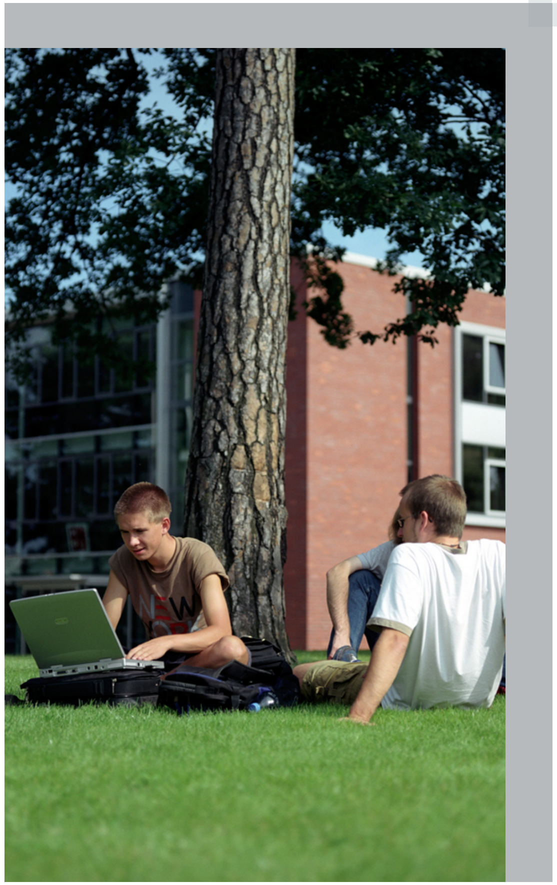
Hasso Plattner Institut
IT Systems Engineering | Universität Potsdam



The Metadata Triumvirate

Social Annotations, Anchor Texts and Search Queries

Michael G. Noll, Christoph Meinel

In this talk

Metadata – "data about data"

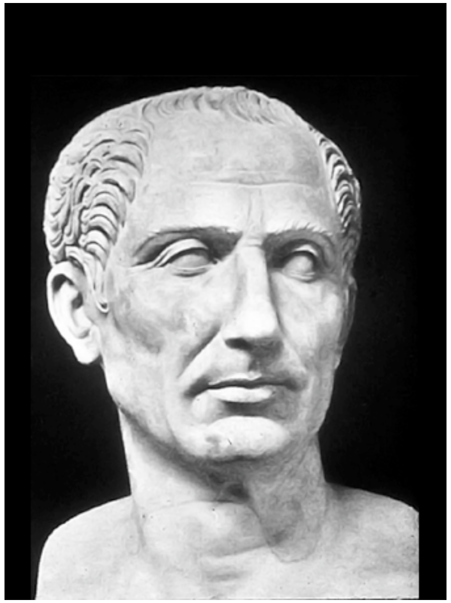
- Here: Web metadata, i.e. data about **WWW documents**
- **Variety of uses** for such metadata in Web information retrieval: indexing, ranking, filtering, …
- **Different types** of Web metadata:

  In this talk, we study and compare 3 very popular ones with the goal to improve our understanding of these metadata types, thereby helping us to improve existing IR algorithms or come up with new ones.

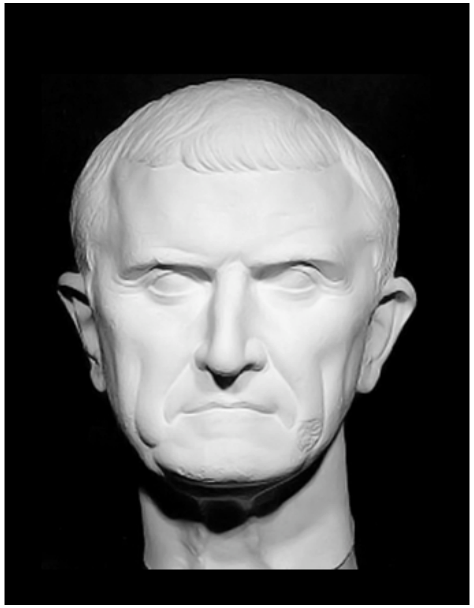
The history of Triumvirates

The Metadata Triumvirate

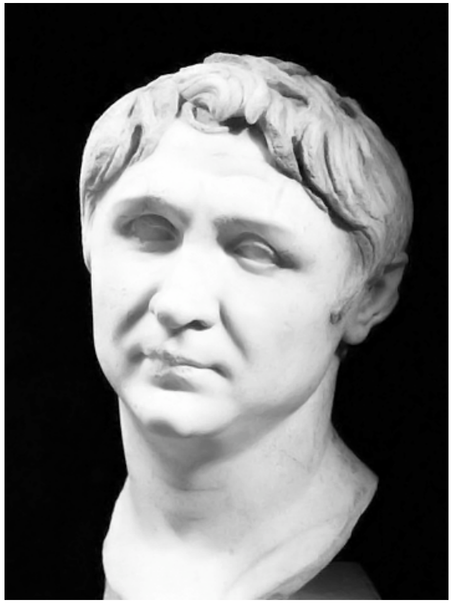
Triumvirate 1.0, 60 BC – “Conquer the World!”



