# One Tag to Bind Them All: Measuring Term Abstractness in Social Metadata

*Dominik Benz*[1], Christian Körner[2], Andreas Hotho[3], Markus Strohmaier[2], Gerd Stumme[1]

*[1]Knowledge and Data Engineering Group (KDE), University of Kassel, Germany*

*[2]Knowledge Management Institute and Know Center, Graz University of Technology, Austria*

*[3]Data Mining and Information Retrieval Group University of Würzburg, Germany*

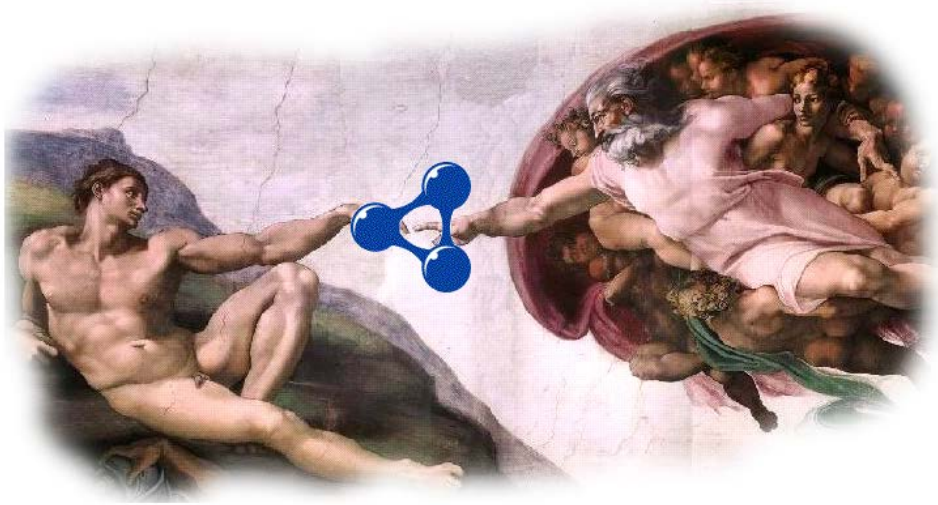
# A cat? Felis silvestris cretensis? An animal?

# A cat? Felis silvestris cretensis? An animal?

- **Abstractness** of terms is crucial to many knowledge management applications
- Manual encoding (e.g. in taxonomies) is expensive

# A cat? Felis silvestris cretensis? An animal?

- **Abstractness** of terms is crucial to many knowledge management applications
- Manual encoding (e.g. in taxonomies) is expensive



