Social Networks and Multimedia Semantics

Peter Mika
Researcher
Yahoo! Research

About me and the lab

- Me
  - 29, born in Budapest, Hungary
  - Ph.D. at the Vrije Universiteit, Amsterdam
    - Thesis: Social Networks and the Semantic Web
  - Researcher at Yahoo! Research Barcelona
    - NLP and semantics
    - Semantic Search
    - Cloud Computing
- Yahoo! Research Barcelona
  - Established January, 2006
  - Led by Ricardo Baeza-Yates
  - Research areas
    - Web Mining
      - content, structure, usage
    - Distributed Web retrieval
    - Multimedia retrieval
    - NLP and Semantics

Yahoo! by numbers (April, 2007)

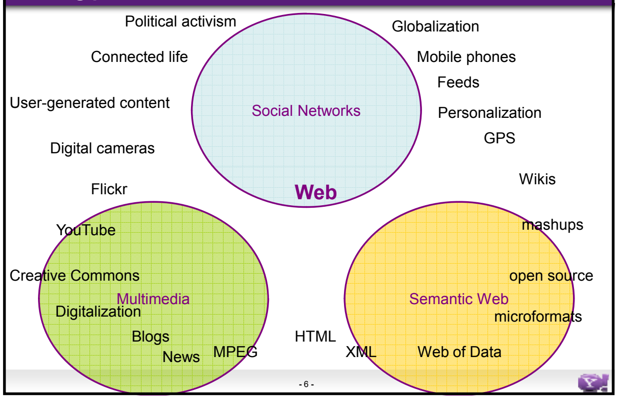
- There are approximately **500 million users** of Yahoo! branded services, meaning we reach 50 percent – or **1 out of every 2 users** – online, the largest audience on the Internet (Yahoo! Internal Data).
- Yahoo! is the most visited site online with nearly **4 billion visits** and **an average of 30 visits per user per month in the U.S.** and leads all competitors in audience reach, frequency and engagement (comScore Media Metrix, US, Feb. 2007).
- Yahoo! accounts for the largest share of time Americans spend on the Internet with 12 percent (comScore Media Metrix, US, Feb. 2007) and **approximately 8 percent of the world's online time** (comScore WorldMetrix, Feb. 2007).
- **Yahoo! is the #1 home page** with 85 million average daily visitors on Yahoo! homepages around the world, an increase of nearly 5 million visitors in a month (comScore WorldMetrix, Feb. 2007).
- Yahoo!'s social media properties (Flickr, delicious, Answers, 360, Video, MyBlogLog, Jumpcut and Bix) have **115 million unique visitors worldwide** (comScore WorldMetrix, Feb. 2007).
- Yahoo! Answers is the largest collection of human knowledge on the Web with more than 90 million unique users and **250 million answers** worldwide (Yahoo! Internal Data).
- There are more than **450 million photos** in Flickr in total and **1 million photos** are uploaded daily. 80 percent of the photos are public (Yahoo! Internal Data).

Yahoo! by numbers (April, 2007)

- **Del.icio.us hits 2 million users** in February, growing more than six times its size from 300,000 users in December 2005 (Yahoo! Internal Data).
- **Yahoo! Mail is the #1 Web mail provider in the world** with 243 million users (comScore WorldMetrix, Feb. 2007) and nearly 80 million users in the U.S. (comScore Media Metrix, US, Feb. 2007)
- Interoperability between Yahoo! Messenger and Windows Live Messenger has formed the largest IM community approaching 350 million user accounts (Yahoo! Internal Data).
- **Yahoo! Messenger is the most popular in time spent** with an average of 50 minutes per user, per day (comScore WorldMetrix, Feb. 2007).
- Nearly 1 in 10 Internet users is a member of a Yahoo! Groups (Yahoo! Internal Data).
- **Yahoo! News is the #1 online news destination** and has reached a new audience high in February with 36.2 million users, 10 million more users than its nearest competitor MSNBC (comScore Media Metrix, US, Feb. 2007).
- Yahoo! is one of only 26 companies to be on both the Fortune 500 list and the Fortune's "Best Place to Work" List (2006).

Overview

The big picture



Agenda

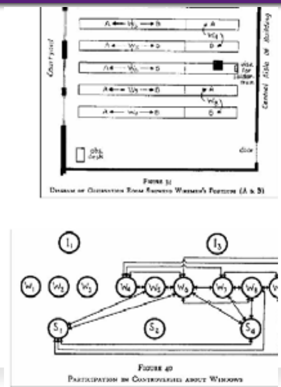
- Part 1: Social Networks and the Semantic Web
  - Investigating social networks on the Web
  - Semantics by emergence
- Part 2: Multimedia Semantics (courtesy of Roelof van Zwol)
  - Media Interaction
  - Media Mining
  - Media Search
- Bonus material: SearchMonkey!

- Research results and related work
- Hopefully ideas for your future work… and discussion

- 7 -

Social Networks on the (Semantic) Web

Network analysis circa 1920



Fast forward to 2003

Making Friendsters in High Places

By Leander Kahney | Also by this reporter

Page 1 of 2 next »

02:00 AM Jul. 17, 2003 PT

Friendster, the popular social-networking service that cleverly assimilates real-life social groups into a large virtual network, just keeps getting bigger.

The service, which opened to the public in March and is still in beta, will hit 1 million users this week, and is expanding at a rate of 20 percent a week, according to the company.

Story Tools

Story Images

Click thumbnails for full-size

"It's growing exponentially," said CEO and founder Jonathan Abrams.

Friendster helps users find dates and new friends by referring people to friends, or friends of friends, or friends

Social Networks on the Web

- **New opportunities for social science**
  - Explicit and implicit social network information
  - Large scale data sets
  - Dynamic data
  - Different modalities (profiles, email, IM, Twitter…)
- **Challenges**
  - Theoretical
    - Friend on the Web = Friend in reality?
  - Technical
    - Extracting information
    - Heterogeneity
    - Quality of data
    - Time and space complexity
  - Pragmatic
    - Ethical and legal challenges
- ➢ **Semantic technologies can help with some of the technical challenges**

- 11 -

SW representations of online social networks

- Friend-of-a-Friend (FOAF): a standard vocabulary for recording personal information in a machine readable format (RDF)
- FOAF documents contain information typically found in SNS and homepages:
  - name, homepage, image, interests, projects, publications, group memberships etc.
  - → extensible through RDF
- Distributed approach
  - FOAF profiles are hosted by the user and usually linked in from his homepage
  - → user retains control

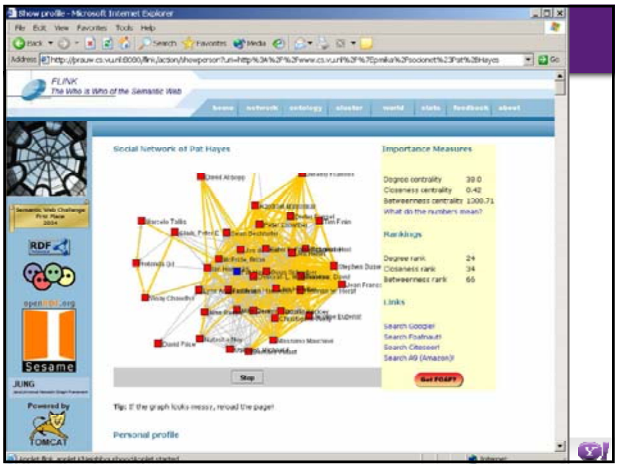
- 12 -

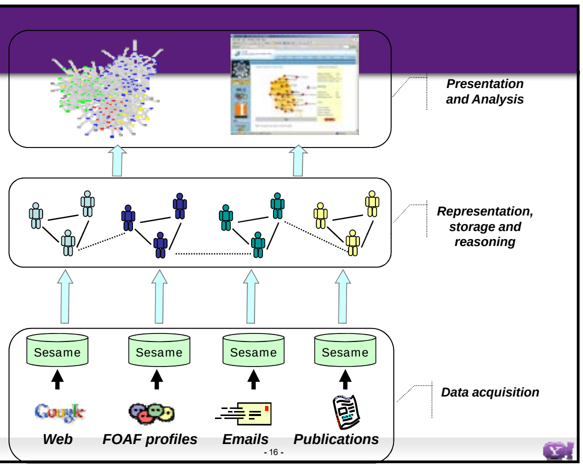
Example

```
<foaf:Person>
    <foaf:name>
        Frank van Harmelen
    </foaf:name>
    <foaf:mbox_sha1sum>
        241021fb0e6289f92815fc210f9e9137262c252e
    </foaf:mbox_sha1sum>
    <foaf:homepage rdf:resource="http://www.cs.vu.nl/~frankh" />
    <foaf:knows>
        <foaf:Person>
            <foaf:mbox rdf:resource="mailto:pmika@cs.vu.nl"/>
            <rdfs:seeAlso  rdf:resource="http://.../~pmika/foaf.rdf />
        </foaf:Person>
</foaf:Person>
```


Flink (2004)

