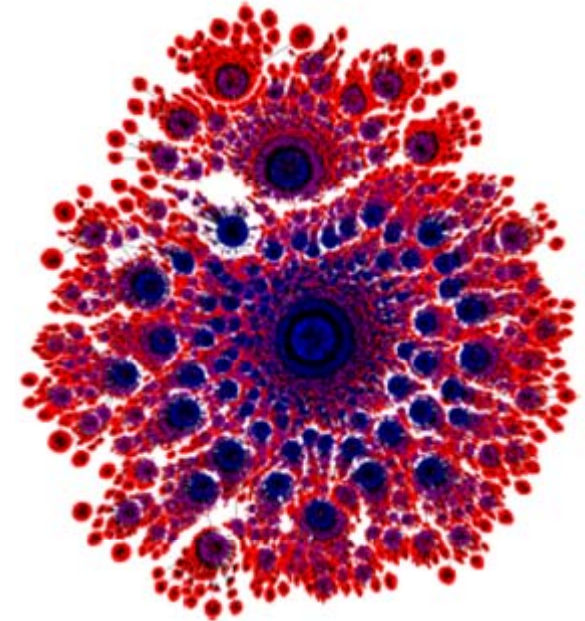
- Hashtag Semantics

Social Tagging

- Tag Relatedness
- Tag Generality
- Tag Hierarchies

Social Navigation

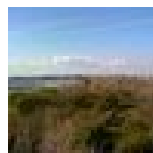
- Navigational Knowledge Engineering



Social Labeling

Example: Twitter

users label short messages
with concepts (hashtags)



healthynewday Healthy New Day
#Obesity: More Weight Equals Longer Hospital
Stays... <http://tinyurl.com/cf6ghm>
27 Mar 09

*depending on
study, up to 30%
of messages
contain labels*

Whether hashtags behave as strong identifiers, and if they could be mapped to concept identifiers in the Semantic Web (URIs)?

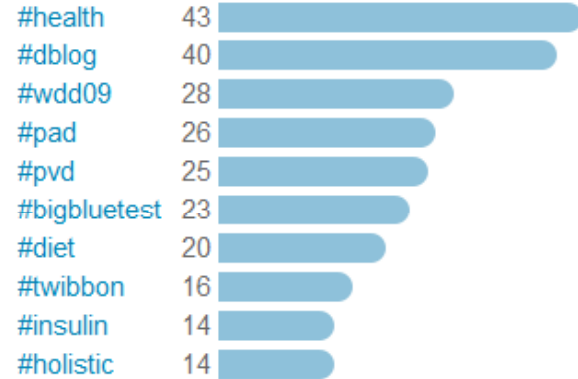
[Laniado and Mika 2010]

Craig's talk this morning: Background knowledge for URLs, to explore other URLs

#diabetes

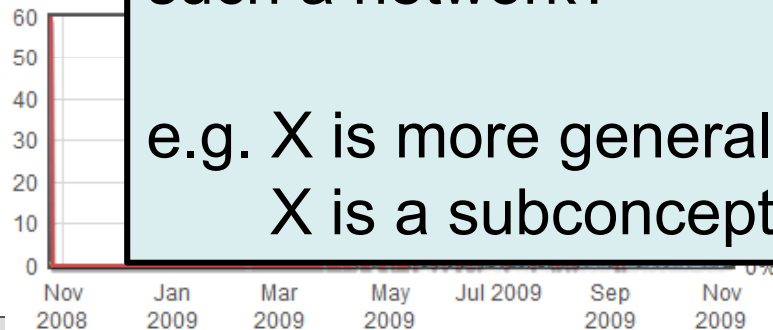
in 1012 tweets

Co-occurrences

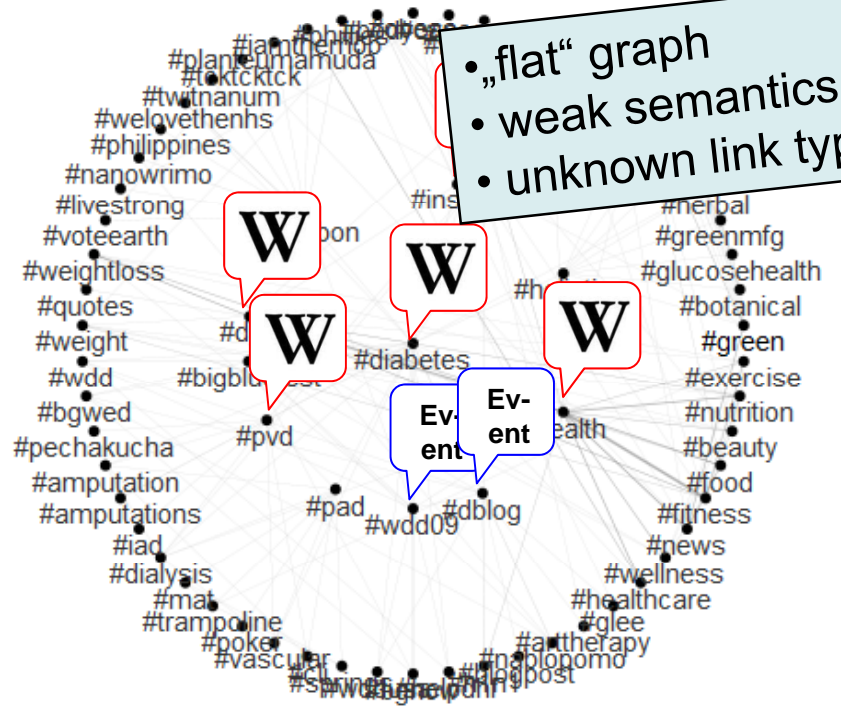


Statistics

Tagness: 10.346%
 Relative position: 65.772%
 Only has
 With link
 Social di

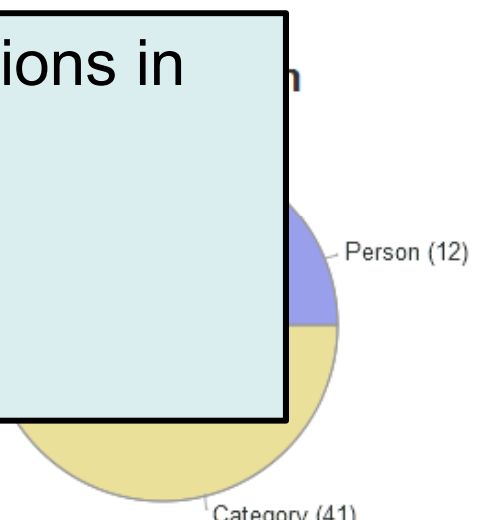


Count Relative count Social diversity



• „flat“ graph
 • weak semantics
 • unknown link types

Can we qualify semantic associations in such a network?
 e.g. X is more general than Y,
 X is a subconcept of Y