# A cat? Felis silvestris cretensis? An animal?

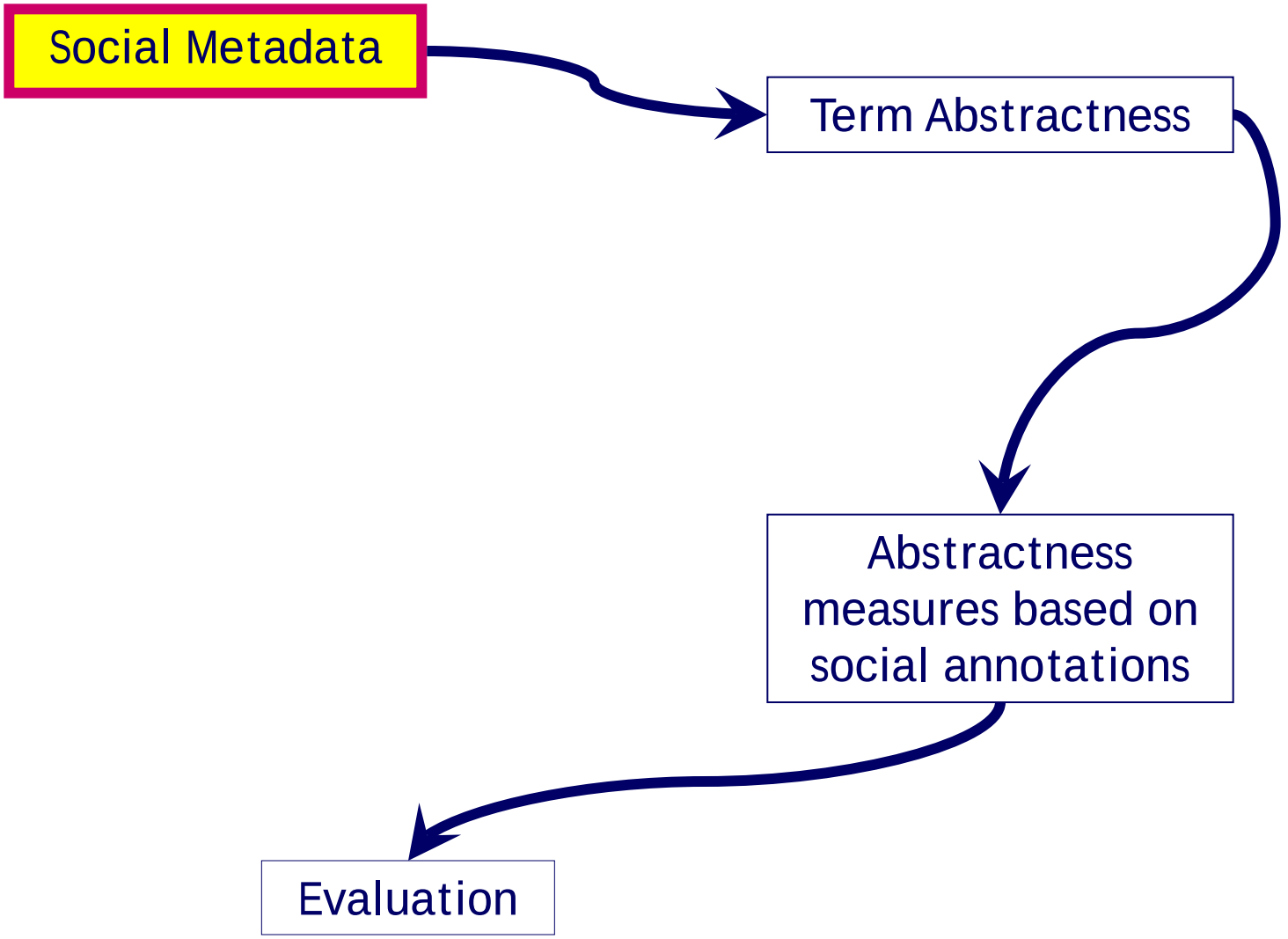
- **Abstractness** of terms is crucial to many knowledge management applications
- Manual encoding (e.g. in taxonomies) is expensive


# A cat? Felis silvestris cretensis? An animal?

- **Abstractness** of terms is crucial to many knowledge management applications
- Manual encoding (e.g. in taxonomies) is expensive