- 1st prize, Semantic Web Challenge 2004
- Social network data collection, aggregation, storage and visualization
  - Data from user profiles and social network services using FOAF
  - Social network mining from the Web, emails
- Semantic Web technology
  - Ontology-based representation
    - Dealing with heterogeneity
  - Ontology-based reasoning
    - Instance unification
- Flink (the website) is a directory of Semantic Web researchers and their works
  - Browse the network of all authors at ISWC '01-'05
  - Emails, publications
  - Carry out analysis, view statistics
  - Download profiles in FOAF format
  - Download networks for analysis
- Open source





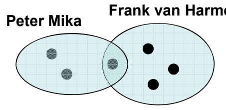
Network mining using search engines

- Given:
  - list of person names
  - parameters
- Algorithm:
  - Filter out persons with two few web pages
  - For each pair of persons
    - Calculate co-occurrence (or average precision
  - Filter again based on tie strength
- Origins:
  - Co-word analysis in bibliometrics
  - Network mining in ReferralWeb

Measures

Jaccard-cooccurence:

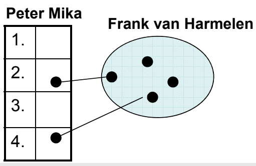
$$\frac{p(A \wedge B)}{p(A \vee B)} = \frac{|A \cap B|}{|A \cup B| - |A \cap B|}$$



Average precision:

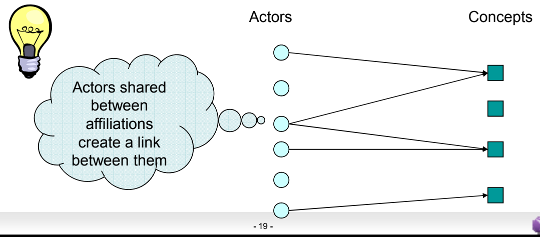
$$P(n) = \frac{\sum_{r=1}^{n} rel(r)}{n}$$

$$P_{ave} = \frac{\sum_{r=1}^{N} P(r) * rel(r)}{N}$$

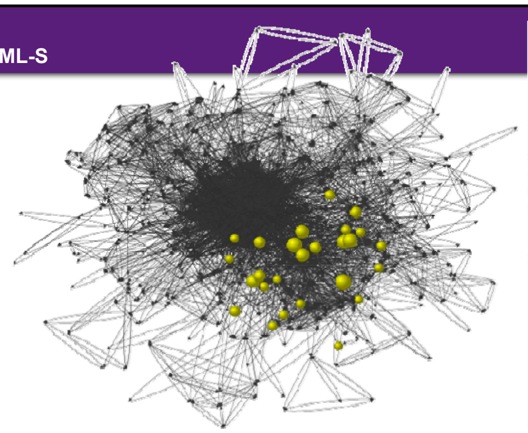


Affiliation network

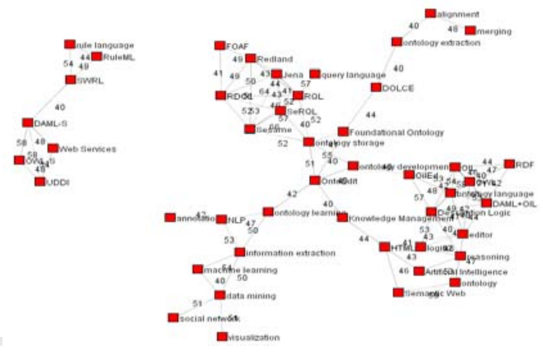
- Bipartite graph (two-mode network)
  - Two sets of nodes, edges run only between nodes



DAML-S



Associations between research topics



Example: identity reasoning

- Source A:
  - Person “F. van Harmelen” is the author of the “Semantic Web Primer”
- Source B:
  - Person “Frank van Harmelen” has the email frankh@cs.vu.nl
- Source C:
  - A person sent an email from frankh@cs.vu.nl to pmika@cs.vu.nl, i.e. they must know each other.
- Conclude: The three Franks are the same person
  - It follows that the author of the Semantic Web Primer knows pmika@cs.vu.nl

Instance matching (smus hing)

- Task: find equivale
- Leibni

All women
until you get to kn
-- the closed

All women are unique
until they turn out to be the same.
-- the open world

$$\forall x \forall y (x = y \leftrightarrow (Px \leftrightarrow Py))$$

- Open vs. closed world
  - OWL: open world, IFPs (max cardinality in general) can lead to sameAs
- Custom reasoner
  - This specific task is poorly supported by DL reasoners
    - Fuzziness, inconsistency
  - Most practical real world rules are outside of DL e.g. Authors of publications are all different

Lessons learned

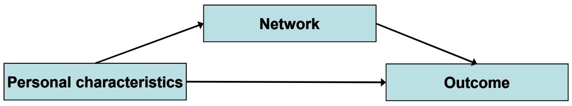
- Quality
  - Social scientists are anxious and rightfully so:
    1. Errors in the extraction of specific cases
       - Syntactic information extraction (Martin Frank, Jim Hendler, Jèrôme Euzenat, York Sure)
       - Analogous to having outlier cases on a questionnaire
    2. General noise
       - Co-occurrence by coincidence
       - Coverage, reliability of the search engine
  - Aggregation, network effects increase robustness
    - See our case study

Lessons learned II.

- Semantic Web technology is a partial match
  - Representation of social relations is difficult
    - Idea: relations as patterns of social interaction
      - P. Mika, Aldo Gangemi: *Descriptions of Social Relations*. 1st Workshop on Friend of a Friend, Social Networking and the (Semantic) Web, Galway, 2004.
  - Equivalence in ontology languages is often too strong
    - Instance unification requires a notion of similarity
  - Missing constructs
- Scalability
  - Addressed by combinations of forward- and backward-chaining reasoning

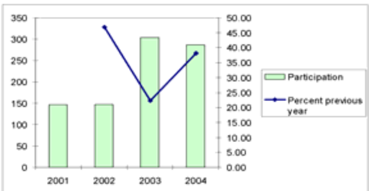
Application: network analysis

- Networks effect substantive outcomes
  - Hypothesis related to the effect of networks on performance
  - Network: features of ego-network, but also position, role
- Research and innovation
  - Outcome: publication performance, patent databases (but also: good ideas)



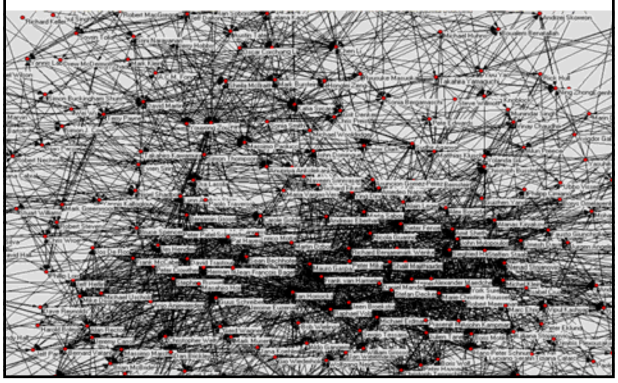
Case study: The Semantic Web community

- Community: the organizers and contributors of SWWS'01, ISWC'02-5 (N=766)
- International, largely academic (79%)

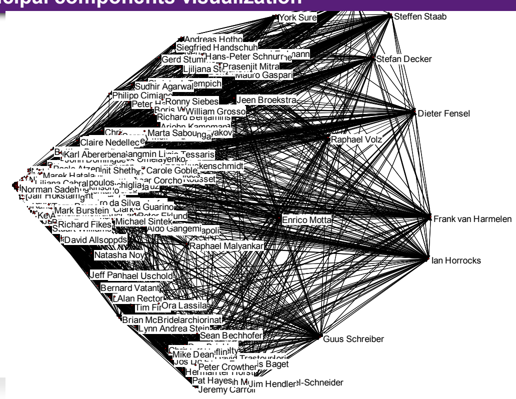


Geographic visualization

Blow-up of the core

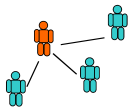


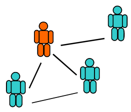
Principal components visualization



Structure

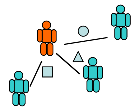
- Access
  – Degree
- Efficiency of access
  – Non-redundancy
- Combination:
  – Structural holes (eff.size)

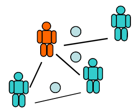




Content

- Access
  – New concepts (ideas)
- Efficiency of access
  – Avg. number of new concepts per degree
  – Avg. number of providers per interest
  – Redundancy in the neighborhood
- Combination:
  – Content holes