Extracting Semantics from Crowds

Motivation

Social Labeling

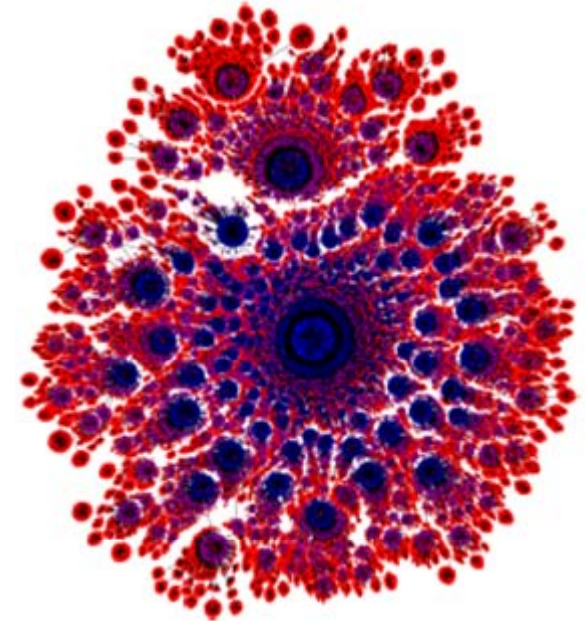
- Hashtag Semantics

Social Tagging

- Tag Relatedness
- Tag Generality
- Tag Hierarchies

Social Navigation

- Navigational Knowledge Engineering



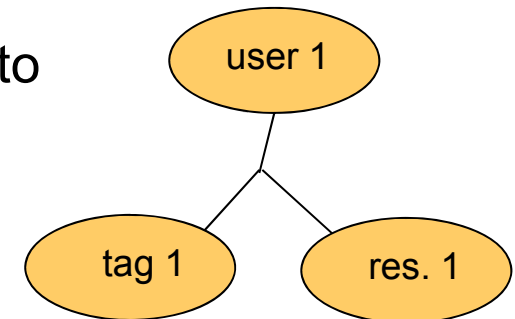
Social Tagging

Example: Delicious



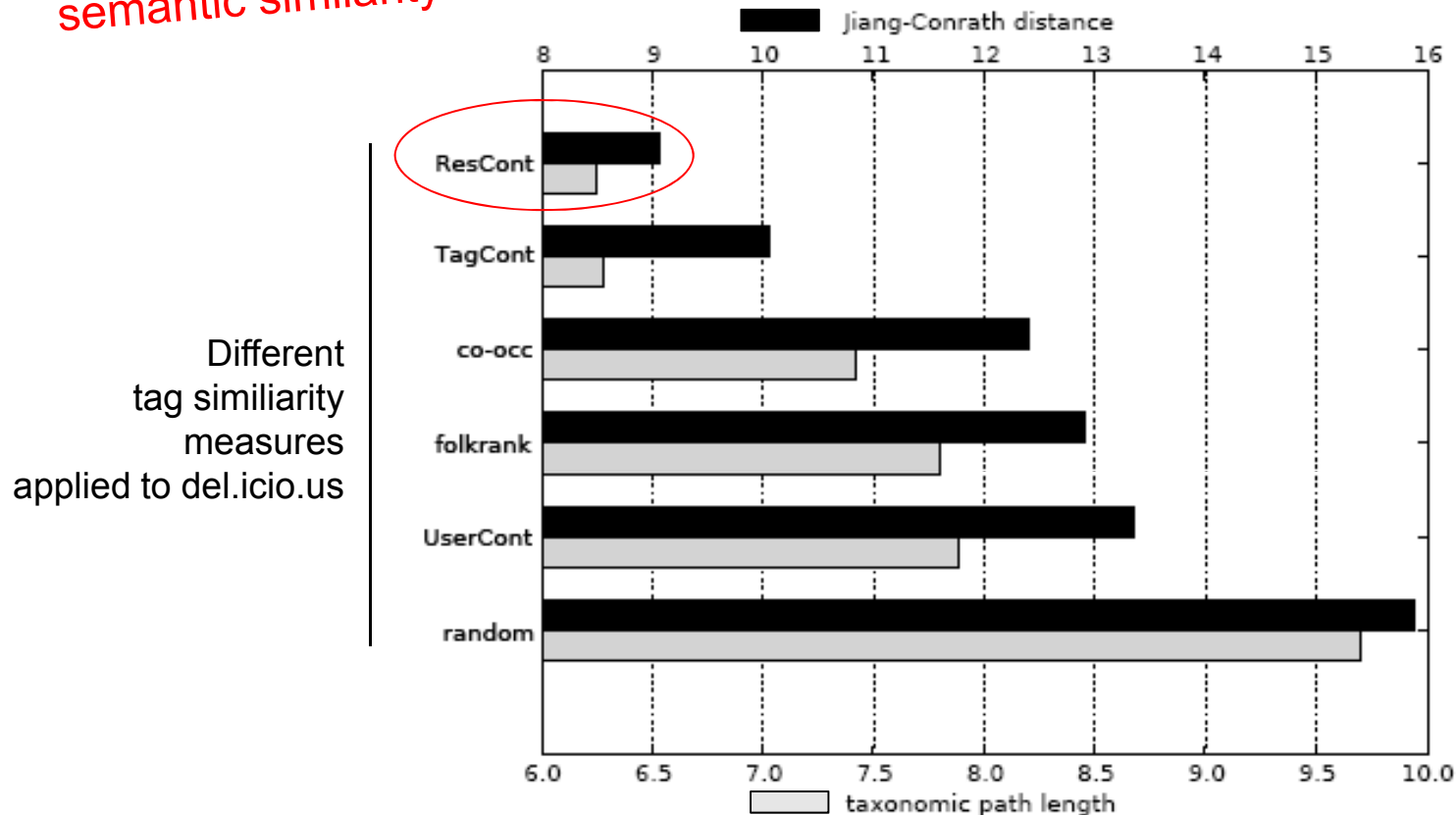
is a tuple $\mathbf{F} := (U, T, R, Y)$ where

- the three disjoint, finite sets U, T, R correspond to
 - a set of persons or users $u \in U$
 - a set of tags $t \in T$ and
 - a set of resources or objects $r \in R$
- $Y \subseteq U \times T \times R$, called set of *tag assignments*



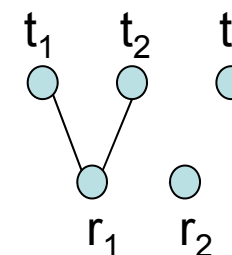
Tag Relatedness

To what extent does tag similarity reflect semantic similarity?

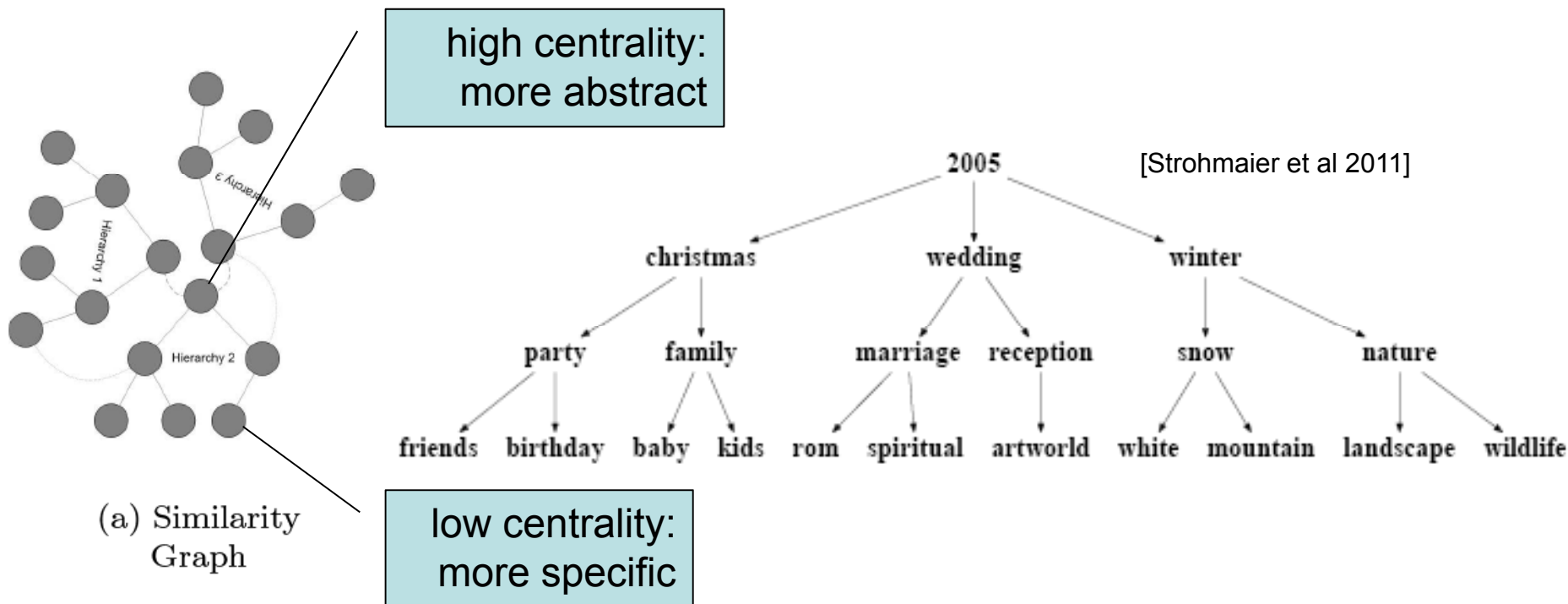


Different tag similarity measures applied to del.icio.us

ResCont



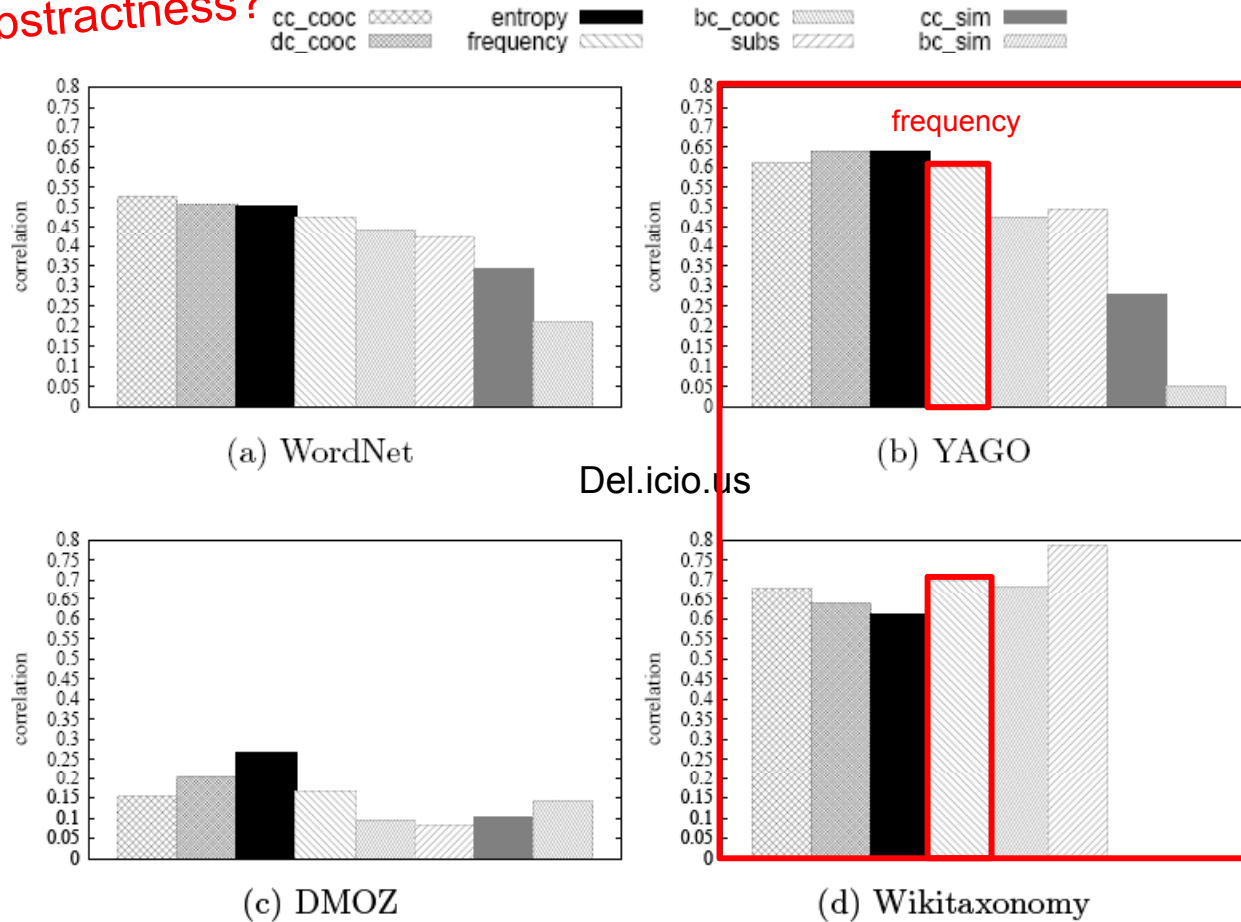
Intuition: Latent Hierarchical Structures



Note: Betweenness centrality is usually difficult to calculate. Calculating all shortest path is usually $O(n^2)$. For all nodes $O(n^3)$. We use an approximation.