➔ **Can term abstractness be inferred from social metadata?**

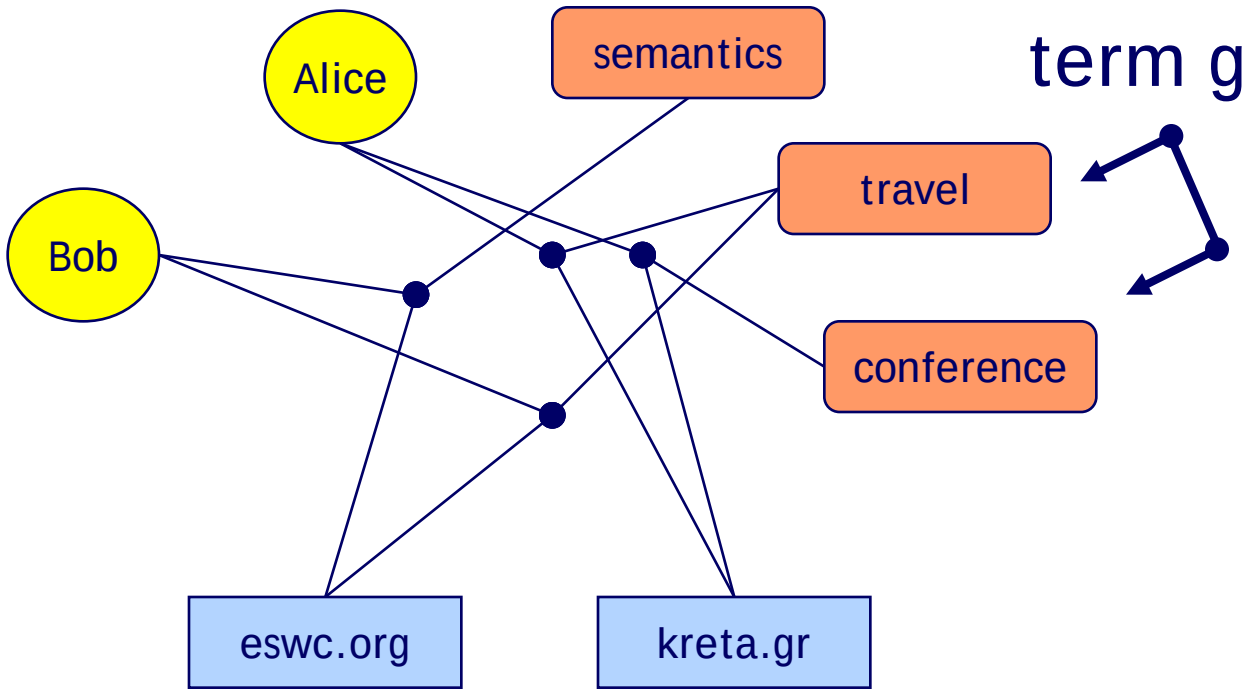
# The Story


Social Metadata
Term Abstractness
Abstractness measures based on social annotations
Evaluation

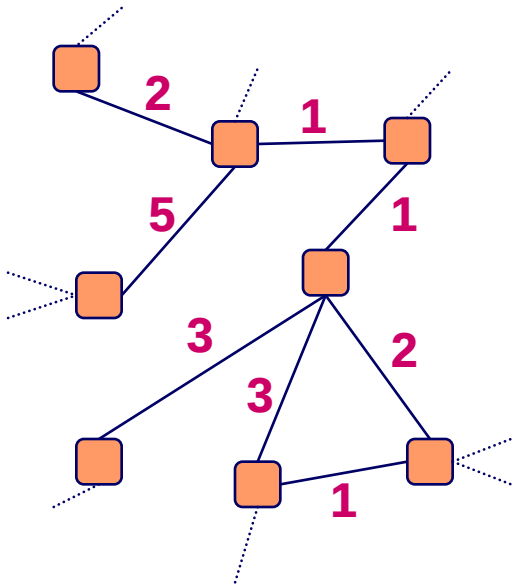
# Terms in Social Metadata


- **social annotations** are widely used
- focus here: **social tagging** as a simple and intuitive way to organize all kinds of resources
- **uncontrolled** vocabulary, tags are „just **strings**“
- formal model: **folksonomy** $F = (U, T, R, Y)$
  - Users $U$, Tags $T$, Resources $R$
  - Tag assignments $Y \subseteq (U \times T \times R)$

# Socially induced term relations




term graphs

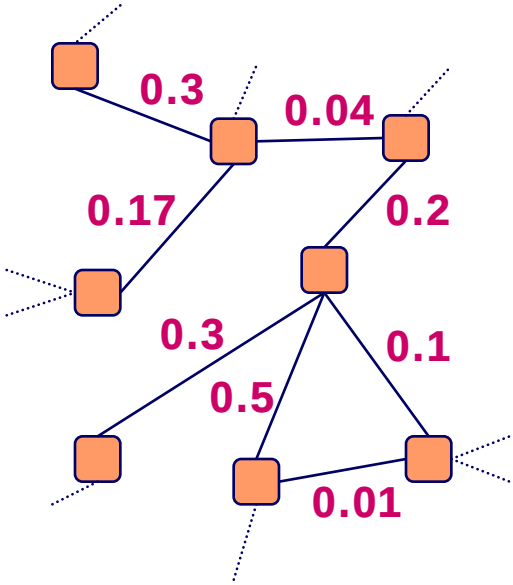



Term Graph **COOC**: edges weighted by co-occurrence:

$$cooc(t_1, t_2) = |\{(u, r) \in U \times R : (u, t_1, r) \in Y \wedge (u, t_2, r) \in Y\}|$$