Network measures vs. real world status

In-degree		Closeness		Structural Holes		Content Hole		Page count		Top publication	
Name	Value	Name	Value	Name	Value	Name	Value	Name	Value	Name	Value
Steffen Staab	106	Ian Horrocks	0.476	Steffen Staab	92	Steffen Staab	257	Dan Brickley	23000	Dieter Fensel	64
Dieter Fensel	105	Dieter Fensel	0.473	Dieter Fensel	90	Frank van Harmelen	249	pat hayes	21500	Mark Musen	58
Ian Horrocks	104	Frank van Harmele	0.473	Ian Horrocks	88	Dieter Fensel	238	Jim Hendler	21300	Steffen Staab	[illegible]
Frank van Harmele	102	Stefan Decker	0.467	Frank van Harmele	84	Ian Horrocks	232	Dieter Fensel	20600	Rudi Studer	47
Stefan Decker	96	Steffen Staab	0.466	Stefan Decker	77	Stefan Decker	198	Ian Horrocks	19200	Tim Finin	46
Rudi Studer	82	Rudi Studer	0.445	Rudi Studer	65	Rudi Studer	171	Eric Miller	17000	Alexander Maedche	45
Guus Schreiber	71	Guus Schreiber	0.438	Guus Schreiber	54	Enrico Motta	121	Frank van Harmele	16800	Ian Horrocks	43
Raphael Volz	59	Jim Hendler	0.431	Tim Finin	44	Katia Sycara	112	Steffen Staab	15200	Stefan Decker	42
York Sure	54	Enrico Motta	0.430	Katia Sycara	41	Guus Schreiber	107	Ora Lassila	14100	Amit Sheth	[illegible]
Tim Finin	52	Peter F. Patel-Sch	0.429	Raphael Volz	40	York Sure	80	Stefan Decker	13100	Katia Sycara	34
Peter F. Patel-Sch	52	Raphael Malyankar	0.429	York Sure	37	Tim Finin	76	Adam Pease	12300	Wolfgang Nejdl	[illegible]
Enrico Motta	52	Masahiro Hori	0.417	Enrico Motta	37	Richard Benjamins	76	Rudi Studer	11700	Frank van Harmelen	32
Jim Hendler	50	Trastour, David	0.416	Peter F. Patel-Schr	36	Heiner Stuckenschi	72	Tim Finin	11100	Enrico Motta	[illegible]
Katia Sycara	48	Raphael Volz	0.414	Jim Hendler	33	Raphael Volz	68	Mike Dean	10300	Nicola Guarino	30
Dan Brickley	44	Jean Francois Bag	0.414	Dan Brickley	31	Matthias Klusch	57	Guus Schreiber	10000	Gio Wiederhold	29

SWWS, ISWC chair (4)

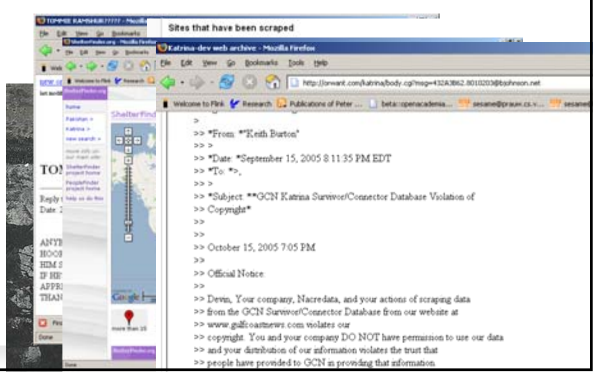
W3C co-chairs (2)

Journal of Web Semantics (4)
IEEE Intelligent Systems (3)

Results

- Cognitive diversity correlates with higher performance beyond the effect of structural diversity
- More details:
  – Peter Mika. *Flink: Using Semantic Web Technology for the Presentation and Analysis of Online Social Networks.* Journal of Web Semantics 3(2), Elsevier, 2005.
  – Peter Mika, Tom Elfring and Peter Groenewegen. *Application of Semantic Technology for Social Network Analysis in the Sciences.* Scientometrics 67:2. Springer, 2006.

Katrina PeopleFinder



2008: New opportunities for research

- More data
  – XFN, FOAF
- Easier access to data
  – Google's Social Graph API
  – OpenSocial, OpenID, OAuth
  – Custom APIs
- Exploring the temporal and spatial dimension of data
  – Change in social networks
  – Social networks mobility

Related streams

- Studies on information diffusion in the blogosphere
- Open Source software communities
- Forums, support groups, UseNet
- Corporate email networks
- Networks of organizations
- Social Networks and Trust
- Social Networks and Recommender Systems
- Analysis of scientific collaboration networks on the Web
- The role of networks in the diffusion of ideas
- Disambiguating personal references on the Web

Related Work

- Lada Adamic and Eytan Adar. How to search a social network. *Social Networks*, 27(3):"187–203", 2005.
- Eytan Adar and Lada A. Adamic. Tracking Information Epidemics in Blogspace. In *Web Intelligence*, Compiegne, France, 2005.
- Daniel Gruhl, Ramanathan V. Guha, David Liben-Nowell, and Andrew Tomkins. Information diffusion through blogspace. In *Proceedings of the 13th International World Wide Web Conference*, pages 491–501, New York, USA, 2004.
- Anjo Anjewierden and Lilia Efimova. Understanding weblog communities through digital traces: a framework, a tool and an example. In *International Workshop on Community Informatics (COMINF 2006)*, Montpellier, France, 2006.
- John C. Paolillo, Sarah Mercure, and Elijah Wright. The Social Semantics of LiveJournal FOAF: Structure and Change from 2004 to 2005. In *Workshop on Semantic Network Analysis (SNA'05)*, 2005.
- Marc A. Smith. Invisible Crowds in Cyberspace: Measuring and Mapping the Social Structure of USENET. In Marc Smith and Peter Kollock, editors, *Communities in Cyberspace*. Routledge Press, London, 1999.
- Derek J. deSolla Price. Networks of scientific papers: The pattern of bibliographic references indicates the nature of the scientific research front. *Science*, 149(3683):510–515, 1965.

Related Work

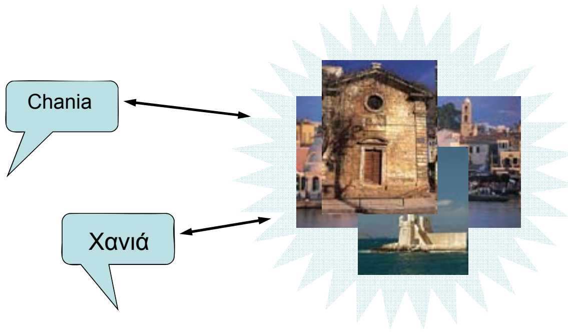
- A.L. Barabási, H. Jeong, Z. Néda, E. Ravasz, A. Schubert, and T. Vicsek. Evolution of the social network of scientific collaborations. *Physica A*, 311(3-4):590–614, 2002.
- Danushka Bollegala, Yutaka Matsuo, and Mitsuru Ishizuka. Disambiguating Personal Names on the Web using Automatically Extracted Key Phrases. In *Proceedings of the 17th European Conference on Artificial Intelligence*, 2006.
- Ronald S. Burt. Structural Holes and Good Ideas (in press). *American Journal of Sociology*, 110(2), 2004.
- Jennifer Golbeck and James Hendler. FilmTrust: Movie recommendations using trust in web-based social networks. In *Proceedings of the IEEE Consumer Communications and Networking Conference*, 2006.
- Yutaka Matsuo, Masahiro Hamasaki, Hideaki Takeda, Junichiro Mori, Danushka Bollegara, Yoshiyuki Nakamura, Takuichi Nishimura, Koiti Hasida, and Mitsuru Ishizuka. Spinning Multiple Social Networks for SemanticWeb. In *Proceedings of the Twenty-First National Conference on Artificial Intelligence (AAAI2006)*, 2006.
- Joshua R. Tyler, Dennis M. Wilkinson, and Bernardo A. Huberman. Email as spectroscopy: automated discovery of community structure within organizations. In *International Conference on Communities and Technologies*, pages 81–96, Deventer, The Netherlands, 2003. Kluwer, B.V.

The Social Side of Semantics

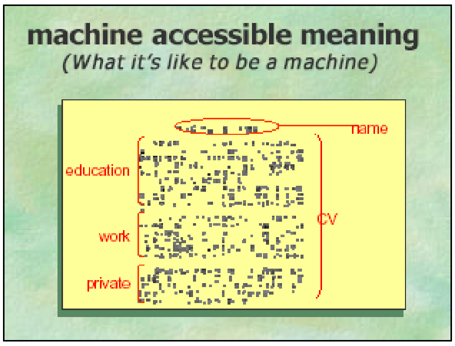
The classic approach to the Semantic Web

- Machines don't understand the Web
- We will annotate it for them using ontologies
  - Ontologies are manually crafted artifacts created by knowledge engineers by acquiring and formalizing the knowledge of experts
- This allows computers to understand the Web's content
  - Interoperability is granted if everyone follows the agreement
- We can search, classify, analyze, predict, reason with the Web's content

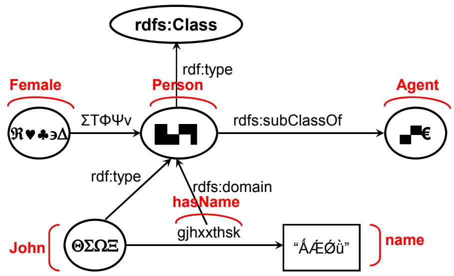
Semantics (Tarski)



What it's like to be a machine?