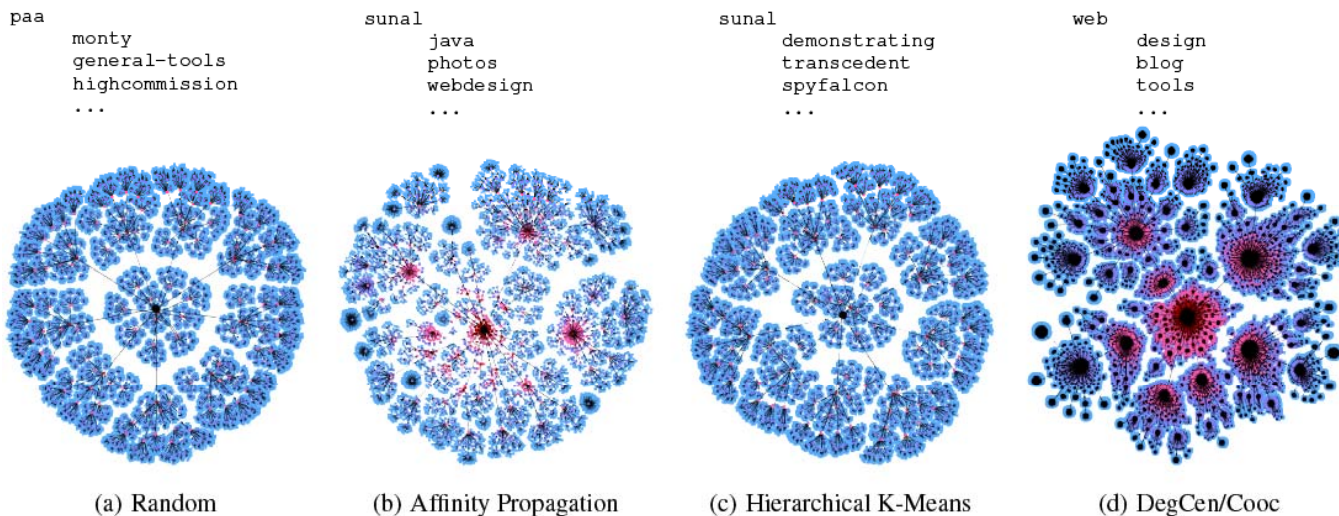
To what extent does tag abstractness reflect semantic abstractness?

Tag Abstractness



with D. Benz et al.

Emergent semantics through hierarchical clustering



on a delicious dataset

Approaches:

- k-means
- Affinity propagation
- Tag generality

Applications:

- user navigation
- ontology learning
- disambiguation

Evaluation:

- Semantic grounding to Golden Standards (e.g. WordNet)

Study of different tagging systems:

BibSonomy, CiteULike ,Delicious, Flickr, LastFM

Semantic and Pragmatic Quality varying greatly.

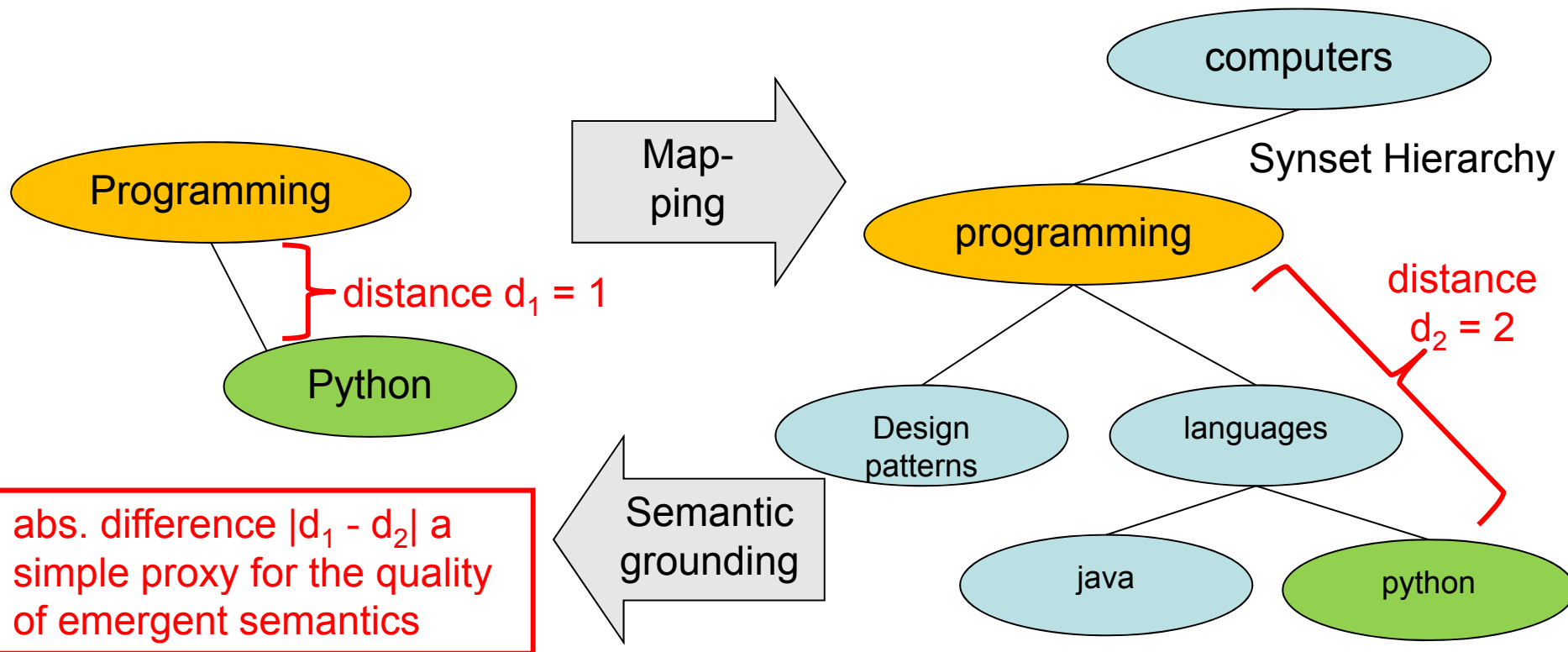
Semantic Validation of Folksonomies

Semantic Networks (Emergent)

via e.g. hierarchical clustering

Semantic Grounding (Golden Standard)

WordNet: a lexical DB for English





To what extent do tag hierarchies reflect semantic hierarchies?

Semantically Evaluating Tag Hierarchies

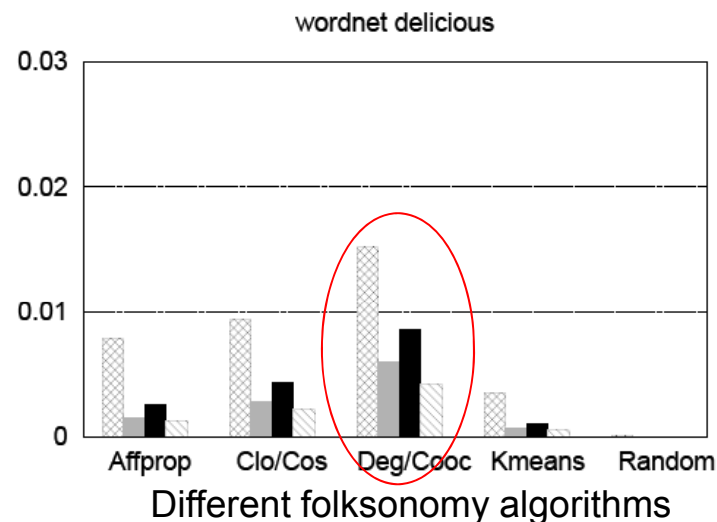
[Dellschaft, Staab 2006]

$$tp(c, \mathcal{F}, \mathcal{T}) = \frac{|ce(c, \mathcal{F}) \cap ce(c, \mathcal{T})|}{|ce(c, \mathcal{F})|}$$

$$tr(c, \mathcal{F}, \mathcal{T}) = \frac{|ce(c, \mathcal{F}) \cap ce(c, \mathcal{T})|}{|ce(c, \mathcal{T})|}$$

$$to(c, \mathcal{F}, \mathcal{T}) = \frac{ce(c, \mathcal{F}) \cap ce(c, \mathcal{T})}{ce(c, \mathcal{F}) \cup ce(c, \mathcal{T})}$$