# Semantic Tag Graphs

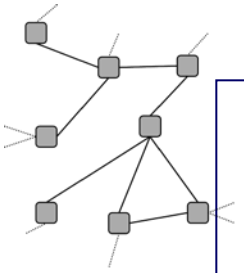
- Term Graph SIM: edges weigted by **Semantic Similarity**
- Represent **tag** as a **vector** in vector space of all resources


- Compute **Cosine Similarity** → „Resource context similarity“

with Cattuto et al: Semantic Grounding of Tag Relatedness in Social Bookmarking Systems (ISWC 2008)

# The Story



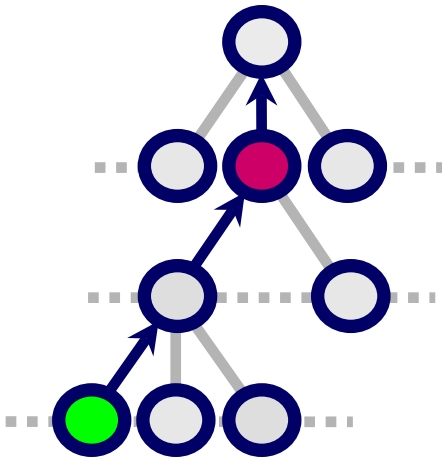
Social Metadata induces term relations (COOC and SIM)

Term Abstractness

Abstractness measures based on social annotations

Evaluation

# Definition of Term Abstractness

- **Term abstractness measure**: Partial Order (based on set of terms T):

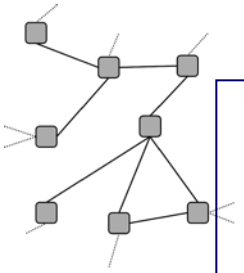
$$\sqsupset \subseteq T \times T$$

- $(t_1, t_2) \in \sqsupset$: „$t_1$ is **more abstract** than $t_2$“
- $\sqsupset_r$ induced by **ranking functions** $r : T \rightarrow \mathbb{R}$

$$r(t_1) > r(t_2) \Rightarrow (t_1, t_2) \in \sqsupset_r$$

| | |
|---|---|
| creature | 0.9 |
| animal | 0.8 |
| cat | 0.6 |
| mouse | 0.6 |
| felis_s | 0.4 |
| felis_s_c | 0.2 |
| petromysc | 0.2 |
| … | |

# The Story



Social Metadata induces term relations (COOC and SIM)