Caesar



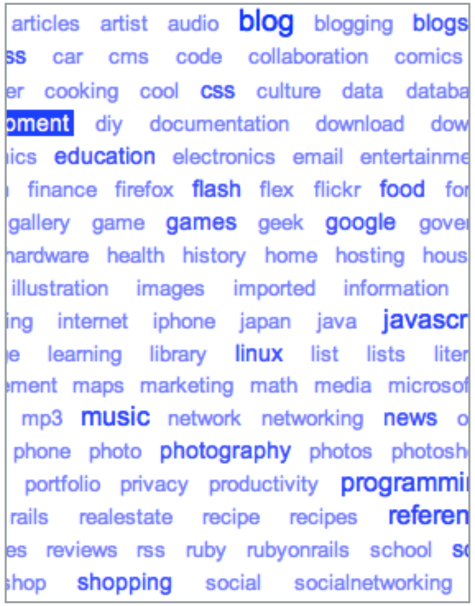
Crassus



Pompeius

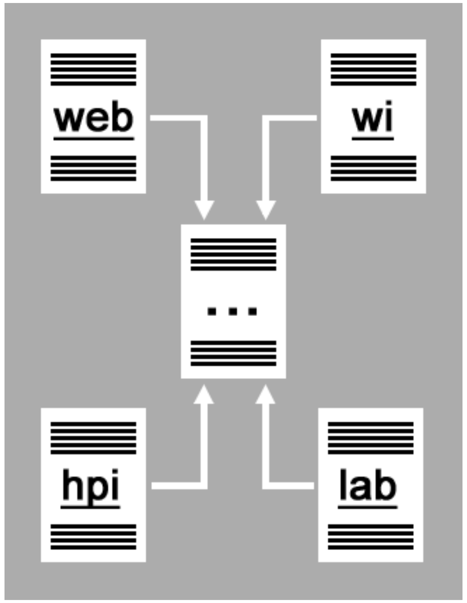
The Metadata Triumvirate

Triumvirate 2.0, 2008 AD – “Conquer the World Wide Web?”

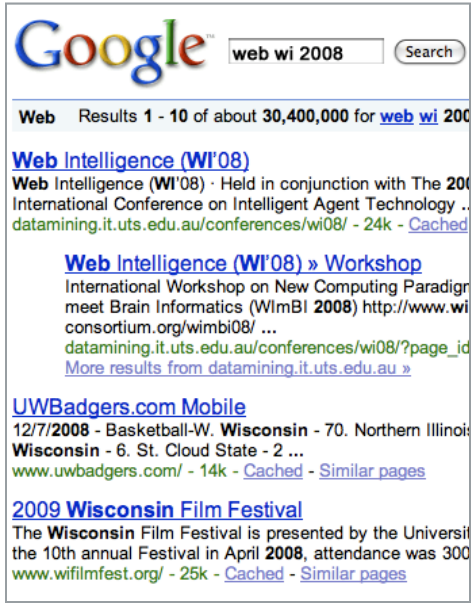


articles artist audio blog blogging blogs
ss car cms code collaboration comics
er cooking cool css culture data databa
pment diy documentation download dow
ics education electronics email entertainme
finance firefox flash flex flickr food fo
gallery game games geek google gove
hardware health history home hosting hous
illustration images imported information
ing internet iphone japan java javascr
e learning library linux list lists liter
ment maps marketing math media microsof
mp3 music network networking news o
phone photo photography photos photosh
portfolio privacy productivity programmi
rails realestate recipe recipes referen
es reviews rss ruby rubyonrails school s
shop shopping social socialnetworking

Social Annotations



Anchor Texts



Google web wi 2008 Search

Web Results **1** - **10** of about **30,400,000** for web wi 200

Web Intelligence (WI’08)
Web Intelligence (**WI**’08) · Held in conjunction with The **20**
International Conference on Intelligent Agent Technology ..
datamining.it.uts.edu.au/conferences/wi08/ - 24k - Cached

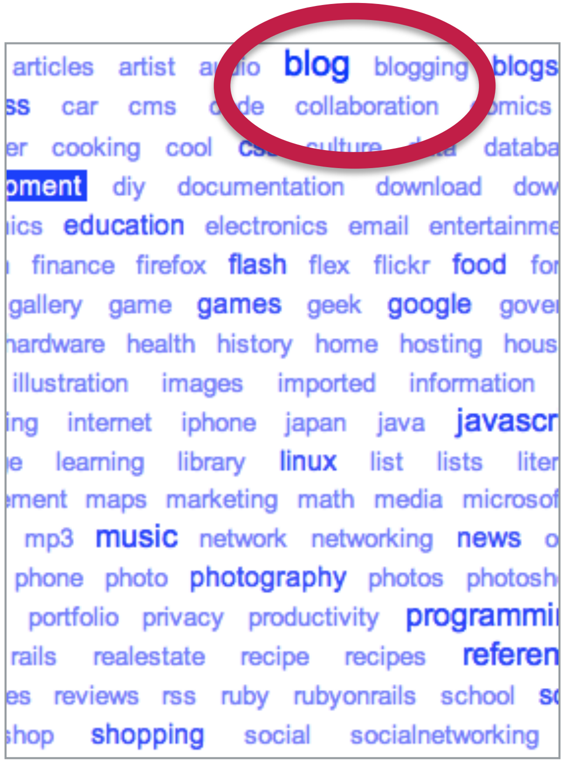
Web Intelligence (WI’08) » Workshop
International Workshop on New Computing Paradigr
meet Brain Informatics (WImBI **2008**) http://www.**wi**
consortium.org/wimbi08/ **...**
datamining.it.uts.edu.au/conferences/wi08/?page_id
More results from datamining.it.uts.edu.au »

UWBadgers.com Mobile
12/7/**2008** - Basketball-W. **Wisconsin** - 70. Northern Illinoi
Wisconsin - 6. St. Cloud State - 2 **...**
www.uwbadgers.com/ - 14k - Cached - Similar pages

2009 Wisconsin Film Festival
The **Wisconsin** Film Festival is presented by the Universit
the 10th annual Festival in April **2008**, attendance was 300
www.wifilmfest.org/ - 25k - Cached - Similar pages