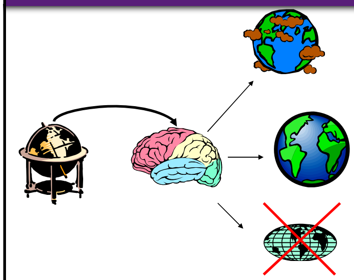


What it's like to be a machine?



(rdfs : subClassOf rdfs : domain rdfs : Class);
$(xxx\ aaa\ yyy) \wedge (aaa\ \mathrm{rdfs:domain}\ zzz) \rightarrow (xxx\ \mathrm{rdf:type}\ zzz)$

The notion of a Universe

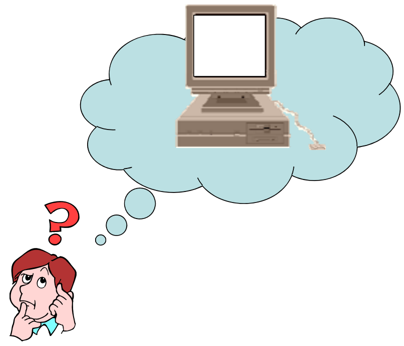


…but the context defines what is the set of possible worlds to start with!

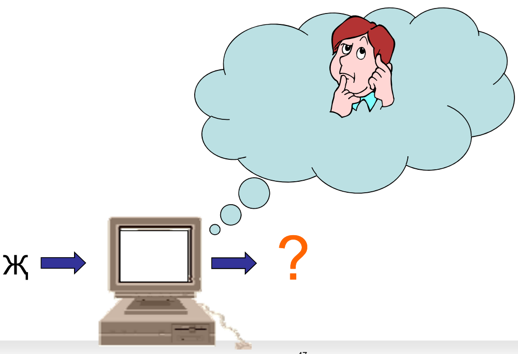
After mapping elements of the model to cognitive schema…

…ontological descriptions rule out possible models of the world…

What it's like to be a computer?



What it's like to be a human?



What it's like to be a human? An Exercise

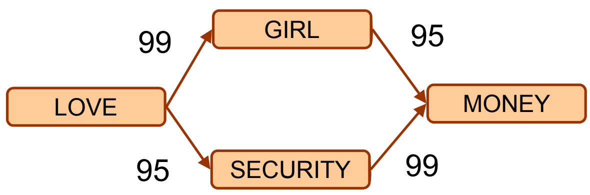
Fill the blank!

LOVE → #^&#)^"/ → MONEY

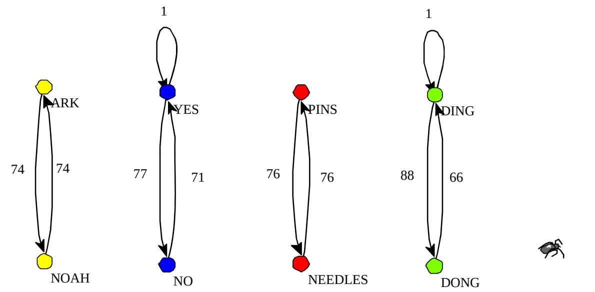
Edinbourgh Associative Thesaurus (EAT)

- Experiment
  - 1973, Edinbourgh university students
  - Participants asked to look at a word (stimulus) and write down the first word it made them think of (response)
  - Responses were then reused as stimuli in the next round of the experiment
  - Stopped when too many words have been accumulated
  - Network encoding: 23219 vertices, 325624 arcs

EAT answers

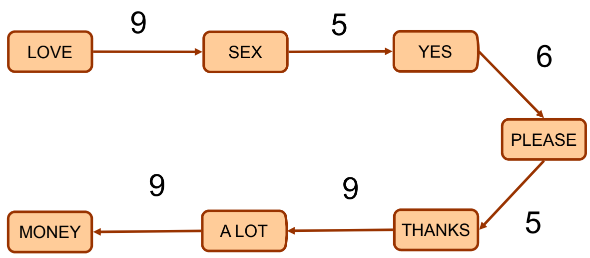


Valued core



Courtesy of Ales Ziberna:Network Analysis of an Associative Thesaurus. Sunbelt, 2004

The shortest paths from "love" to "money" on network without lines with weights below 5



Important concepts

Vertex	Sum of inward line weights	Vertex	Sum of inward line weights
MONEY	4387	TREE	2019
WATER	3299	GOOD	1988
FOOD	2918	HOUSE	1972
ME	2515	BIRD	1896
MAN	2435	UP	1891
CAR	2434	CHURCH	1881
SEA	2224	TIME	1802
SEX	2154	FIRE	1795
HORSE	2100	SHIP	1762
DOG	2073	MUSIC	1722

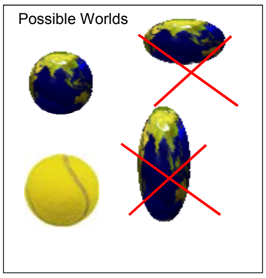
Problem*-1

- Knowledge is situated
  - Interpretation by association is context-dependent, not absolute
  - Acknowledged by RDF Semantics
  - Holds for ontologies: repeat the EAT experiment with a different community!
- A large part of this context is the social context
  - The original community where the ontology was created and in which it's directly interpretable
  - As in *Def. ontology: shared, formal representation of the conceptualizations of a community*

Required: incorporating the social context into the model of ontologies

Formal semantics

- Universe, Interpretation
- Entailment independent of interpretation
- *However, the remaining set of possible worlds* is dependent on the interpreter



Why bother?

- Ontology (re)interpretation is at the core of the ontology mapping problem
- Even if we don't transfer ontologies across domains, they still suffer from ontology drift
- The kind of associative knowledge contained in EAT is missing from current linguistic, philosophical ontologies

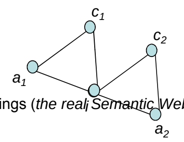
Idea

Start simple!

1. A graph model of ontologies based on tripartite graphs of actors, concepts and instances
   - An extension of the current ontology models with an explicit account for agents
   - A *social-semantic* network
2. Emergent semantics
   - General idea: observe semantics in the way agents interact (use concepts)
     - Bottom-up ontologies
     - Semantics = syntax + statistics

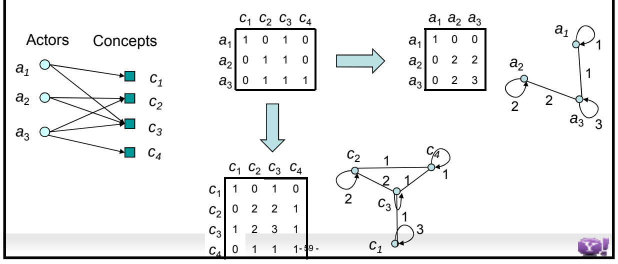
Representation

- Tripartite graph with hyperedges
  - Edge ~ an actor associates a concept with a certain instance
- Analogy: folksonomies
  - Actors: users
  - Concepts: tags
  - Instances: objects
  - Example: del.icio.us, Flickr, 43Things (*the real Semantic Web out there*)



Emergence

1. Create three weighted bipartite graphs (affiliation networks)
2. Dichotomize
3. Fold each bipartite graph into two (weighted) regular graphs
4. Normalize (e.g. Jaccard-coefficient)
5. Filter



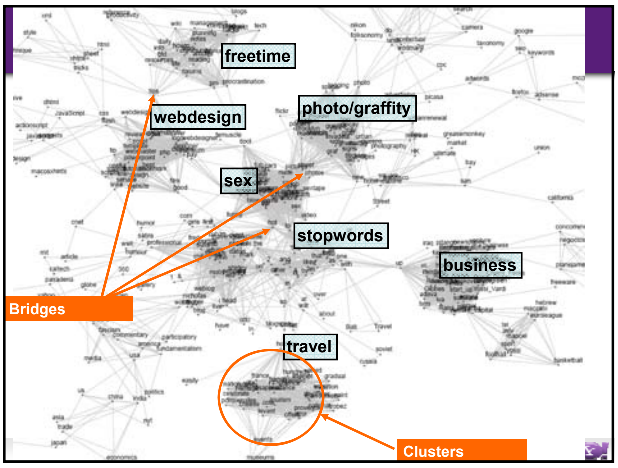
Outcome

- Two of the resulting graphs represent associations between concepts
  1. In the O_{ci} network concepts represent sets of items, relationships are based on overlaps in item-sets
     - Ignores actors
  2. In the O_{ac} network concepts represent sets of actors(!), relationships are based on overlaps in communities of actors
     - Ignores items

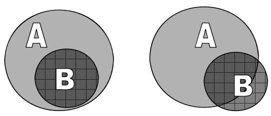
While the first is familiar, the second is something new.