Formal definition of term abstractness measure

Abstractness measures based on social annotations

Evaluation

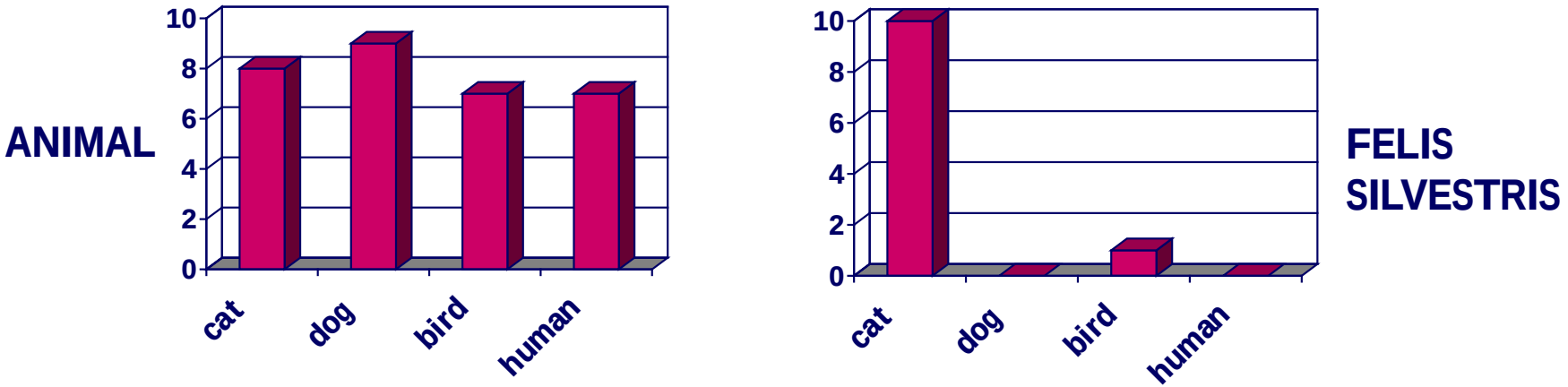
# Folksonomy-based generality ranking

- **Frequency** („the more often used, the more general“):

$$freq(t) = |\{(u, t', r) \in Y : t = t'\}|$$

- **Entropy** of Co-occurrence Distribution:

$$entr(t) = -\sum_{t' \in cooc(t)} p(t'|t) \log p(t'|t)$$



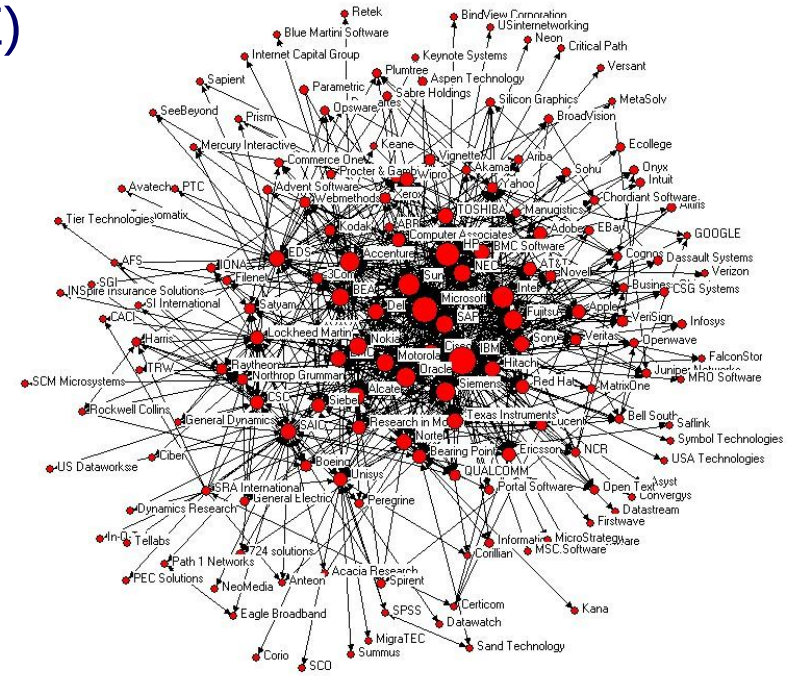
# Folksonomy-based generality ranking (2)

- **Centrality [Heyman2006]:**
  - „importance“ within graph G = (V,E)
  - *„more important, more general“*
  - Degree, Closeness, Betweenness

$$dc(v) = \frac{d(v)}{|V|-1}$$

$$cc(v) = \frac{1}{\sum_{t \in V \setminus v} d_G(v,t)}$$

$$bc(v) = \sum_{s \neq v \neq t \in V} \frac{\sigma_{st}(v)}{\sigma_{st}}$$




- Computed on **COOC** and **SIM** graphs

# Folksonomy-based generality ranking (3)

- **Statistical Model of Subsumption [Schmitz2006]:**
  - **Tag t probably subsumes t‘ if**

$$p(t|t') > \xi \wedge p(t'|t) < \xi$$

**for given threshold $\xi$ and**

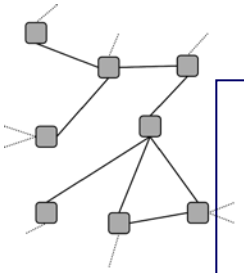
$$p(t'|t) = \frac{cooc(t',t)}{\sum_{t'' \in cooc(t)} cooc(t'',t)}$$

```
P1: animal cat
P2: animal dog
P3: animal cat
P4: animal mouse
P5: animal mouse
P6: animal cat
P7: animal bird
P8: animal dog
```

p(animal|cat) = 3/3 = 1

p(cat|animal) = 3/8 = 0.375

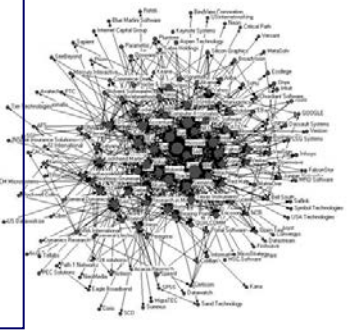
# The Story



Social Metadata induces term relations (COOC and SIM)