TP TR TF TO
 TP..Taxonomic precision TR..Taxonomic recall TF..Taxonomic F measure TO...Taxonomic overlap



Holds with other knowledge bases (Yago, Wikitaxonomy) and datasets (bibsonomy, citeulike, lastfm less so)

[Dellschaft, Staab 2006] On How to Perform a Gold Standard Based Evaluation of Ontology Learning (2006), by Klaas Dellschaft, Steffen Staab, In Proceedings of the 5th International Semantic Web Conference (ISWC'06)

Limitations and Opportunities

Pragmatics influence semantics:

1. Tagging behavior effects preciseness of emerging semantics [Körner et al. 2010]
2. Social networks influence resulting semantic networks [Wang and Groth 2010]
3. Stream types effect stream semantics [Wagner and Strohmaier 2010]

Different motivations for tagging

The diagram illustrates two different tagging motivations for a YouTube video. At the top, two boxes represent 'User A' and 'User B'. Below them, a horizontal bar shows the tags assigned by each user. User A's tag is 'entertainment.video.selected', which is highlighted in a grey box. An arrow points from this box to a larger grey box on the left containing the text 'This seems to be a category!'. User B's tags are 'funny', 'music', 'hilarious', 'video', and 'youtube', which are also highlighted in a grey box. An arrow points from this box to a larger grey box on the right containing the text 'These seem to be keywords!'. In the center is a screenshot of a YouTube video player. The video title is 'Crazy Indian Video... Buffalaxed!'. The video shows a man in a red shirt dancing with a group of people in red and white outfits. The video player interface includes the YouTube logo, a search bar, and playback controls. At the bottom of the video player, it shows '★★★★★ 54.312 Bewertungen' and '15.776.926 Aufrufe'.

How do Semantics Emerge?

Are they Influenced by the Pragmatics of Tagging?

Different styles of tagging:

To *categorize* or to *describe* resources [Strohmaier et al. 2010]

	Categorizer (C)	Describer (D)
Goal	later browsing	later search
Change of vocabulary	costly	cheap
Size of vocabulary	limited	Open
Tags	subjective	objective

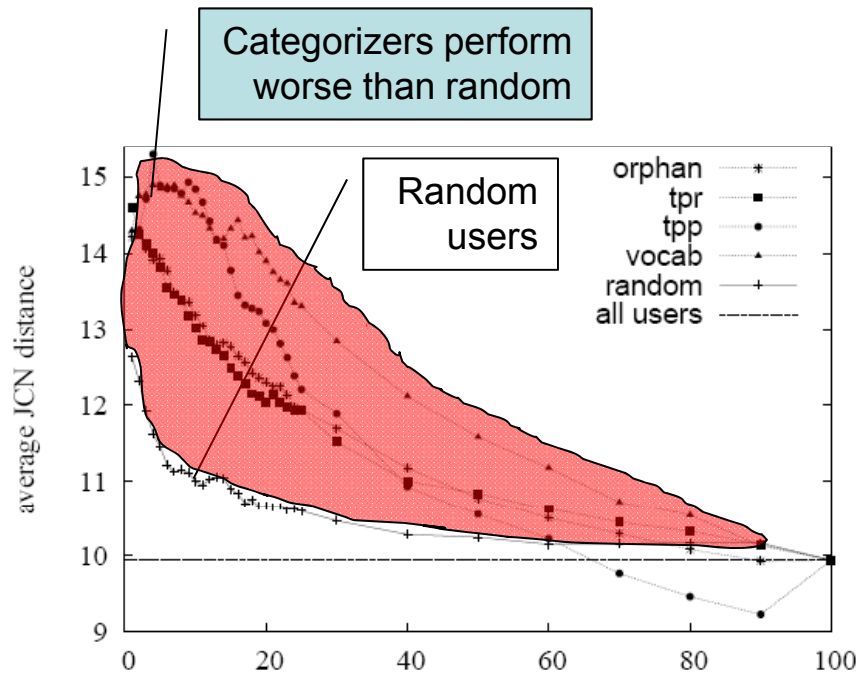
Example tag clouds

art blog book boys celebrity design design_blog
 design_magazine design_shop education entertainment
fashion_blog fashion_blog_men fashion_blog_sneakers
fashion_brand fashion_brand_bags fashion_brand_eyewear
 fashion_brand_shoes fashion_community fashion_designer
fashion_magazine fashion_model fashion_photography
fashion_shop fashion_show fashion_streetstyle food
 fragrance games health iklan local music news
 Errors es_reviews español español,espanol,blog espanol. events
 everyday examples exce **excel** Excel_Functions Excel2007 Exceler
 excelets ExcelPoster Excl exxel experts face ferrero fertility file
 filemaker files finance financial financialanalysis firefoxrss Flags
 flash Flash_Drawing flickr_blogging **flip** flooding flowcharts flowmap
 forex Formats formulas **forum** forums foul foun fractal france Free
 freelancer freelancers Freeway freelancers french Friends **fun**
 functions **gallery** Gallery gambling games ganar Gantt gapminder

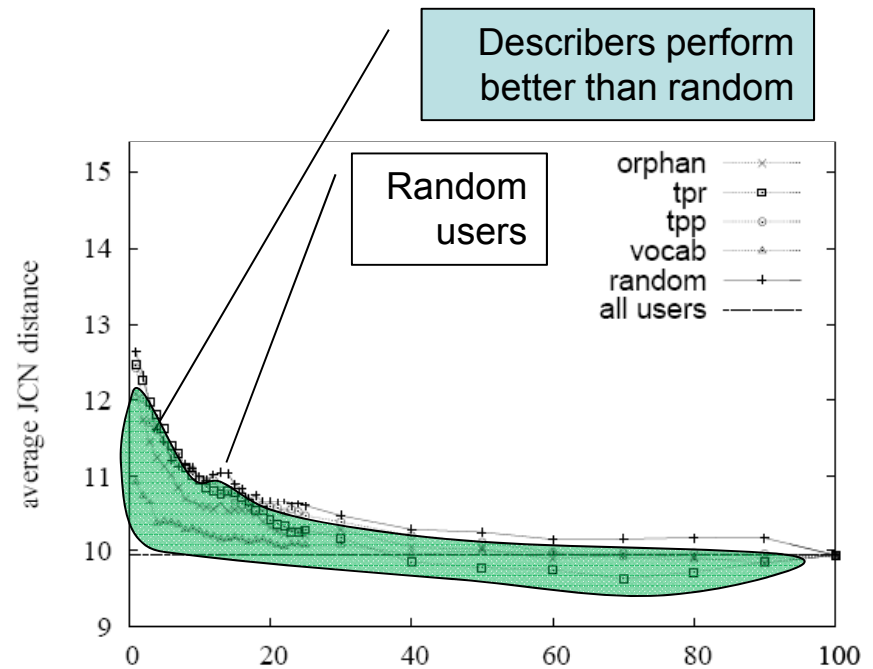
Semantic Assumption:
Categorizers produce more precise emergent semantics than Describers.

Example 1

Descriptors outperform categorizers on precision of emergent tag semantics



(C) → percentage of included users
(a) CF_i^m .



(D) → percentage of included users
(b) DF_i^m .

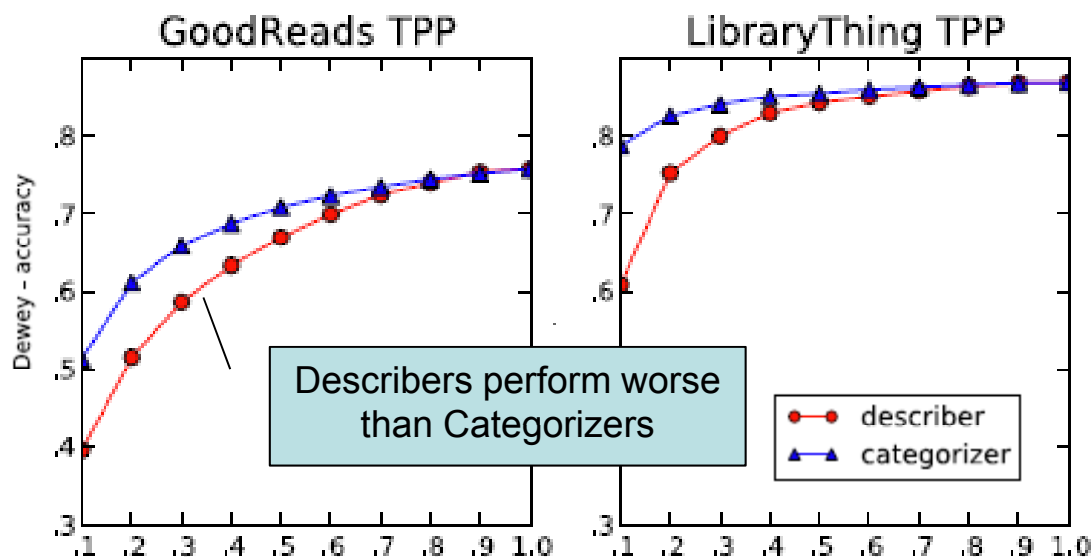
Example 2

Categorizers outperform describers on social classification accuracy

Overview of the Dewey Decimal Classification

The ten main classes are:

- 000 Computers, information & general reference
- 100 Philosophy & psychology
- 200 Religion
- 300 Social sciences
- 400 Language
- 500 Science
- 600 Technology
- 700 Arts & recreation
- 800 Literature
- 900 History & geography