Case study 1: del.icio.us

- Social bookmarking application
  - Technology aware web community
  - 30k users in December, 2004
  - Latest items made available through RSS
- Dataset:
  - ~52 k unique annotations
  - ~30 k URLs
  - ~10 k users
  - ~30 k unique tags
- "Messy" data
  - Ambiguity
  - Multiplicity (synonyms, multi-lingual)
  - Entry limitations

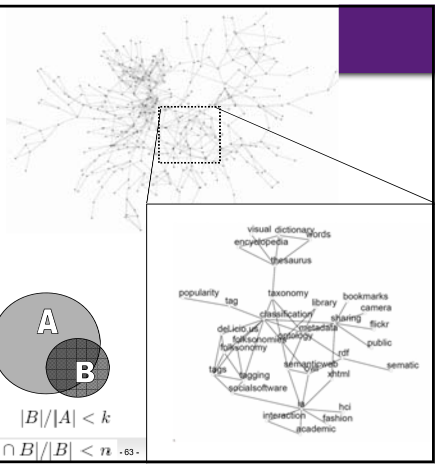


Broader	Narrower
rss	atom
cmyk	rgb
cell	umts, wcdma, ev-do
phone	cell
ajax	json
xml	xslt
rdf	owl
flickr	gmail, picasa
ruby	rails
mac	iphoto
java	j2ee
google	gds
search	a9, engine
linux	ubuntu, gnome
flash	actionscript
flickr	liekr, photoset
javascript	xmlhttprequest, dom, sarissa



$|B|/|A| < k$

$|A \cap B|/|B| < n$



Results

- When looking at co-occurrence of terms (O_{ci})
  - Network reflects language use
  - Better for clustering, determining ambiguity of terms and finding synonyms
- When looking at community overlaps (O_{ac})
  - Network reflects the domain
  - Better for finding broader/narrower terms, non-trivial relationships

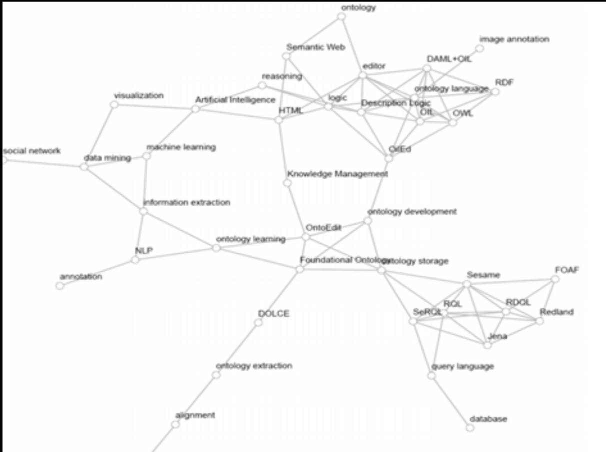
Remember: in the second case
it doesn't matter whether the concepts are used on the same items
(or how many items are classified under a concept)

Case study 2: Community-based ontology extraction

- Idea: turn it into an ontology extraction technique!
- Application of the model to the Web:
  - Actors: members of some community
  - Concepts: set of pre-selected concepts
  - Items: web pages
- Obtaining O_{ac}
  1. Associate actors with concepts (Google)
  2. Apply folding
- Obtaining O_{ci}
  1. Associate concepts with concepts (Google)

Evaluation

- Community
  - ISWC authors (N=706)
    - flink.semanticweb.org
  - List of 60 terms selected from ISWC proceedings
- E-mail survey
  - 30+ AI researchers – most of them members of the community
  - *In terms of the associations between the concepts, which ontology of Semantic Web related concepts do you consider more accurate?*



Results

- Findings:
  1. O_{ac} is considered more representative than O_{ci}
  2. Those in the community agree more than those outside
  3. Those in the (theoretical) core of the Semantic Web community agree even more!

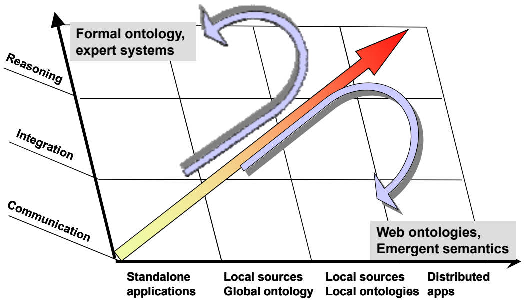
- Note: not a (simple) disambiguation effect.

	N	O_{ac}	O_{ci}	Ratio	Sign.
All	30	22	8	73.3%	0.0055
ISWC	23	18	5	78.3%	0.0040
ISWC-core	15	13	2	86.7%	0.0032

Summary: An alternative to the classic way to semantics

- Logic is a useful tool in capturing semantics but not enough
  - Logic alone cannot capture meaning no matter how powerful the language is
  - Ontology = logic plus social agreement (commitment)
    - The agreement provides the grounding
  - Web ontologies and web ontology languages are typically very weak due to the scale of the Web
- But is it necessary to agree in advance? It turns out, machines can learn agreements.
  - Emergent Semantics: learning semantics based on the usage of symbols
  - Semantics = syntax + statistics

Meta-summary



Peter Mika and Hans Akkermans. *Towards a new synthesis of ontology technology and knowledge management.* Knowledge Engineering Review 19(4), Cambridge University Press, 2005.

Related work

- Analysis, modelling:
  - Golder, S. and Huberman, B. A. (2006) Usage patterns of collaborative tagging systems. Journal of Information Science, 32(2):198--208.
  - Ciro Cattuto, Vittorio Loreto and Luciano Pietronero. Semiotic Dynamics and Collaborative Tagging PNAS 104, 1461 (2007)
  - Other works by the European TAGORA project (tagora-project.eu)
- Applications in (mm) search, recommendation, spam detection:
  - Challenges in Searching Online Communities. Amer-Yahia, Sihem ; Benedikt, Michael ; Bohannon, Philip, IEEE Data Eng. Bull., 2007
  - X. Wu, L. Zhang, and Y. Yu, "Exploring social annotations for the semantic web," in *WWW '06: Proceedings of the 15th international conference on World Wide Web.* New York, NY, USA: ACM Press, 2006, pp. 417-426.
- Related streams:
  - Query log analysis (query graphs) in IR
  - Ontology learning by natural language processing

Multimedia Semantics

Roelof van Zwol
roelof@yahoo-inc.com
Yahoo! Research Barcelona

Multimedia Research

- Goal:
  - Deploy collective knowledge present in social media properties to provide a better user (search) experience.
- Focus:
  - **Media Interaction**: creating the incentives for users
  - **Media Mining**: extracting knowledge from user generated content
  - **Media Search**: enhancing the user experience through novel search assistants, recognizing visual concepts, and offering diversity in search results for ambiguous topics.

Media Interaction

Flickr: Who's Looking?

Video Tag Game

Flickr: Who is Looking?

Roelof van Zwol

ACM Web Intelligence
November 2007

About Flickr

- On-line photo sharing service
- > 2 Billion photos uploaded
- > 8.5 Million Web-users registered
- > 2.500 photos uploaded per minute
- > 12.000 photos served per second, at peak times

Who is looking?

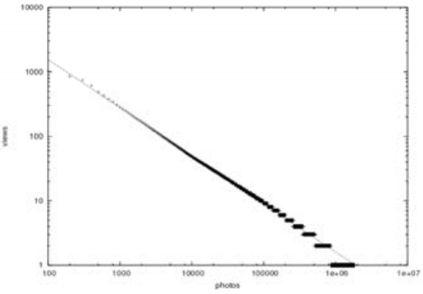
- A characterisation of usage behaviour on Flickr, with focus on:
  - **When?**
    - **Temporal characteristics**
  - Who?
    - Social
  - Where?
    - Spatial
- Not about "why do we tag?"
  - Social incentives
    - G.W. Furnas et al. "Why do tagging systems work?"
    - C. Marlow et al. "Ht06, tagging paper, taxonomy, flickr academic article, to read"
    - M. Ames and M. Naaman. "Why we tag: motivations for annotation in mobile and online media"

Data Collection

- Analysis is based on:
  - **HTTP access logs of Flickr, spanning a 60 day period**
    - 1.83 Million public photos
      - uploaded in the first 10 days
      - and their views in the consecutive 50 days
    - limited to the detailed photo views on Flickr:
      - .*flickr.com/photos/<owner id>/<photo id>/?.*
  - **Data collected through public Flickr API:**
    - flickr.photos.getInfo
    - flickr.photos.getAllContexts
    - flickr.contacts.getPublicList
  - **Mapping service from IP to long/lat coordinates**

Characterisation of Photo Views

- 1.83 million photos; 6.72 million views
- Power law - the probability of having x visits is proportional to $x^{-0.7}$