Formal definition of term abstractness measure

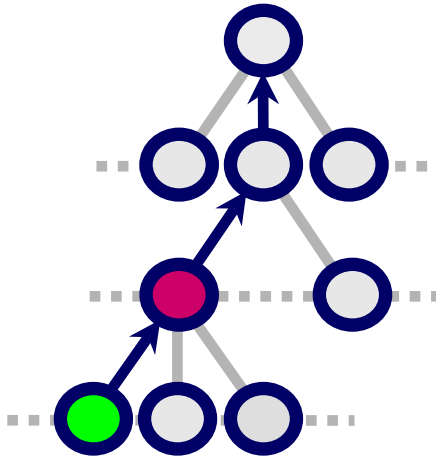
Frequency, Entropy, Centrality and Statstical subsumption are folksonomy-based generality rankings



Evaluation

# Which generality ranking is best?

- Idea: Compare partial orders with hierarchical information encoded in **reference taxonomies**



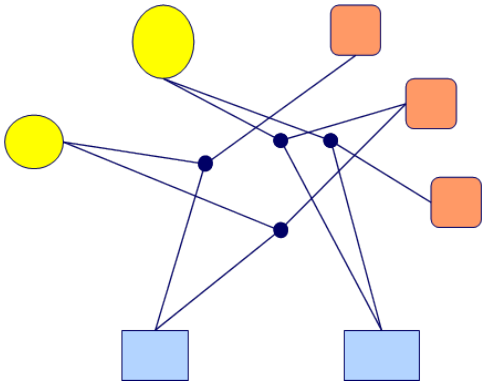
| | # concepts | # is-a relations |
|---|---|---|
| **Wordnet** | *79,690* | *81,866* |
| **Yago** | *244,553* | *249,465* |
| **DMOZ** | *767,019* | *767,018* |
| **Wikitaxonomy** | *2,445,974* | *4,447,010* |

- Comparison metric: **gamma rank coefficient** (based on sets of concordant pairs **C** / discordant pairs **D**):

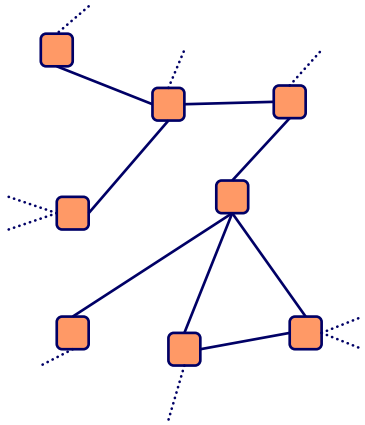
$$CR(\sqsupset, \sqsupset_*) = \frac{|C| - |D|}{|C| + |D|}$$

# Social Metadata used

- **Folksonomy** crawled from Delicious in 2006
  - 667,128 users
  - 2,454,546 tags
  - 18,782,132 resources
  - 140,333,714 tag assignments