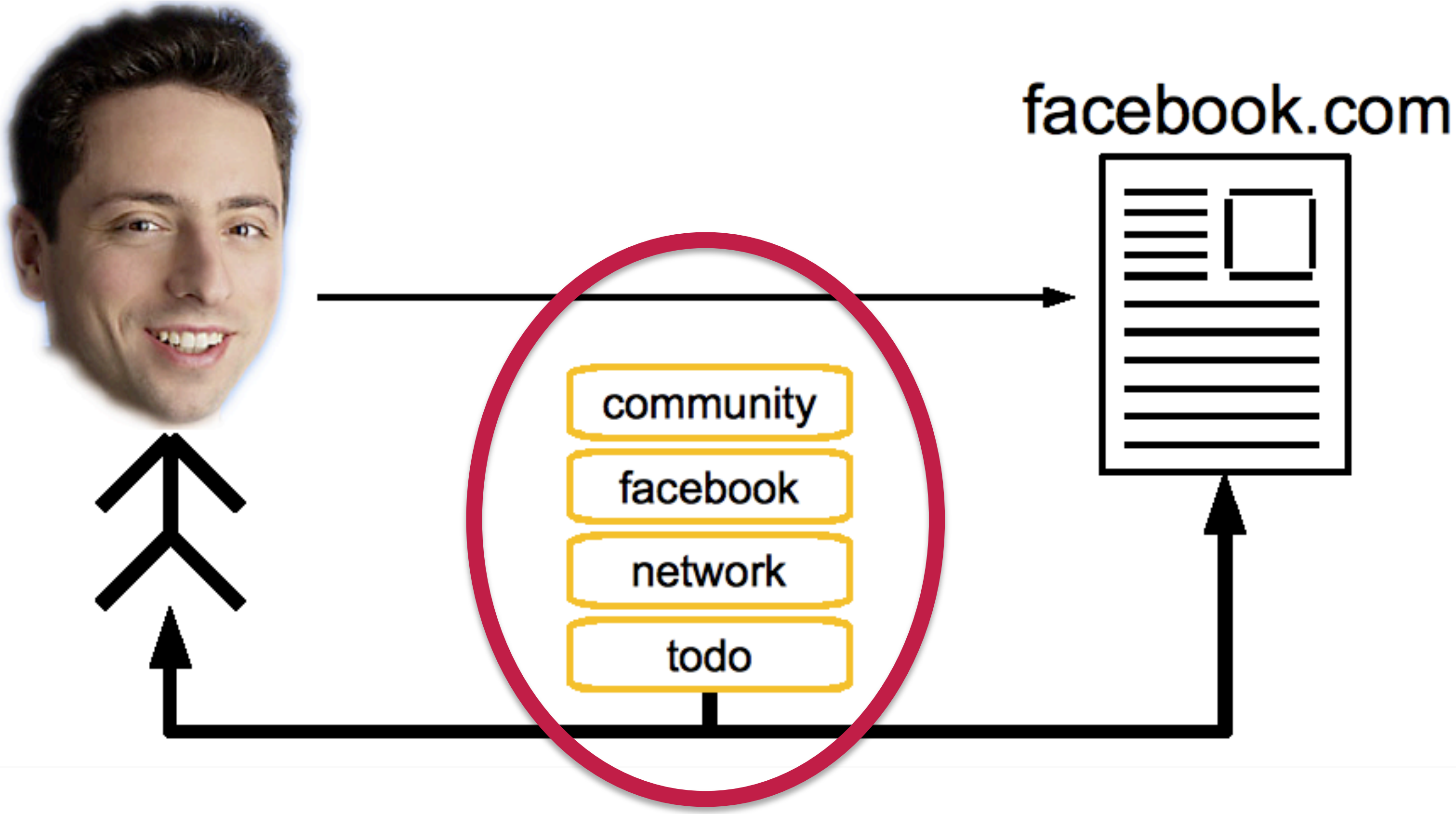
Search Queries

Social Annotations


articles artist audio blog blogging blogs
ss car cms code collaboration comics
er cooking cool css culture data databa
pment diy documentation download dow
ics education electronics email entertainme
finance firefox flash flex flickr food fo
gallery game games geek google gove
hardware health history home hosting hous
illustration images imported information
ing internet iphone japan java javascr
e learning library linux list lists liter
ment maps marketing math media microsof
mp3 music network networking news o
phone photo photography photos photosh
portfolio privacy productivity programmi
rails realestate recipe recipes referen
es reviews rss ruby rubyonrails school sc
shop shopping social socialnetworking

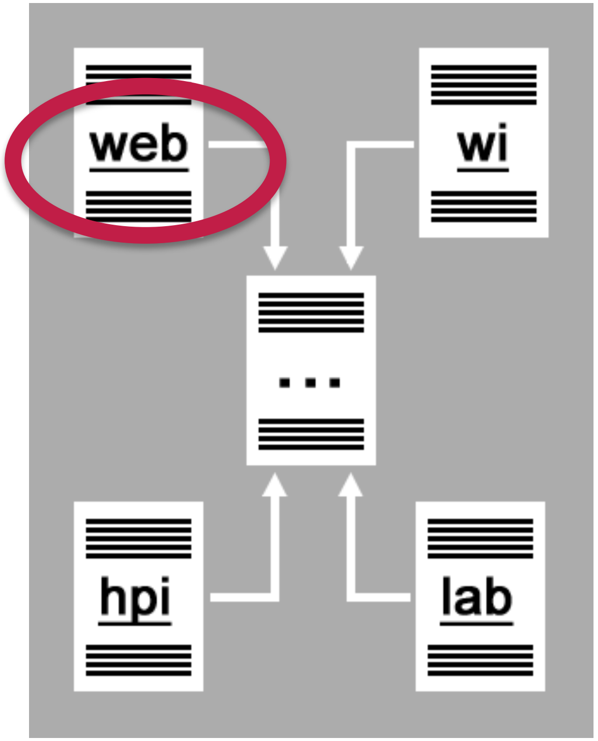
- **Definition of a social annotation:** list of “tags” (words) with which a social bookmark has been annotated
- Derived from **user-provided metadata**
- What does the social annotation **“web, conference, sydney, 2008”** tell about the user and the annotated document?
- Used for Web search personalization, emerging semantics, content classification, expert identification, …