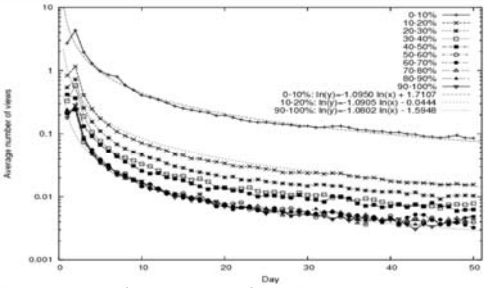


Characterisation of Photo Views

slice	views
0-10%	3,802,875
10-20%	812,131
20-30%	515,532
30-40%	365,712
40-50%	312,270
50-60%	182,856
60-70%	182,857
70-80%	182,856
80-90%	182,857
90-100%	182,857

- Dividing the collection into equal **slices**, based on the number of photos
- Where slice 0-10% contains the top 10% most frequently viewed photos
- Emphasize on the skewedness of the distribution of photo views: 0-10% slice already covers >50% of all views

Characterisation of Photo Views



- **The average number of photo views per day for the *slices* over a 50 day period.**
  - The declining trend followed by each of the slices can be modelled by an exponential decay
  - The number of views on day x after being uploaded is proportional to $e^{-1.1x}$

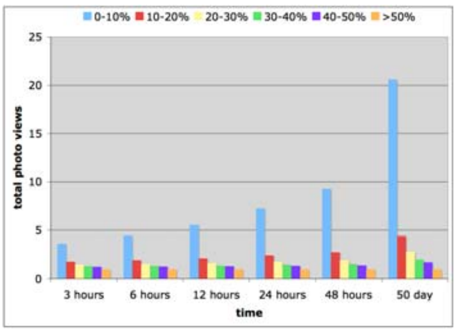
Characterisation of Photo Views

- Focus on first 48 hours
- Shows similar behaviour for different trends (slices)
- After 48 hours, a photo already received ~50% of the total number of views it will receive after 50 days
- Moreover, popular photos are already discovered within 3 hours after being uploaded

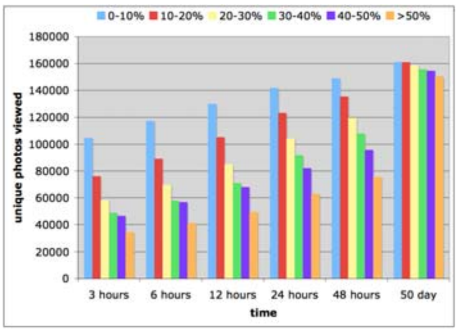
	3 hours		6 hours		12 hours	
slice	avg	std	avg	std	avg	std
0-10%	3.63	8.25	4.44	12.51	5.55	18.66
10-20%	1.77	0.97	1.92	1.05	2.11	1.12
20-30%	1.47	0.67	1.54	0.7	1.62	0.73
30-40%	1.3	0.46	1.33	0.47	1.36	0.48
40-50%	1.25	0.43	1.27	0.44	1.3	0.46
>50%	1	0	1	0	1	0

	24 hours		48 hours		50 day	
slice	avg	std	avg	std	avg	std
0-10%	7.24	26.5	9.28	37.6	20.6	87.7
10-20%	2.43	1.22	2.75	1.28	4.4	0.7
20-30%	1.77	0.77	1.92	0.79	2.8	0.4
30-40%	1.44	0.5	1.52	0.5	2	0
40-50%	1.35	0.48	1.39	0.49	1.7	0.45
>50%	1	0	1	0	1	0

Characterisation of Photo Views



Characterisation of Photo Views



Applications

- What can you do with this knowledge?
  - **Predict** the popularity of a photo (using temporal, and social indicators)
  - Develop **caching** strategies for frequently viewed media content
  - Develop a hybrid model for serving multimedia content that implements a **P2P storage strategy** for *in-frequently* viewed content, in combination with a **content distribution network** for serving *popular* media content

by: thepres6

Video Tag Game

Roelof van Zwol, Lluis Garcia, Georgina Ramirez, Borkur Sigurbjornsson

World Wide Web conference, April 2008

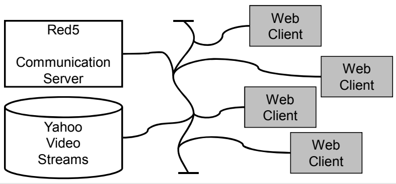
Public launch: Q3 2008

About & Motivation

- About
  - Time-based annotation of streaming video, in a multi-player game
- Motivation
  - To collect dense, time-based annotations of video
  - Investigate users accuracy when tagging streamed video
  - Enable retrieval of video-fragments

How?

- Set-up
  - In a multi-player game setting
  - Tagging of streaming video
  - Temporal scoring mechanism, that rewards tag-agreement between users
- Architecture



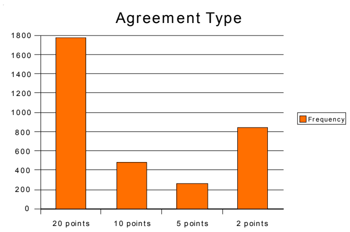
Video Tag Game

- Temporal Scoring Mechanism:
  - If two players agree on a tag, the players get points
  - More points should be rewarded for a tag if the difference in time between two players, submitting that tag, is smaller
  - Entering the same tag twice within a short period of time should not be rewarded (for that user, others can however benefit)

Video Tag Game

Video Tag Game

- SIGIR demo:
  - 27 games / 59 players / 5890 tags
  - 0.57 agreement (3360 scoring tags), on avg. 12.88 points per agreed tag:



Media Mining

Flickr Tag Recommendation based on Collective Knowledge

Resolving Tag Ambiguity

Syntactic Classification of Tags

Flickr Tag Recommendation based on Collective Knowledge

Borkur Sigurbjornsson, and Roelof van Zwol

World Wide Web conference, April 2008

Motivation

- I went to Barcelona, took a photo, tagged it:
  "Sagrada Familia"
- 2 years later I want to find the photo
  - query: church barcelona gaudí
  - 0 pictures found
- Task:
  - Help users to provide rich annotations

Flickr Annotations

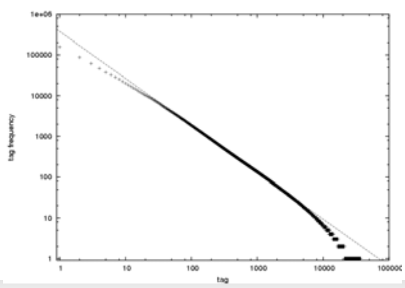
- Characteristics:
  - Most photos have few tags
  - Few photos have many tags

Tags per photo	Percentage of photos[1]
1	30%
2-3	34%
4-6	23%
> 6	13%

[1] based on a random sample of 100 million tagged Flickr photos

Flickr Tag Frequency

- Few tags are used to describe many photos
- Most tags are used to describe few photos



Collective Knowledge

- Many users annotate photos of "La Sagrada Familia":
  - Sagrada Familia, Barcelona
  - Sagrada Familia, Gaudi, architecture, church
  - church, Sagrada Familia
  - Sagrada Familia, Barcelona, Spain
- Derived collective knowledge:
  - Barcelona, Gaudi, church, architecture

Tag Co-occurrence Statistics

- Input: A snapshot of 100M public photos on Flickr, with annotations
- Approach is based on probabilistic framework
  - Assume an photo is labelled with a set of tags $T = \{t_a, t_b, ...\}$
  - Define $I(T)$ as the number of photos that contain the tag set T
  - For any pair of tags t_i, t_j, we denote the number of image co-occurrences by $I(t_i \cap t_j)$
  - Estimate the probability that a tag, t_i, appears in presence of tag t_j, by calculating:

$$p(t_i|t_j) = \frac{I(t_i \cap t_j)}{\sum_k I(t_k \cap t_j)}$$

  - Examples:
    - P(barcelona|sagradafamilia) = 0.46
    - P(sagrada familia|gaudi) = 0.14

Tag Co-occurrence Statistics

- Probabilistic framework cont'd:
  - Estimate the probability that any one tag is used on an image by:

$$p(t_i) = \frac{\sum_j I(t_i \cap t_j)}{\sum_{j,k} I(t_k \cap t_j)}$$

  - Objective is to calculate the probability of a tag in any context, e.g. a set of tags T:

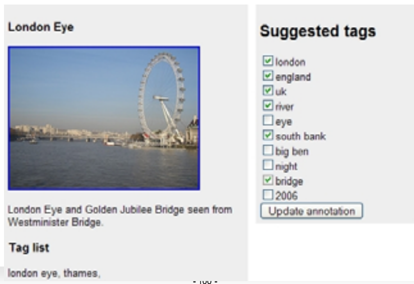
$$p(T|t_i) = \prod_{t \in T} p(t|t_i)$$

$$p(t_i|T) = \frac{p(T|t_i)p(t_i)}{p(T)} = \frac{p(t_i)\prod_{t\in T} p(t|t_i)}{\sum_j p(t_j)\prod_{t \in T} p(t|t_j)}$$

    - P(Sagrada Familia | {church, Barcelona})=0.67

Tag Recommendation System

- Task: Given a partially annotated photo, recommend additional annotations
- Approach: Use the aggregated annotation term co-occurrence



Summary

- Tagging is sparse but diverse
  - Few tags per photo
  - Tag frequency distribution follows a power law
- Use the collective knowledge to recommend tags
  - For 68% of photos our first suggestion is good
  - For 94% of photos we provide a good suggestion among top 5
  - For top 5 suggestions, 54% are good
- Future work
  - Use additional data sources (User profile, social contacts)
    - TagSuggest 2.0P
  - Use light weight image features

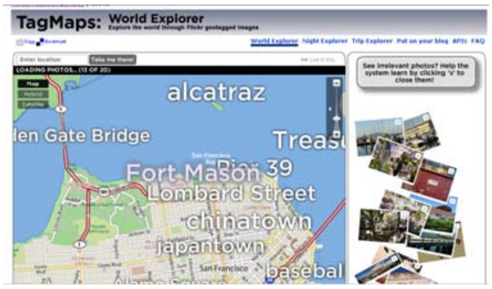
Related work: temporal tag extraction

- Interestingness of a tag within a time window
  - TF-IDF like measure
- Efficient computation for interactive visualization
- http://research.yahoo.com/taglines/

M. Dubinko, R. Kumar, J. Magnani, J. Novak, P. Raghavan, and A. Tomkins. Visualizing tags over time. In Proc. WWW, pp. 193–202. ACM, 2006.