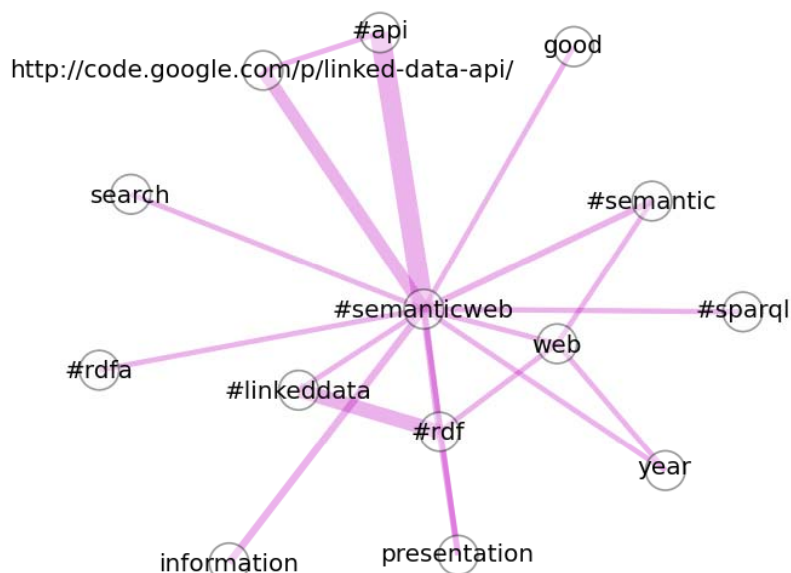


Supervised multi-class SVM

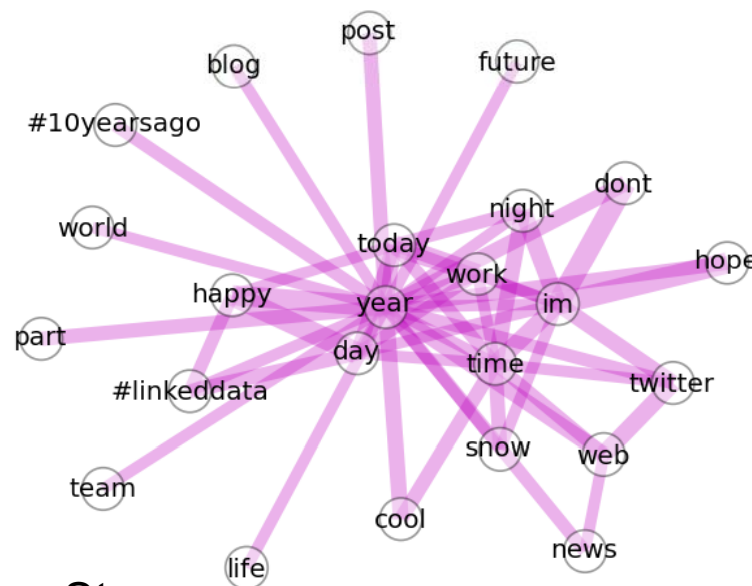
Example 3

Type of stream aggregation effects semantics

... the type of aggregation:
 Hashtag stream aggregations are more robust against external disturbances than user list streams



Hashtag Stream
 $O_R(RU_a)S(R_h)$



User List Stream
 $O_R(RU_a)S(R_{UL})$

Extracting Semantics from Crowds

Motivation

Social Labeling

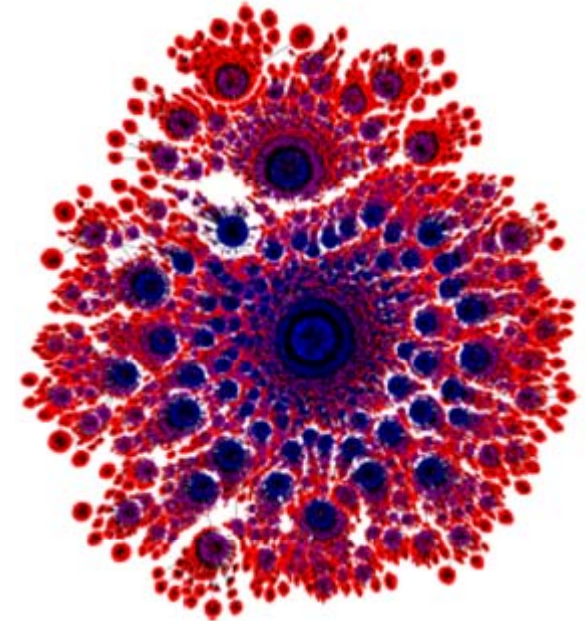
- Hashtag Semantics

Social Tagging

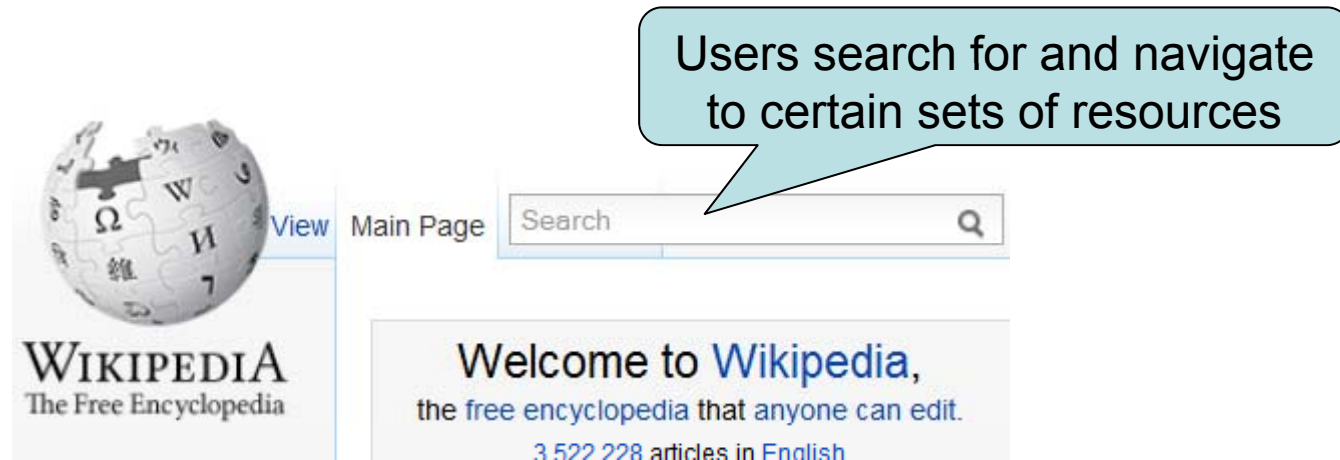
- Tag Relatedness
- Tag Generality
- Tag Hierarchies

Social Navigation

- Navigational Knowledge
- ### Engineering



Social Navigation



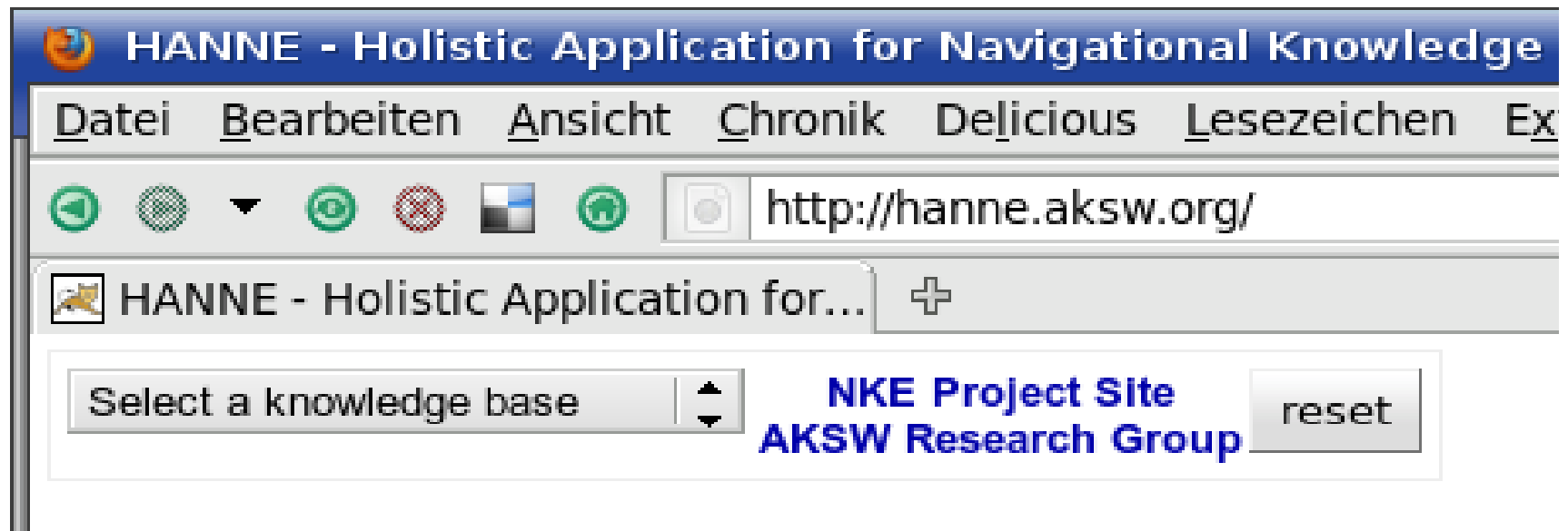
Can we produce ontological constructs as a byproduct of navigational activities of users ?