Social Annotations



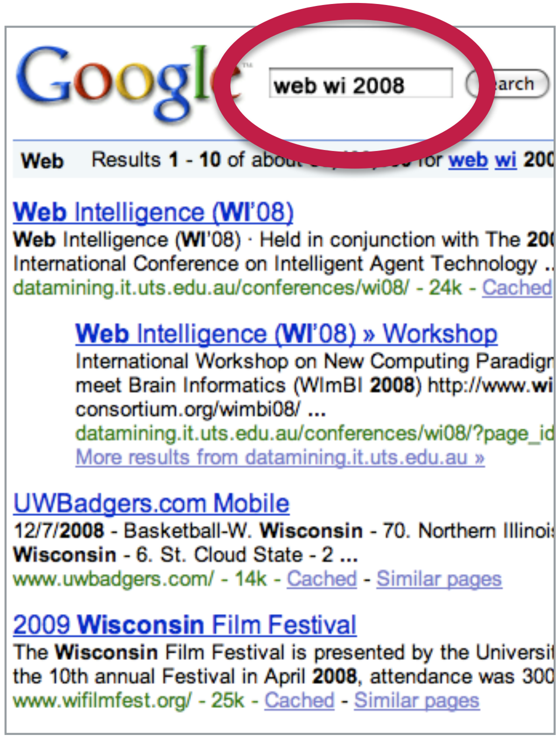
Sergey's social annotation of Facebook.com

Anchor Texts



- **Definition of an anchor text:** words within <a>…</a> HTML element
- Derived from **Web link structure**
- What do the anchor texts **“web”, “wi”, “hpi”, “lab”** tell about the linked page?
- Used for gaining more information about the linked Web pages, for improving indexing and ranking techniques, …

Search Queries



web wi 2008

Web Results **1** - **10** of abou for web wi 200

Web Intelligence (WI'08)
Web Intelligence (**WI**'08) · Held in conjunction with The **20**
International Conference on Intelligent Agent Technology ..
datamining.it.uts.edu.au/conferences/wi08/ - 24k - Cached

Web Intelligence (WI'08) » Workshop
International Workshop on New Computing Paradigr
meet Brain Informatics (WImBI **2008**) http://www.**wi**
consortium.org/wimbi08/ **...**
datamining.it.uts.edu.au/conferences/wi08/?page_id
More results from datamining.it.uts.edu.au »

UWBadgers.com Mobile
12/7/**2008** - Basketball-W. **Wisconsin** - 70. Northern Illinoi
Wisconsin - 6. St. Cloud State - 2 **...**
www.uwbadgers.com/ - 14k - Cached - Similar pages

2009 Wisconsin Film Festival
The **Wisconsin** Film Festival is presented by the Universit
the 10th annual Festival in April **2008**, attendance was 300
www.wifilmfest.org/ - 25k - Cached - Similar pages

- **Definition of a search query:**
  search keywords of the user's query
- Derived from search query logs,
  i.e. **user interactions**
- What does the search **"web wi 2008"** tell about the searcher or the clicked search result document?
- Used for query rewriting, user profiling, extracting semantics, …

Questions we want to answer

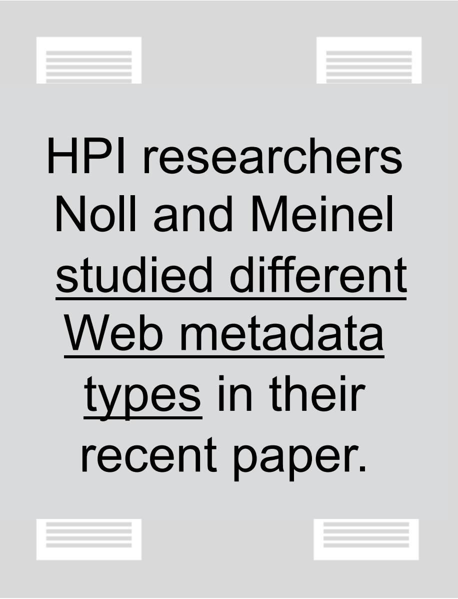
Questions we want to answer

How do these different types of metadata compare?

animals architecture art august austral
birthday black blackandwhite blue bost
canon car cat chicago china chi
scotland sea seattle show sky snow s
taiwan texas thailand tokyo toronto tour
vacation vancouver washington wate

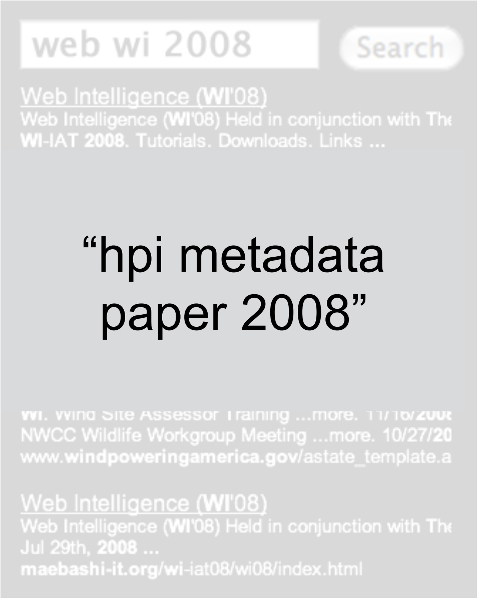
“metadata, paper, social web, hpi, research, 2008”

Social Annotations



HPI researchers Noll and Meinel <u>studied different Web metadata types</u> in their recent paper.

Anchor Texts



Web Intelligence (WI'08)
Web Intelligence (WI'08) Held in conjunction with Th
WI-IAT 2008. Tutorials. Downloads. Links ...