Related work: spatial tag extraction

- http://tagmaps.research.yahoo.com/



Alexander Jaffe, Mor Naaman, Tamir Tassa, Marc Davis. Generating Summaries and Visualization for Large Collections of Geo-Referenced Photographs. *In proceedings* The 8th ACM SIGMM international workshop on Multimedia information retrieval (MIR '06), Santa Barbara, CA, USA, 2006.

- 103 -

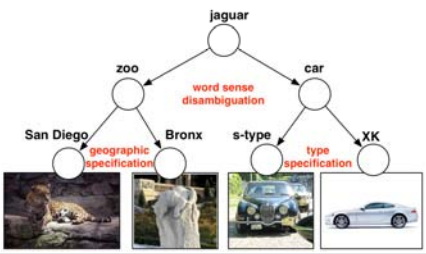
Resolving Tag Ambiguity

Malcolm Slaney, Kilian Weinberger, and Roelof van Zwol

ACM Multimedia
November 2008

Resolving Tag Ambiguity

- The objective of this research is to determine when additional tags are needed. Two scenarios:
  1. A tag set has an **ambiguous** meaning
  2. The tag set is not sufficiently **specific**



- 105 -

Resolving Tag Ambiguity

- Two contributions:
  1. A statistical approach is proposed to measure the ambiguity of a tag set, and the user is only interrupted, when the ambiguity score is above a certain threshold
  2. The method introduces pair wise disambiguation to recommends two tags that would reduce the ambiguity of the existing tag set the most

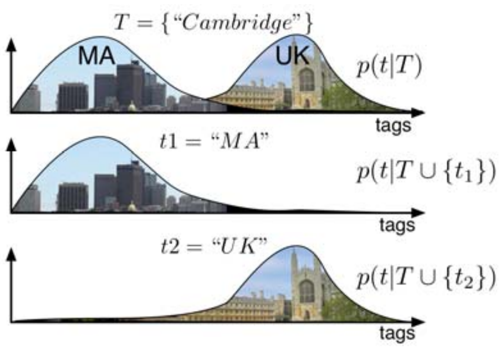
- 106 -

Resolving Tag Ambiguity

- Intuition:
  - A tag set is ambiguous if it can appear in two different tag contexts
    - Geographic locations, time-based events, languages, topical, social, or any combination of the mentioned contexts ("Java": location, programming language, coffee, etc.)
  - Example: "Cambridge"
    - Considered ambiguous, based on spatial context
    - Tag suggestions: "Massachusetts" or "United Kingdom"
    - Alternative tag suggestion "university" is highly relevant, but will not resolve the ambiguity.
- Approach:
  - Extends the probabilistic framework of TagSuggest, and uses a *weighted symmetric KL divergence* for detecting pairs of tags that have the largest impact on reducing the ambiguity

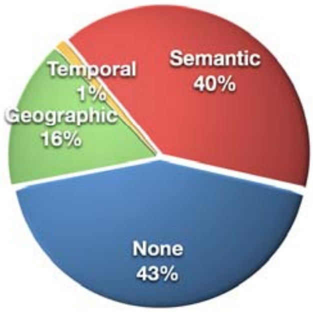
- 107 -

Resolving Tag Ambiguity



- 108 -

Results



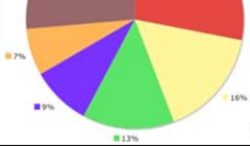
Classifying Flickr Tags
using Open Content Resources

Roelof van Zwol, Borkur Sigurbjornsson, Simon Overell

GeoClass: Identifying Geo-related Content
Joint development with Y!Geo team

Syntactic Classification

- Objective: syntactic classification of tags using open source content (Wordnet, Wikipedia, ODP, etc.)
- Assign tag semantics using WordNet broad categories

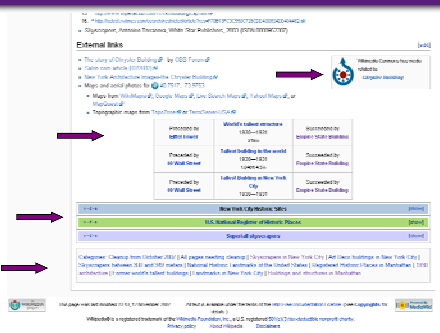


  - Paris :: location
  - Eiffel Tower :: artifact
  - Coverage: 52% of tag volume

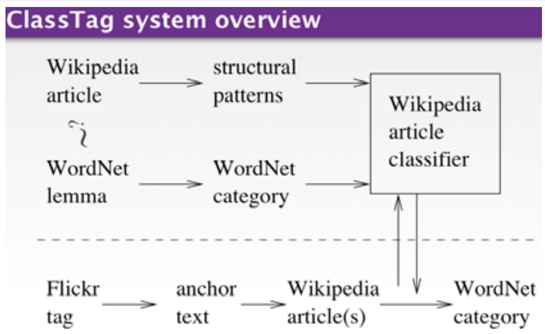
How…

- To extend coverage of syntactic classification?
  - Based on classification of Wikipedia pages
  - Mapping from tags to classified wikipedia pages
  - Upperbound for coverage: 78.6% of the tag volume
- How to classify wikipedia pages?
  - Use structural patterns found in Wikipedia pages
    - templates and categories
  - Achieved extended coverage: 68% of the tag volume

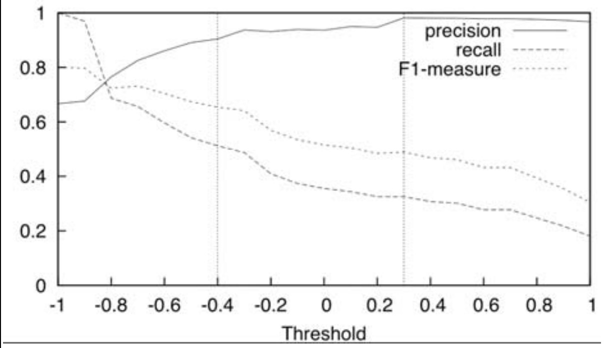
Example



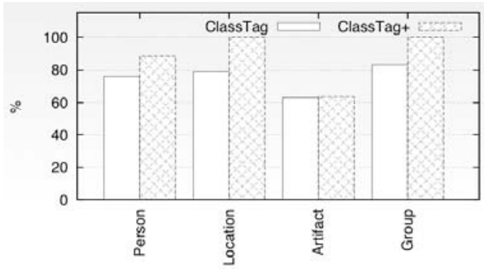
System



Performance



TagClass



REST-API

<tagclass tag="iwo jima">
<classification source="wordnet" class="location" instanceof="island" rank="1" />
<classification source="wordnet" class="act" instanceof="amphibious assault" rank="2"/>
<classification source="wikipedia" class="location" rank="1" support="0.80"/>
<classification source="wikipedia" class="act" rank="2" support="0.10"/>
<classification source="wikipedia" class="artifact" rank="3" support="0.07"/>
</tagclass>

<tagclass tag="bigapple" >
<classification source="wikipedia" class="location" rank="1" support="0.79"/>
<classification source="wikipedia" class="act" rank="2" support="0.20"/>
</tagclass>

Media Search

TagExplorer

Diversifying Image Search Results

TagExplorer

Borkur Sigurbjornsson, and Roelof van Zwol

Public launch: Q3 2008

Motivation

- Tag cloud: a visual depiction of user-generated tags used typically to describe the content of web sites.