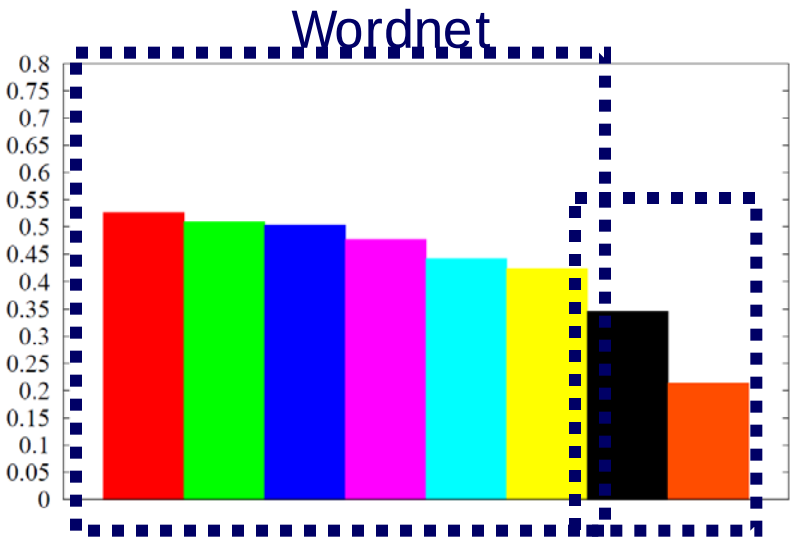


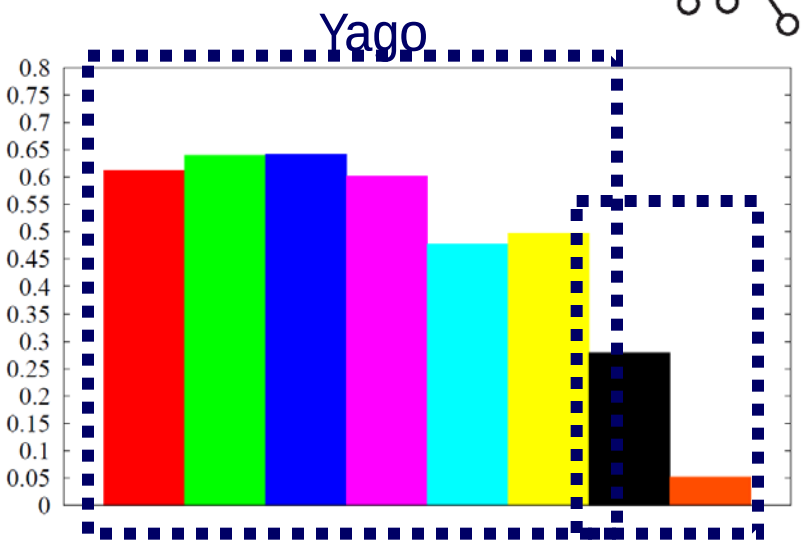
- Extracted term graph **COOC**:
  - 892,749 nodes, 38,210,913 edges
  - Edge filter: Co-occurrence threshold 2