“hpi metadata paper 2008”

WI. Wind Site Assessor Training ...more. 11/16/200
NWCC Wildlife Workgroup Meeting ...more. 10/27/20
www.windpoweringamerica.gov/astate_template.a

Web Intelligence (WI'08)
Web Intelligence (WI'08) Held in conjunction with Th
Jul 29th, 2008 ...
maebashi-it.org/wi-iat08/wi08/index.html

Search Queries

Questions we want to answer

Five questions

- Q1: **Volume** of data per single metadata item?
- Q2: **New data** per metadata type?
- Q3: **Homogeneous** or **heterogeneous** metadata?
- Q4: **Similarity** between metadata types?
- Q5: Usefulness for **classification** of web documents?

Experimental Setup

Experimental data

We created our own experimental data set "CABS120k08" in 2008

- Bootstrapped by an intersection of **AOL500k** and **Open Directory Project**

  + targeted **Web crawl**
  + scraping **Delicious**
  + retrieving **Google PageRank**

  = metadata for **120,000** web documents

Overview of CABS120k08
120,000 web documents
2,600,000 search queries
85,000 categories
2,200,000 anchor texts
1,300,000 social annotations
120,000 PageRank scores

Data set (500 MB) is available for download at:
http://www.michael-noll.com/cabs120k08/

Experimental Results

Q1: Volume of data per single metadata item?

"Does a social annotation provide more data than an anchor text?"
or: "How much data do users provide when using a specific metadata type?"

Experimental results

Approach

- Measure size of a single metadata item by its “length”
- Definitions of length for…
  - Social annotation → number of **tags**
  - Anchor text → number of **words**
  - Search query → number of **search keywords**

Experimental results

Mean length

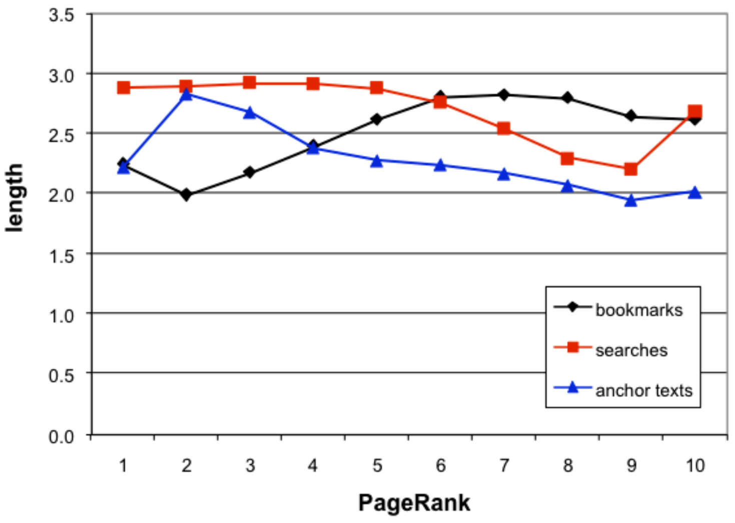
Social annotation:	2.49
Anchor text:	2.43
Search query:	2.89

→ Surprisingly, **2.x** seems to be a “magic number” for user behavior across different problem domains (social bookmarking, hyperlink creation, Web search). Human psychology?

Experimental results

Correlation of length with document popularity:

- **positively** for social annotations
- **negatively** for anchor texts and search queries



→ Anchor texts provide more metadata for less popular documents, whereas social annotations do so for popular ones

Q2: New data per metadata type?

"How helpful is an analysis of a given metadata type
for discovering new information?"

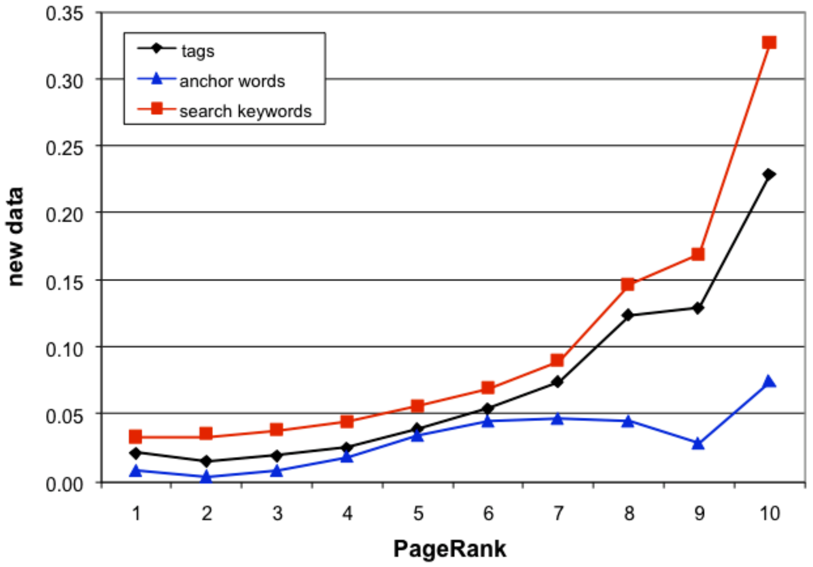
Experimental results

Approach

- Measure “novelty” of data provided by each metadata type
- Novelty is defined as the **percentage of unique terms** which are **new to a Web document**, i.e. terms that are not already present in the document’s <TITLE>, <BODY>, plus selected HTML metadata
- For example, to retrieve a Web document in a search for “biology” even though the query term “biology” is not part of the document’s HTML content.

Experimental results

- Generally, the amount of new information is **relatively low**
- ≤ 6% for 90% of documents
- Search queries >> social annotations >> anchor texts



→ Compared to anchor texts, social annotations are a better source of new data

→ However, similarity between social annotations and anchor texts (as we see later) is rather low = they provide **different** data, so both are useful!

Q3: Homogeneous or heterogeneous metadata?

“Is the data of each metadata type consistent/diverse/chaotic…?”