Motivation

- Limitation of tag clouds
  - Only work at a collection level or on individual tags, not at level of tag sets
  - Lacks all structure
- Innovation by TagExplorer
  - Exploits *tag co-occurrence*, to enable the user to explore a tag space
  - Provides *semantic break-up* to facilitate human interpretation

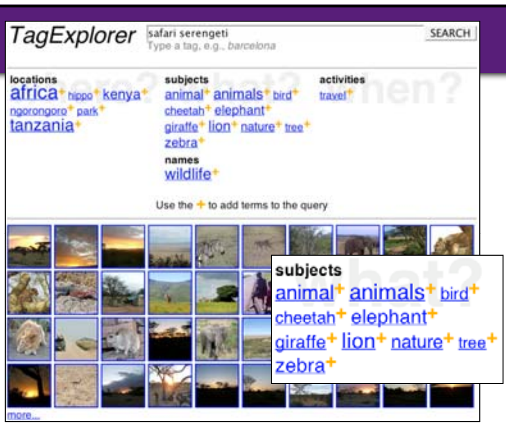
Approach

- Combines:
  - Tag semantics
    - Dual level
      - Where?, When?, What?
      - Nouns in WordNet broad categories
        - » Location, artifact, activities, event, person, group, etc.
    - Other schemas can be applied.
  - Tag co-occurrence analysis
    - For a given set of tags - a keyword based query - a set of related tags is derived.

Approach

Approach

- Combines:
  - Tag semantics
    - Dual level
      - Where?, When?, What?
      - Nouns in WordNet synset
        - » Location, artifact, activities, event, person, group, etc.
    - Other schemas can be applied.
  - Tag co-occurrence analysis
    - For a given set of tags - a keyword based query - a set of related tags is derived.



Applications

- Facilitate “endless browsing” concept
- Search Assistant
  - General search
    - Dual level: “Where?, When?, What?” and WordNet broad categories
  - Vertical search
    - For example video.yahoo.com:
      - Action, Art & Animation, Entertainment & TV, Food, Games, How-To, etc.

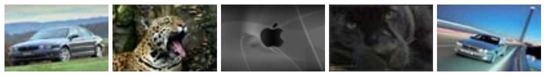
Diversifying Image Search Results

Roelof van Zwol, Vanessa Murdock, Lluis Garcia, Georgina Ramirez, Reinier van Leuken

ACM Multimedia Information Retrieval
October 2008

Dimensions of Diversity

- **Topical diversity**
  Query: "Jaguar"



- **Visual diversity**
  Query: "Jaguar X-type"

- Other dimensions: spatial, temporal, social

Topical Diversity

- Collection: 6M public photos from Flickr
  - Title, description and tags
- Retrieval models
  - Query Likelihood (full index, tags only)
  - Relevance model (full index, tags only, dual index)
- Topics
  - 95 topics extracted from Flickr search logs
  - 25 ambiguous topics

Topical Diversity

- Blind pooling, 51.000 images judged for relevance.
- Two step assessment:
  - Binary relevance judgement
  - Sense classification
- Measured inter- assessor agreement for 20% of topics
  - >85% for all topics
  - most topics >90%

Topical Diversity

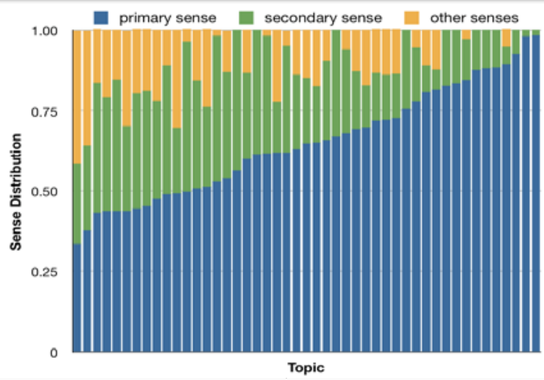
- Retrieval performance
  - Unambiguous topics

Model	P@1	P@5	P@10	P@15	P@20	P@25	P@50
Query Likelihood	0.747	0.733	0.733	0.719	0.709	0.701	0.667
Query Likelihood (Tags Only)	**0.779**	0.749	0.720	0.712	0.703	0.700	0.673
Relevance Model	0.758	0.743	0.720	0.708	0.706	0.699	0.677
Relevance Model (Tags Only)	**0.779**	0.726	0.717	0.719	0.714	0.710	**0.683**
Relevance Model (Dual Index)	0.768	**0.754**	**0.739**	**0.726**	**0.719**	**0.716**	0.680

  - Ambiguous topics

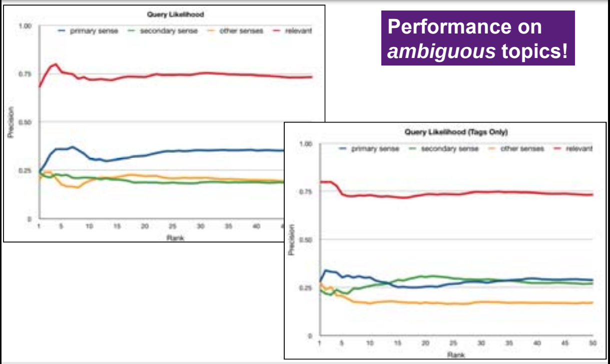
Model	P@1	P@5	P@10	P@15	P@20	P@25	P@50
Query Likelihood	0.680	0.760	0.720	0.725	0.734	0.744	0.734
Query Likelihood (Tags Only)	0.800	0.736	0.732	0.720	0.736	0.736	0.734
Relevance Model	0.720	0.760	**0.768**	**0.784**	**0.788**	**0.792**	**0.778**
Relevance Model (Tags Only)	**0.840**	0.728	0.744	0.741	0.756	0.752	0.735
Relevance Model (Dual Index)	0.720	**0.776**	**0.768**	0.755	0.754	0.760	0.763

Topical Diversity



Topical Diversity

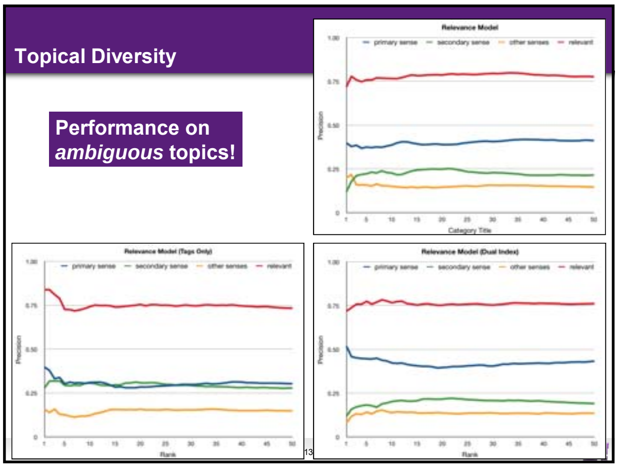
Performance on *ambiguous* topics!



- 133 -

Topical Diversity

Performance on *ambiguous* topics!



Boosting Image Retrieval through Aggregating Search Results based on Visual Annotations

Ximena Olivares, and Roelof van Zwol

ACM Multimedia
November 2008

Image Object Retrieval

- Visual annotations in Flickr

- 136 -

Content-based retrieval (step by step)

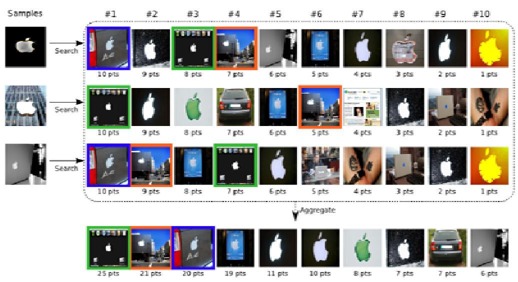
- Extracting visual features from image
  - ≈ words in a document

- K-means clustering of the visual features.
  - ≈word stemming
- Eliminate large clusters
  - ≈stopword removal
- Apply vector space model
- Spatial coherence filter

- 137 -

Image Object Retrieval

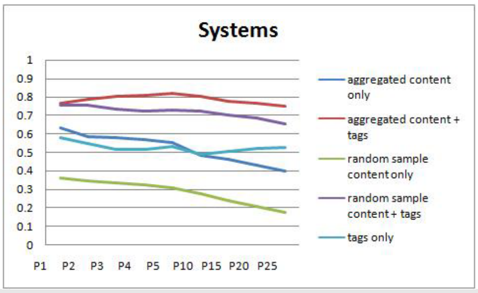
- Rank aggregation, using the visual annotations



- 138 -

Retrieval Performance

- Retrieval performance



Publications

R. van Zwol. Flickr: Who is looking? ACM Web Intelligence 2007.

R. van Zwol, Lluis Garcia, G. Ramirez, B. Sigurbjornsson, and M. Labad. Video Tag Game. WWW 2008.

B. Sigurbjornsson and R. van Zwol. Flickr tag recommendation based on collective knowledge. WWW 2008.

M. Slaney, K. Weinberger, and R. van Zwol. Resolving tag ambiguity. ACM Multimedia 2008

R. van Zwol, V. Murdock, L. Garcia, and G. Ramirez. Diversifying image search with user generated content. In ACM MIR 2008.

X. Olivares, M. Ciaramita, and R. van Zwol. Boosting image retrieval through aggregating search results based on visual annotations. ACM Multimedia 2008

Questions?

http://photosoup.games.yahoo.com/