Navigational Knowledge Engineering:
A light-weight methodology for low-cost knowledge engineering by a massive user base.

Prototype: HANNE

HANNE: Holistic Application for Navigational
Knowledge Engineering

<http://aksw.org/Projects/NKE>

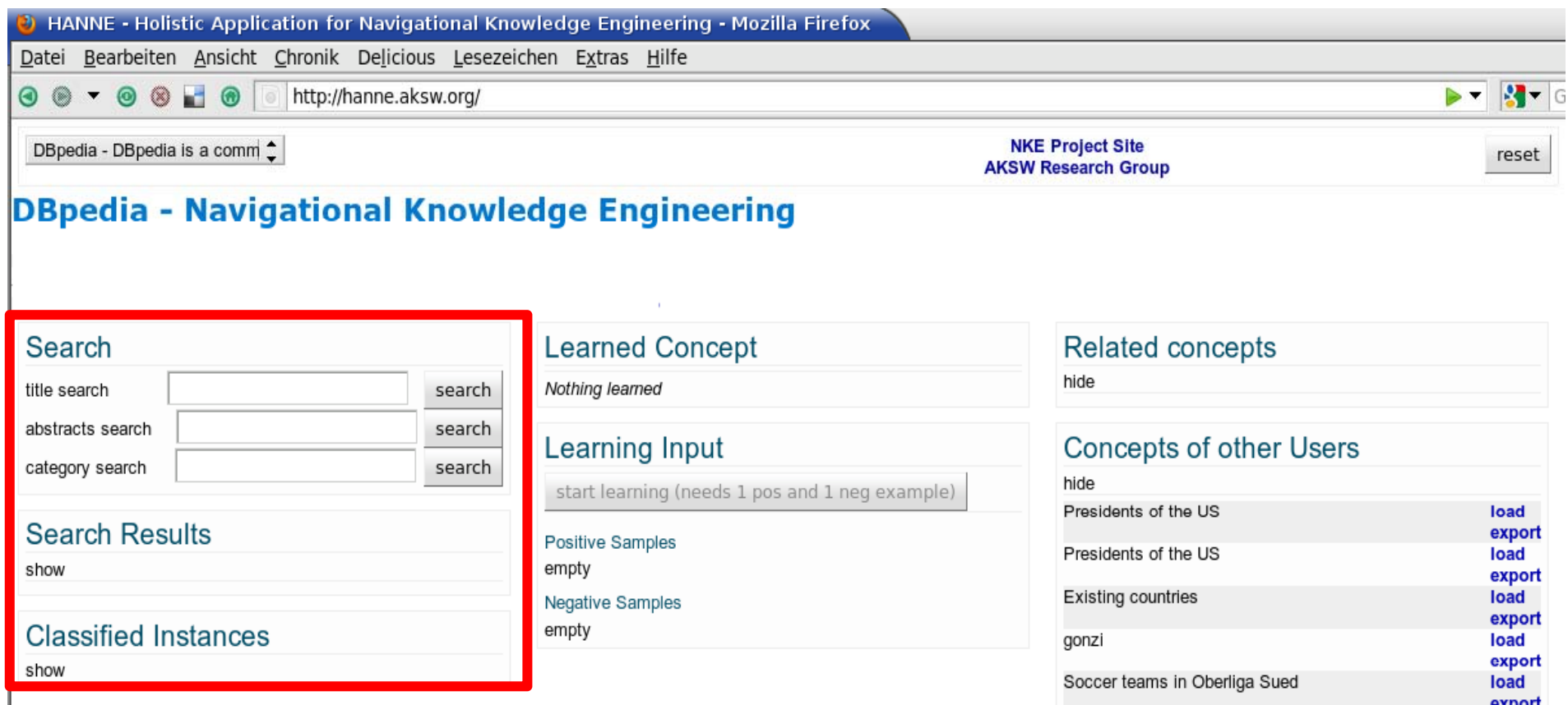


with S. Hellmann, J. Lehmann, C. Stadler, J. Unbehauen, University of Leipzig

Navigational Knowledge Engineering

<http://aksw.org/Projects/NKE>

Example: Extending DBPedia with NKE



The screenshot shows the HANNE web application in a Mozilla Firefox browser. The address bar shows <http://hanne.aksw.org/>. The page title is "DBpedia - Navigational Knowledge Engineering".

The interface includes several sections:

- Search:** A red box highlights this section, which contains three search input fields: "title search", "abstracts search", and "category search", each with a "search" button.
- Search Results:** A section with a "show" button.
- Classified Instances:** A section with a "show" button.
- Learned Concept:** A section displaying "Nothing learned".
- Learning Input:** A section with a "start learning (needs 1 pos and 1 neg example)" button and two empty lists: "Positive Samples" and "Negative Samples".
- Related concepts:** A section with a "hide" button.
- Concepts of other Users:** A section with a "hide" button and a list of concepts with "load" and "export" buttons:

Presidents of the US	load	export
Presidents of the US	load	export
Existing countries	load	export
gonzi	load	export
Soccer teams in Oberliga Sued	load	export

Navigational Knowledge Engineering

Search

title search

abstracts search

category search

Search Results

hide

Positive Matches

Hertha BSC [more...](#)
 Hertha Berliner Sport-Club von 1892 (commonly known as Hertha BSC or Hertha Berlin) is a German football club based in Berlin. A founding member of the German Football Association in Leipzig in 1900, the club has a long history as Berlin's best-supported side and competes today in the first division Bundesliga. + -

Lokomotive Leipzig [more...](#)
 1. FC Lokomotive Leipzig is a German football club based in the city of Leipzig, Saxony and may be more familiar to many of the country's football fans as the historic side VfB Leipzig, the first national champions of Germany. They currently play in the 5th tier of the German football league system. + -

Zentralstadion [more...](#)
 The Zentralstadion (Central Stadium), located in Leipzig, Saxony, Germany, is the premier football facility in the former East Germany. It has served as the home of FC Sachsen Leipzig which will soon make way for RB Leipzig the latest football team property of energy drink maker Red Bull. A new licensing agreement with the company will see the facility renamed. + -

FC Sachsen Leipzig [more...](#)
 FC Sachsen Leipzig is a German football club playing in Leipzig, Saxony. The roots of the club go back to 1899 and the founding of Britannia Leipzig. Following World War I, a 1919 merger with FC Hertha 05 Leipzig created Leipziger Sportverein 1899. Another predecessor side, SV Tura Leipzig, was formed in 1932 and just six years later, in 1938, joined with 1899 to create Tura 1899 Leipzig. + -

Learned Concept

Nothing learned

Learning Input

start learning (needs 1 pos and 1 neg example)

Positive Samples
empty

Negative Samples
empty

Choose initial **positive** and **negative** examples from the search result.

Here we are looking for Football Clubs in Saxony, a region in Germany.

<http://aksw.org/Projects/NKE>

The Problem of Inductive Concept Learning

U is a universal set of objects

C is a concept: a subset of objects in U $C \subseteq U$

To learn a concept C means to learn to recognize objects in C, to be able to tell whether

$x \in C$ for each $x \in U$

Example:

U may be the set of all patients in a register, and $C \subseteq U$

The set of all patients having a particular disease

Inductive Concept Learning

Inductive concept learning Given a set \mathcal{E} of positive and negative examples of a concept \mathcal{C} , find a hypothesis \mathcal{H} , expressed in a given concept description language \mathcal{L} , such that:

- every positive example $e \in \mathcal{E}^+$ is covered by \mathcal{H} ,
- no negative example $e \in \mathcal{E}^-$ is covered by \mathcal{H} .

To test the coverage, the function

covers(\mathcal{H}, e)

Returns the value true if e is covered by \mathcal{H} , and false otherwise.

HANNE & The Problem of Inductive Concept Learning

given:

- **Background knowledge** (OWL/DL knowledge base)
- **positive** and **negative** examples of a concept (instances)

find:

- A hypothesis (expressed as OWL class descriptions) that covers all **positive** and no **negative** examples

Search

title search	<input type="text"/>	search
abstracts search	Football Leipzig	search
category search	<input type="text"/>	search

Search Results

hide

Positive Matches

Nicky Adler more...

Nicky Adler (born 23 May 1985 in Leipzig) is a German footballer. In 2003 he moved from the amateur ranks into the professional game as a player for 1860 Munich. Adler played twice for the Under-19's national team, eleven times for the Under-20s (scoring four goals in the process) and 68 times in lower-league football before starting out as a senior player. Adler made his debut in the top flight of German football on 1 September 2007 at Energie Cottbus.

+ -

Bernd Stange more...

Bernd Stange is a German football manager currently managing Belarus. Stange started playing at an early age and was called into the East German youth team. He continued to play for Chemie Gnaschwitz in the lower divisions until 1965 and later a year at Vorwärts Bautzen before joining HSD DHfK Leipzig, playing until retiring in 1970.

+ -

Mario Strikers Charged more...

Mario Strikers Charged, known as Mario Strikers Charged Football in PAL regions and Mario Power Soccer in the Koreas, is a sports video game developed by Canadian developer Next Level Games and published by Nintendo for the Wii. This game was announced at the 2006 Games Convention in Leipzig, Germany as the sequel to Super Mario Strikers for the Nintendo GameCube. It was released on May 25, 2007 in Europe, June 7, 2007 in Australia and July 30, 2007 in North America.

+ -

Learned Concept

Nothing learned

Learning Input

start learning

Positive Samples

Lokomotive Leipzig more...