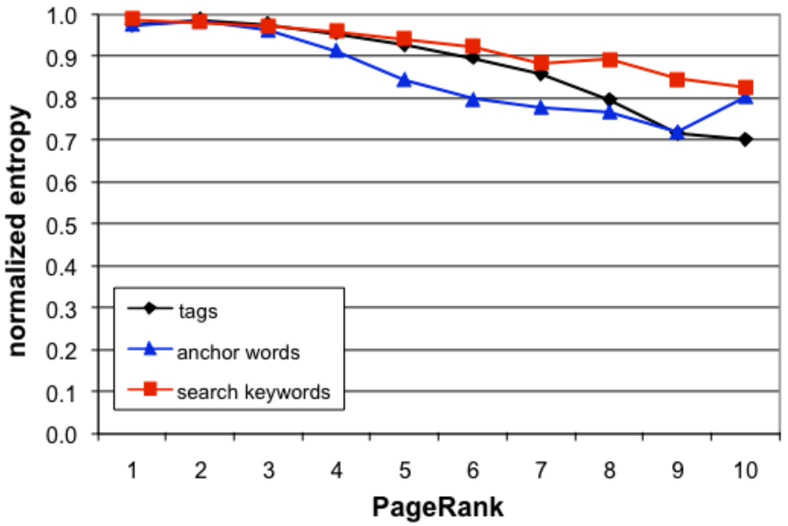
Experimental results

Approach

- Measure “diversity” of data **within** a given metadata type
- **Entropy** is used to measure diversity based on terms and term counts
- Note: Scoring a **high diversity** can indicate both **positive** (capturing different perceptions/meanings of content) and **negative** results (no consensus, noise).

Experimental results

- Strong **negative correlation** with document popularity for all types: With increasing popularity, diversity of information decreases.
- Highest diversity for search queries: most “random” task, formulating good queries, spelling corrections ?
- Social annotations more diverse than anchor texts



→ Potential advantage for social annotations as they might capture information and meanings that anchor texts miss (cf. Bao et al. WWW 2007).

Q4: Similarity between metadata types?

"How similar is the data provided by these metadata types?"

Experimental results

Approach

- Study the **interrelations** between metadata types
- **Pairwise cosine similarity** is used to measure similarity
- Preprocessing of terms: splitting (“new_york”), stemming, stop words

Experimental results

	Social annotat.	**Anchor texts**	**Search queries**	**Categories**
Social annotat.	x	0.126	0.126	**0.189**
Anchor texts	0.126	x	**0.193**	0.103
Search queries	0.126	**0.193**	x	0.102
Categories	**0.189**	0.103	0.102	x

Highest similarities for two pairs:

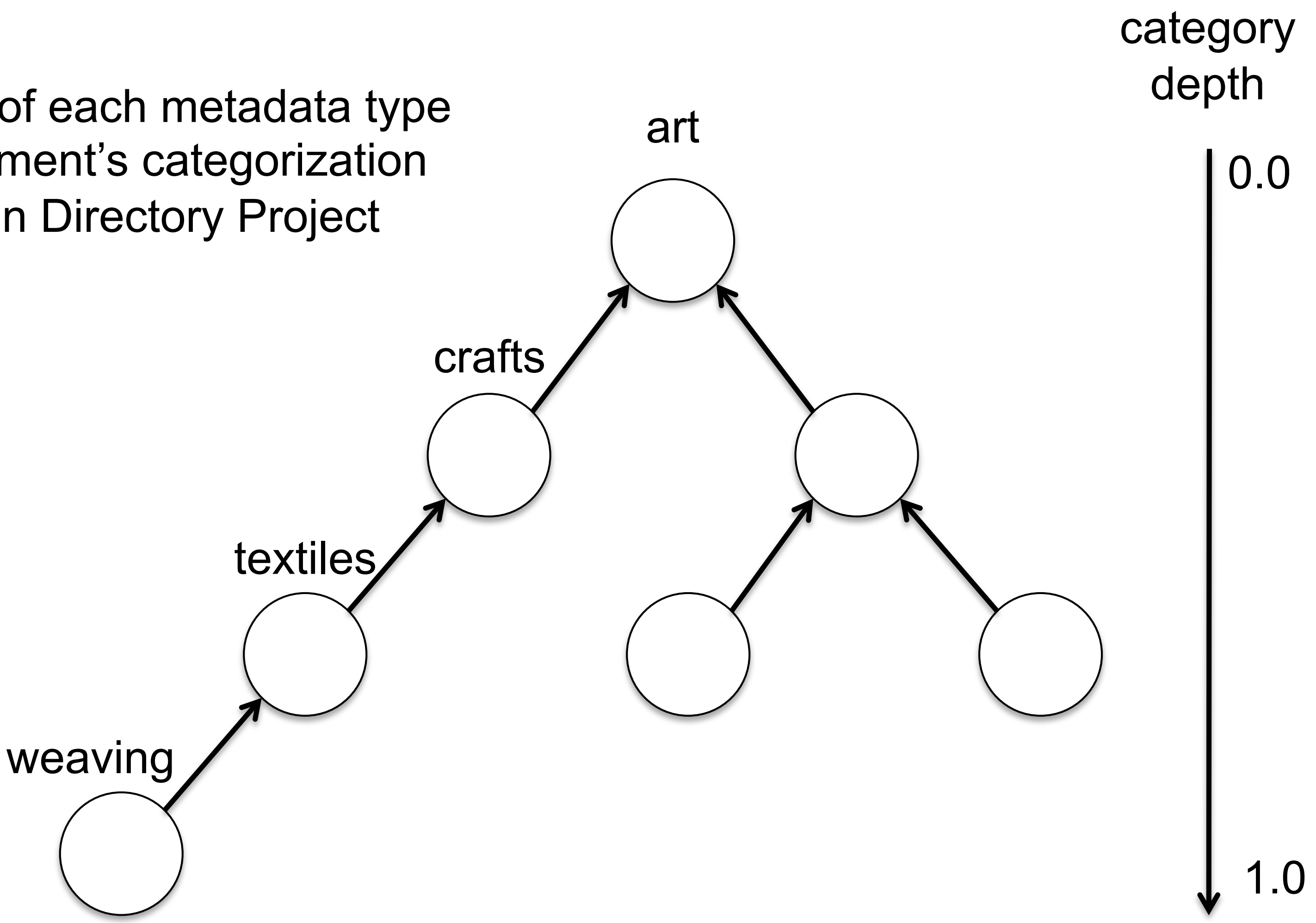
- sim(**social annotations, categories**) = 0.189 → “better” for classification?
- sim(**anchor texts, search queries**) = 0.193 → “better” for Web search?