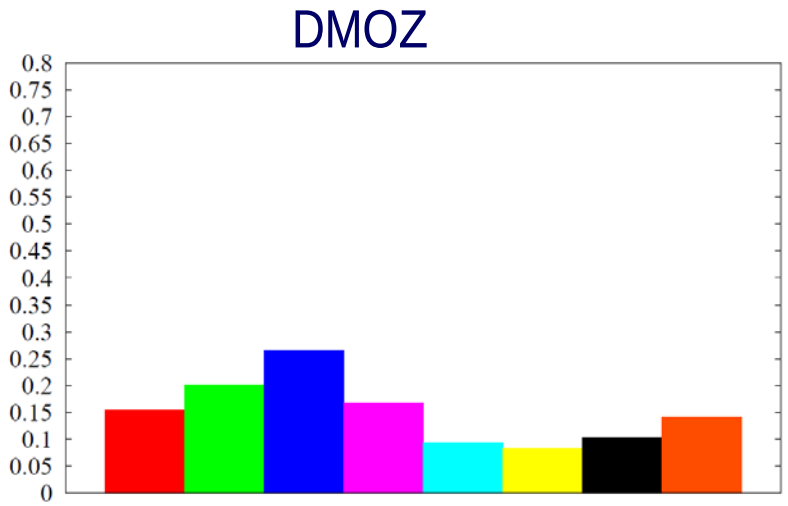


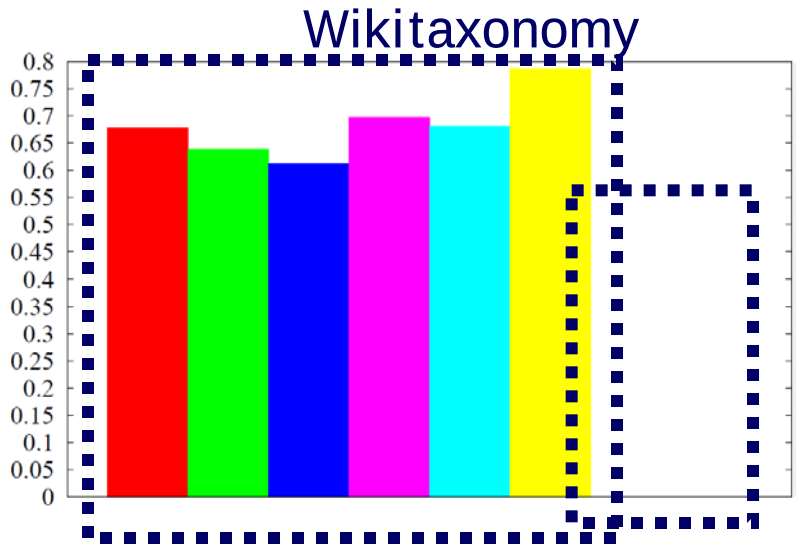
- Extracted term graph **SIM**:
  - 10,000 nodes, 405,706 edges
  - Edge filter: Similarity threshold 0.04

# Results

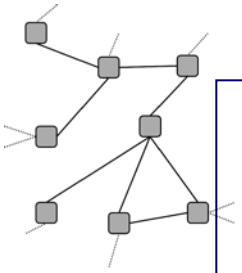




cc_cooc
dc_cooc
entropy
frequency
bc_cooc
subs
cc_sim
bc_sim

# Discussion & Implications

- In general: no „clear winner“, but measures based on **similarity graph** perform **worse**
- **Cooccurrence graph better** source for subsumption information
- **Simple** measures like **frequency** already a **good indicator** of generality
  - popularity/generality problem is less severe!
- Relevant for **ontology learning**, tag recommendation, query expansion, assisted browsing, …
- Future work: prior Tag Sense Disambiguation

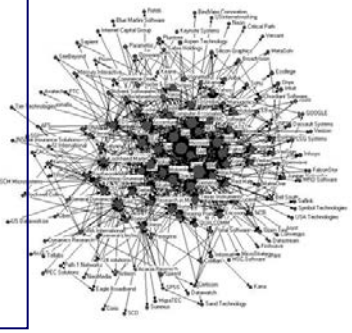
# The Story



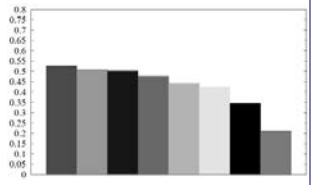
Social Metadata induces term relations (COOC and SIM)

Formal definition of term abstractness measure

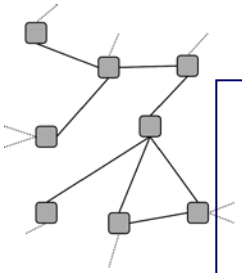
Frequency, Entropy, Centrality and Statstical subsumption are folksonomy-based generality rankings



Simple folksonomy-based measures approximate well term generality



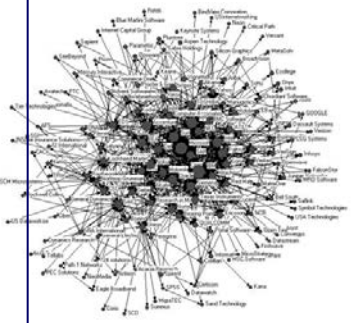
# The Story

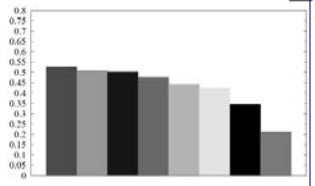


Social Metadata induces term relations (COOC and SIM)

Formal definition of term abstractness measure


Frequency, Entropy, Centrality and Statstical subsumption are folksonomy-based generality rankings





Simple folksonomy-based measures approximate well term generality

Thanks!
benz@cs.uni-kassel.de