1. FC Lokomotive Leipzig is a German football club based in the city of Leipzig, Saxony and may be more familiar to many of the country's football fans as the historic side VfB Leipzig, the first national champions of Germany. They currently play in the 5th tier of the German football league system.

x

FC Sachsen Leipzig more...

FC Sachsen Leipzig is a German football club playing in Leipzig, Saxony. The roots of the club go back to 1899 and the founding of Britannia Leipzig. Following World War I, a 1919 merger with FC Hertha 05 Leipzig created Leipziger Sportverein 1899. Another predecessor side, SV Tura Leipzig, was formed in 1932 and just six years later, in 1938, joined with 1899 to create Tura 1899 Leipzig.

x

Negative Samples

Hertha BSC more...

Hertha Berliner Sport-Club von 1892 (commonly known as Hertha BSC or Hertha Berlin) is a German football club based in Berlin. A founding member of the German Football Association in Leipzig in 1900, the club has a long history as Berlin's best-supported side and competes today in the first division Bundesliga.

x

Zentralstadion more...

The Zentralstadion (Central Stadium), located in Leipzig, Saxony, Germany, is the premier football facility in the former East Germany. It has served as the home of FC Sachsen Leipzig which will soon make way for RB Leipzig the latest football team property of energy drink maker Red Bull. A new licensing agreement with the company will see the facility renamed.

x

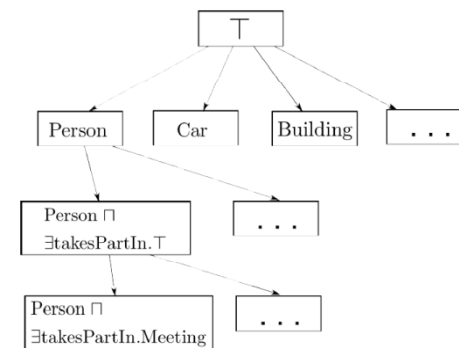
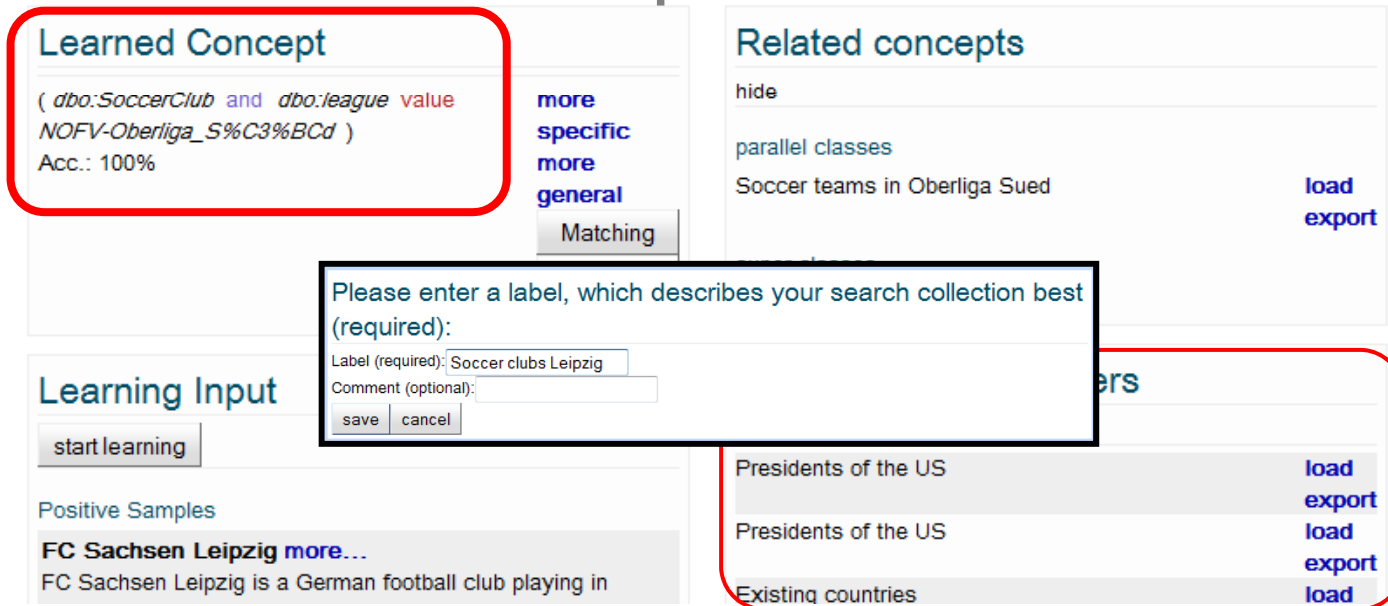


Figure 2: Illustration of a search tree in OCEL.

Based on the extension, ICL searches for suitable hypotheses.

Concepts Learned



Learned Concept

(*dbo:SoccerClub* and *dbo:league* value *NOFV-Oberliga_S%C3%BCd*)
Acc.: 100%

more specific
more general

Matching

Related concepts

hide

parallel classes
Soccer teams in Oberliga Sued

load export

Please enter a label, which describes your search collection best (required):

Label (required): Soccer clubs Leipzig
Comment (optional):

save cancel

Learning Input

start learning

Positive Samples

FC Sachsen Leipzig more...
FC Sachsen Leipzig is a German football club playing in

Presidents of the US load export
Presidents of the US load export
Existing countries load

- The Learned Concept is shown in Manchester OWL Syntax
- The user can retain the concept for later retrieval.
- Saved concepts are displayed as social navigation suggestions. Can be used to enrich existing knowledge base.

Result

Classified Instances 13 instances

hide (displaying result 1-13)

FSV Zwickau more...
 FSV Zwickau is a German football club located in Zwickau, Saxony. Today's club claims as part of its complex heritage sides that were East Germany's first champions: 1948 Ostzone winners SG Planitz and 1950 DDR-Oberliga champions ZSG Horch Zwickau.

1. FC Gera 03 more...
 1. FC Gera 03 is a German football club located in Gera, Thuringia.

SC Borea Dresden more...
 SC Borea Dresden are a German football club from the city of Dresden, Saxony. The club dropped the name FV Dresden-Nord on July 1, 2007 and adopted its current name to help encourage new sponsorship support. Boreas is the Greek god of the North Wind.

Budissa Bautzen more...
 Budissa Bautzen is a German football club from Bautzen, Saxony. Founded as Fußball Klub Budissa Bautzen on 24 May 1904, the club was part of East German competition after World War II.

VfB Auerbach more...
 VfB Auerbach is a German football club from the city of Auerbach, Saxony. The club also has a tennis department.

FC Eilenburg more...
 FC Eilenburg is a German football club from the city of Eilenburg, Saxony. The association was created January 1, 1994 as the successor side to Sportverein Möritz. The new club joined the Landesliga Sachsen (V) on the strength of SV's Bezirksliga Sachsen (VI) title. A poor 13th place finish immediately returned the side to sixth tier play where they would remain for three seasons.

VfB Germania Halberstadt more...
 VfB Germania Halberstadt is a German football club from Halberstadt, Saxony-Anhalt.

ZFC Meuselwitz more...
 Zipsendorfer Fußballclub Meuselwitz is a German football club from Meuselwitz, Thuringia.

VfB Pößneck more...
 VfB Pößneck is a German football club from the city of Pößneck, Thuringia with a membership of roughly 400.

SSV Markranstädt more...
 SSV Markranstädt is a German football club from the city of Markranstädt, Saxony near Leipzig. It is

Learning Input

Positive Samples

Lokomotive Leipzig more...
 1. FC Lokomotive Leipzig is a German football club based in the city of Leipzig, Saxony and may be more familiar to many of the country's football fans as the historic side VfB Leipzig, the first national champions of Germany. They currently play in the 5th tier of the German football league system.

FC Sachsen Leipzig more...
 FC Sachsen Leipzig is a German football club playing in Leipzig, Saxony. The roots of the club go back to 1899 and the founding of Britannia Leipzig. Following World War I, a 1919 merger with FC Hertha 05 Leipzig created Leipziger Sportverein 1899. Another predecessor side, SV Tura Leipzig, was formed in 1932 and just six years later, in 1938, joined with 1899 to create Tura 1899 Leipzig.

Negative Samples

Hertha BSC more...
 Hertha Berliner Sport-Club von 1892 (commonly known as Hertha BSC or Hertha Berlin) is a German football club based in Berlin. A founding member of the German Football Association in Leipzig in 1900, the club has a long history as Berlin's best-supported side and competes today in the first division Bundesliga.

Zentralstadion more...
 The Zentralstadion (Central Stadium), located in Leipzig, Saxony, Germany, is the premier football facility in the former East Germany. It has served as the home of FC Sachsen Leipzig which will soon make way for RB Leipzig the latest football team property of energy drink maker Red Bull. A new licensing agreement with the company will see the facility renamed.

AZ (football club) more...
 AZ, an acronym for Alkmaar Zaanstreek, is a football club from the city of Alkmaar, Nethenands. They are the current Dutch Eredivisie champions.

Charlton Athletic F.C. more...
 Charlton Athletic Football Club (also known as The Addicks) is a professional association football club based in Charlton, in the London Borough of Greenwich. The club was founded on 9 June 1905, when a number of youth clubs in the South-East London area, including East Street Mission and Blundell Mission, combined to form Charlton Athletic Football Club.