Q5: Usefulness for classification of web documents?

"How helpful are these metadata types for classification tasks?"

Classification

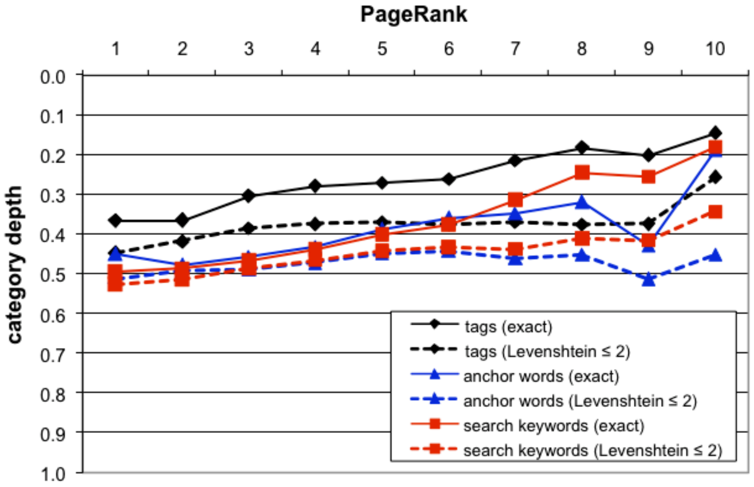
Approach:
Matching data of each metadata type against a document's categorization trees from Open Directory Project



Hasso Plattner Institut

Experimental results

- Strong **negative correlation** with document popularity for all types: With increasing popularity, broader classification scores are achieved.
- Social annotations are “used” for broader classification than anchor texts and search queries