Useful properties:

- Biased towards high recall
- Scales well: Number of training examples is more important than the size of the background knowledge

With only 2 positives and 4 negatives, it is possible to find 13 more instances, which are football clubs situated close to Saxony, Germany

Mock up (1)



WIKIPEDIA
The Free Encyclopedia

- Main page
- Contents
- Featured content
- Current events
- Random article
- Donate to Wikipedia

- Interaction
 - Help
 - About Wikipedia
 - Community portal
 - Recent changes
 - Contact Wikipedia

Toolbox

Special page

Soccer Leipzig



Search results

From Wikipedia, the free encyclopedia

Socccr Leipzig

Search

Content pages [Multimedia](#) [Help and Project pages](#) [Everything](#) [Advanced](#)

Results **1-20** of **255** for **Soccer Leipzig**

You may create the page "[Soccer Leipzig](#)", but consider checking the search results below to see whether it is already covered.

For search help, please visit [Help:Searching](#).

RB Leipzig

RB **Leipzig** is a German football Category:German football clubs | club based in ... External links : [com/soccer/leipzig/de/home.html](#) RB **Leipzig**



...
12 KB (1,560 words) - 16:27, 9 December 2010

Dieter Kühn

Dieter Kühn (born July 4, 1956 in **Leipzig** , Saxony) is a former football (**soccer**) player from East Germany , who won the silver medal with ...



3 KB (229 words) - 15:51, 25 October 2010

1. FC Lokomotive Leipzig

FC Lokomotive **Leipzig** is a German football club based in the city of **Leipzig** in Saxony ... External links : [abseits-soccer.com/clubs/leipzig...](#)



19 KB (2,679 words) - 00:19, 11 December 2010

Sebastian Albert

Sebastian Albert (born 26 February 1987 in **Leipzig**) is a German footballer who ... [com/soccer/leipzig/de/rb-leipzig-team-spieler-detail_10.](#)



Learning Input

start learning

Positive Samples

RB Leipzig more...

RB Leipzig (RasenBallSport Leipzig e. V.) is a German football club based in Leipzig, formed in 2009. The club plays in the tier-five NOFV-Oberliga Süd in 2009-10.



Lokomotive Leipzig more...

1. FC Lokomotive **Leipzig** is a German football club based in the city of Leipzig, Saxony and may be more familiar to many of the country's football fans as the historic side VfB Leipzig, the first national champions of Germany. They currently play in the 5th tier of the German football league system.



Negative Samples

Sebastian Albert more...

Sebastian Albert (born 26 February 1987 in Leipzig) is a German footballer who last played for Hansa Rostock.



Mock up (2)

Hello, [Sign in](#) to get personalized recommendations. New customer? [Start here.](#)

Sponsored by [Discover Card](#)

[Your Amazon.com](#) | [Today's Deals](#) | [Gifts & Wish Lists](#) | [Gift Cards](#)

[Your Account](#) | [Help](#)

[Shop All Departments](#)

Search

[Computer & Accessories](#)

Brands
Bestsellers
Accessories
Laptops
Desktops
Drives & Storage
Printers & Ink
Connected Home
Software
Deals

Department

- < Any Department
- < Electronics
- < Computers & Accessories
- < Memory Cards & External Storage
- External Hard Drives**

Electronics > Computers & Accessories > Memory Cards & External Storage > External Hard Drives > "hard drive"

Related Searches: [external hard drive](#), [internal hard drive](#), [portable hard drive](#).

Showing 1 - 24 of 1,620 Results Sort by [Relevance](#)


Shipping Option (What's this?)

Any Shipping Option

Free Super Saver Shipping

Brand

- Western Digital (177)
- Seagate (136)
- Hitachi (43)
- Toshiba (52)
- Iomega (80)
- LaCie (84)
- Oyen Digital (22)
- [See more...](#)



Western Digital WD Elements 1 TB USB 2.0 Desktop External Hard Drive WDBAAU0010HBK-NESN

Buy new: ~~\$129.99~~ **\$72.99**


[11 new](#) from \$72.99

[2 used](#) from \$80.00

Get it by **Tuesday, Dec 14** if you order in the next **37 hours** and choose one-day shipping.

Eligible for **FREE** Super Saver Shipping.

★★★★☆ (794)



Western Digital My Passport Essential SE 1 TB USB 3.0 and USB 2.0 Ultra Portable External Hard Drive WDBACX0010BBK-NESN (Midnight Black)

Buy new: ~~\$169.99~~ **\$129.99**


[28 new](#) from \$119.00

Get it by **Tuesday, Dec 14** if you order in the next **38 hours** and choose one-day shipping.


Eligible for **FREE** Super Saver Shipping.

★★★★☆ (39)


Positive Examples



Seagate Expansion 500 GB USB 2.0 Portable External Hard Drive ST905004EXA101-RK



Toshiba Canvio Plus 500 GB USB 2.0 Portable External Hard Drive E05A050CAU2XX (Raven Black)




Toshiba Canvio Plus 500 GB USB 2.0 Portable External Hard Drive E05A050CAU2XX (Raven Black)

Buy new: ~~\$94.39~~ **\$59.99**

[37 new](#) from \$59.99

[1 used](#) from \$60.00

Get it by **Tuesday, Dec 14** if you order in the next **38 hours** and choose one-day shipping.



Seagate Expansion 500 GB USB 2.0 Portable External Hard Drive ST905004EXA101-RK


Buy new: ~~\$89.00~~ **\$69.99**

[46 new](#) from \$59.98


[3 used](#) from \$55.95

Get it by **Tuesday, Dec 14** if you order in the next **38 hours** and choose one-day shipping.

Negative Examples



Western Digital WD Elements 1 TB USB 2.0 Desktop External Hard Drive WDBAAU0010HBK-NESN



Western Digital My Passport Essential SE 1 TB USB 3.0 and USB 2.0 Ultra Portable External Hard Drive WDBACX0010BBK-NESN (Midnight Black)

Extracting Semantics from ...

Motivation

Social Labeling

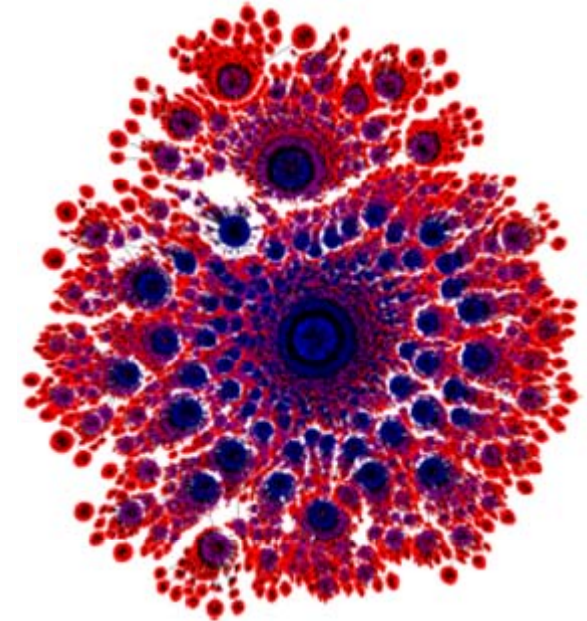
- Hashtag Semantics

Social Tagging

- Tag Relatedness
- Tag Generality
- Tag Hierarchies

Social Navigation

- Navigational Knowledge Engineering



→ Conclusions

Objectives

Provide (some) answers to the following questions:

- **What** is the difference between extracting semantics from text vs. extracting semantics from crowds?
- **Why** should we study crowds and crowd behavior from a semantic computing perspective?
- **How** can semantics be extracted from online crowd behavior, such as
 - ...from Social Labeling
 - ...from Social Tagging
 - ...from Social Navigation
- **What** are the implications for semantic computing research?

Social Computation for the Web of Data

Crowds provide some advantages over semantic extraction from text.

It is through the process of **social computation**, i.e. the combination of **social behavior** and **algorithmic computation**, that semantic structures emerge.

In order to extract semantics from crowds, **understanding users' behavior** and its impact on emerging semantic structures is important.

Shaping or classifying certain users' behavior are ways of increasing preciseness of emerging semantic structures.

Crowd Semantics

based on [Hovy 2011]

Def: A concept C is a list of triples

$$C = \{(r_1 w_1 s_1) (r_2 w_2 s_2) \dots (r_n w_n s_n)\}$$

where $r_i \in \{\text{Relations}\} = R$, e.g., *:subj*, *:agent*, *:color-of*

$w_i \in \{\text{Words}\} = \text{vocabulary}$, e.g., *happy*, *run*, *apple*

$s_i \in [0,1]$

and each w_i has been associated with C through the relation r_i , with a strength of association s_i that is computed under some measure.

Summary

We can observe that:

- semantic structures can be obtained as a byproduct of online crowd behavior (labeling, tagging, navigation)
- these structures can approximate structures in reference knowledge bases (DBPedia, WordNet, etc)

but:

- pragmatics influences resulting semantics
- semantic preciseness remains a challenge

An Agenda for Semantic Computing Research

Utilizing online behavior of crowds for the construction, maintenance and enrichment of large-scale semantic structures

Thank You.

Acknowledgements

co-authors and collaborators

D. Helic, C. Körner, J. Pöschko, R. Kern, C. Trattner, H.P. Grahsl, C. Wagner (Graz University of Technology, Austria), D. Benz, G. Stumme (U. of Kassel, Germany), A. Hotho (U. of Würzburg, Germany), S. Hellmann, J. Lehmann, C. Stadler, J. Unbehauen (U. of Leipzig, Germany), L. Hong (Parc, USA) and A. Zubiaga (UNED, Spain)