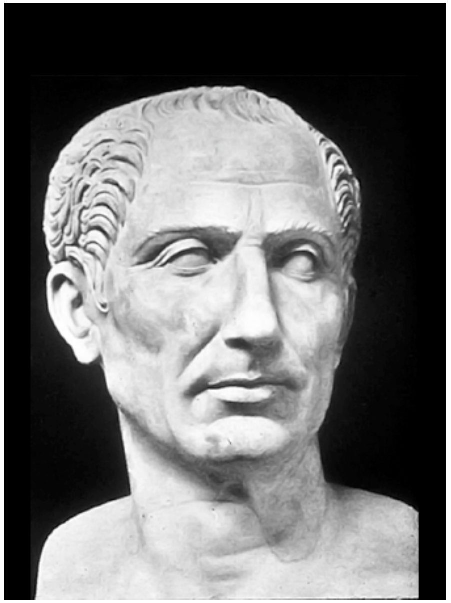


→ Of all three, social annotations seem to be the best at classification tasks

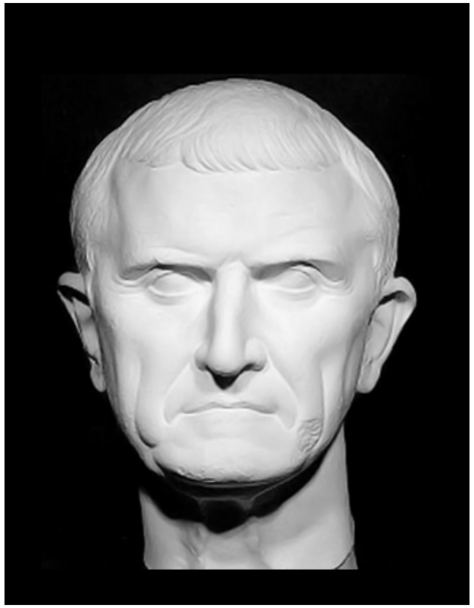
Conclusions

The Metadata Triumvirate

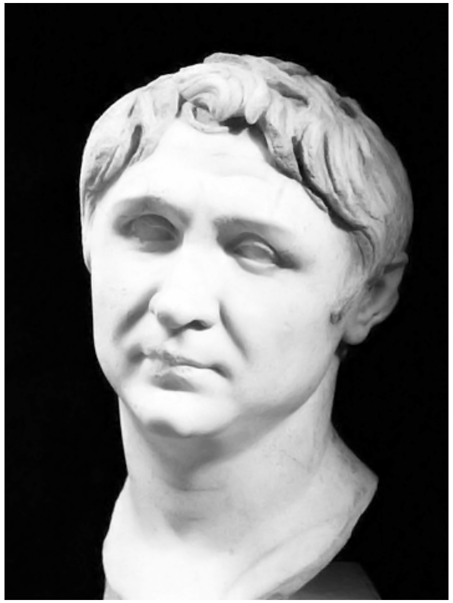
Triumvirate 1.0



Caesar



Crassus



Pompeius

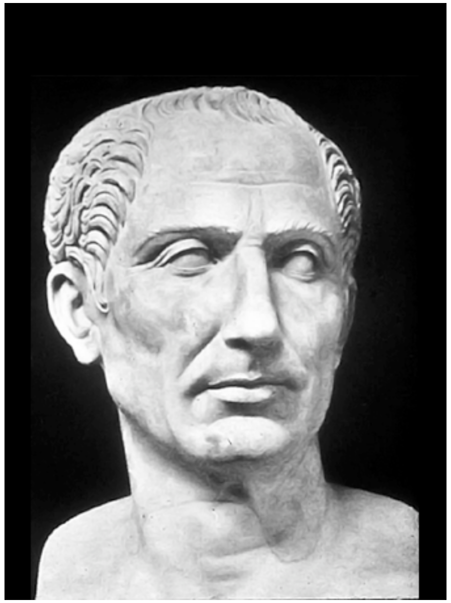
The Metadata Triumvirate

Worked out quite well…

Roman Empire, 44 BC

The Metadata Triumvirate

…however…



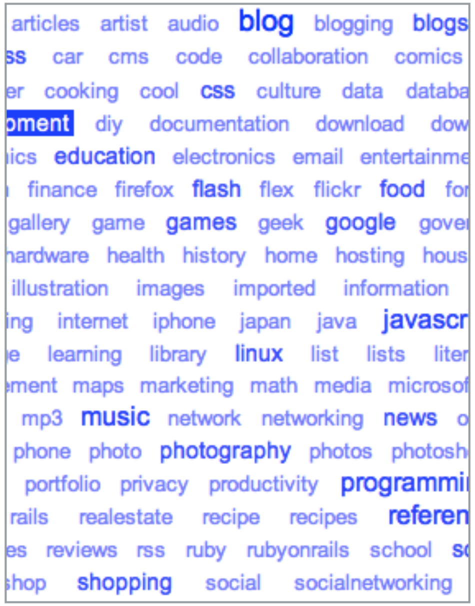
Caesar

Crassus

Pompeius

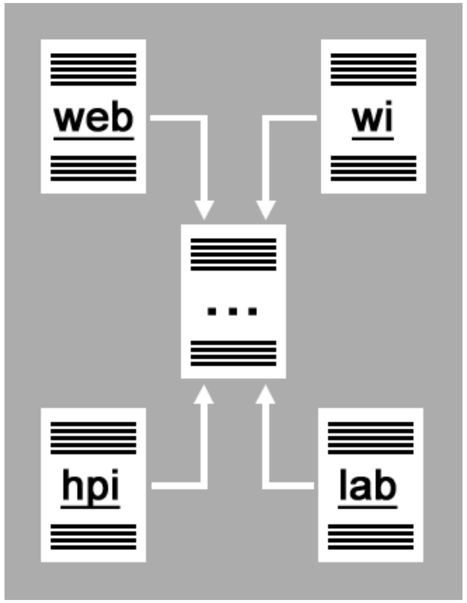
The Metadata Triumvirate

Metadata Triumvirate – no casualties (yet)!

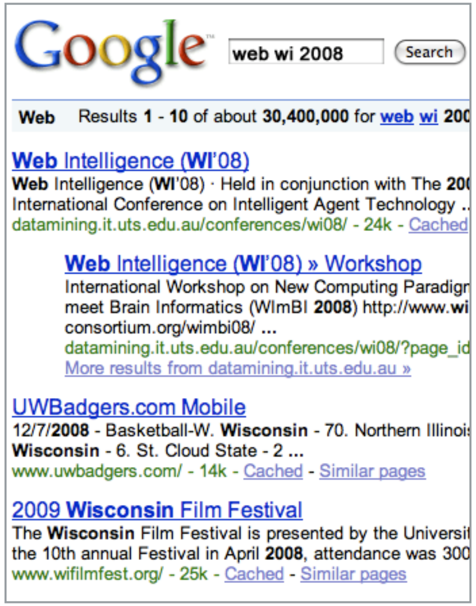


articles artist audio **blog** blogging **blogs**
ss car cms code collaboration comics
er cooking cool **css** culture data databa
pment diy documentation download dow
ics **education** electronics email entertainme
finance firefox **flash** flex flickr **food** for
gallery game **games** geek **google** gover
hardware health history home hosting hous
illustration images imported information
ing internet iphone japan java **javascr**
e learning library **linux** list lists liter
ment maps marketing math media microsof
mp3 **music** network networking **news** o
phone photo **photography** photos photosh
portfolio privacy productivity **programmi**
rails realestate recipe recipes **referen**
es reviews rss ruby rubyonrails school **sc**
hop **shopping** social socialnetworking

Social Annotations



Anchor Texts



Google web wi 2008 Search

Web Results **1** - **10** of about **30,400,000** for web wi 200

Web Intelligence (WI'08)
Web Intelligence (**WI**'08) · Held in conjunction with The **200**
International Conference on Intelligent Agent Technology ..
datamining.it.uts.edu.au/conferences/wi08/ - 24k - Cached

Web Intelligence (WI'08) » Workshop
International Workshop on New Computing Paradigr
meet Brain Informatics (WImBI **2008**) http://www.**wi**
consortium.org/wimbi08/ ...
datamining.it.uts.edu.au/conferences/wi08/?page_id
More results from datamining.it.uts.edu.au »

UWBadgers.com Mobile
12/7/**2008** - Basketball-W. **Wisconsin** - 70. Northern Illinoi
Wisconsin - 6. St. Cloud State - 2 ...
www.uwbadgers.com/ - 14k - Cached - Similar pages

2009 Wisconsin Film Festival
The **Wisconsin** Film Festival is presented by the Universit
the 10th annual Festival in April **2008**, attendance was 300
www.wifilmfest.org/ - 25k - Cached - Similar pages

Search Queries

Conclusions

- First study to compare social annotations, anchor texts and search queries directly on a large volume of real-world data
- Starting point for future research
- Research data set CABS120k08, available for free download: http://www.michael-noll.com/cabs120k08/

Contact info:
Michael G. Noll
michael.noll@hpi.uni-potsdam.de
Hasso Plattner Institute
Potsdam, Germany