Controlled Vocabulary versus Tags

Do experts choose the right names…

… to label the wrong things?

Paul Heymann and Hector Garcia-Molina

Department of Computer Science

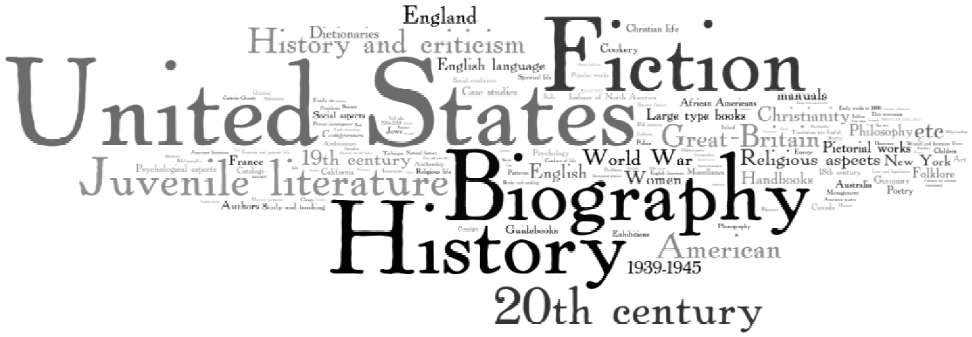
Stanford University

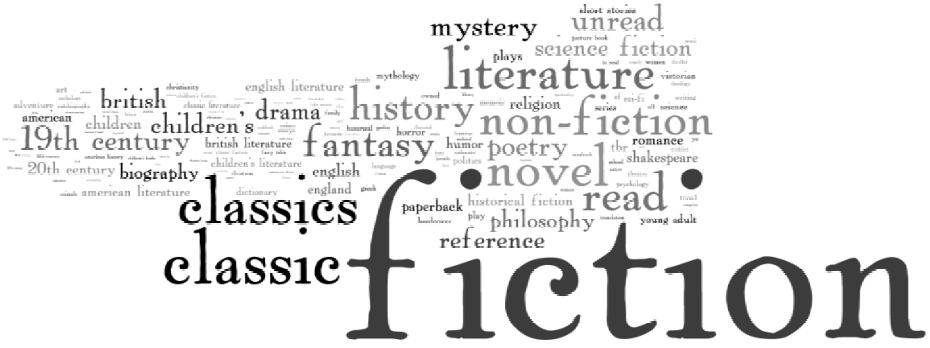
This Paper

Controlled Vocabulary (LCSH)

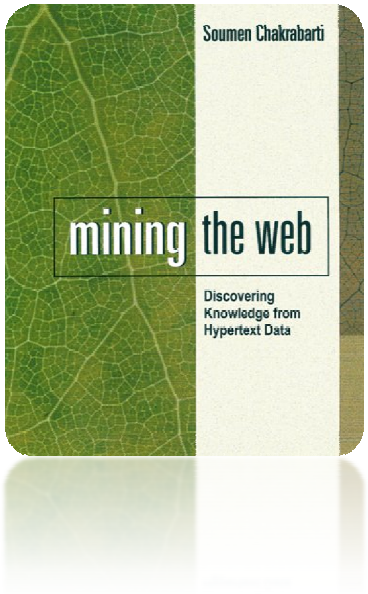
VS

Uncontrolled Vocabulary (Tagging)



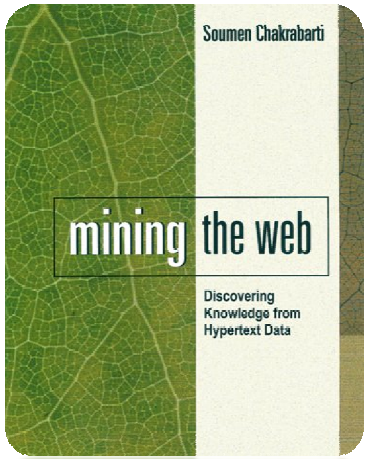


Example: Mining the Web



Example: Mining the Web

Controlled Vocabulary



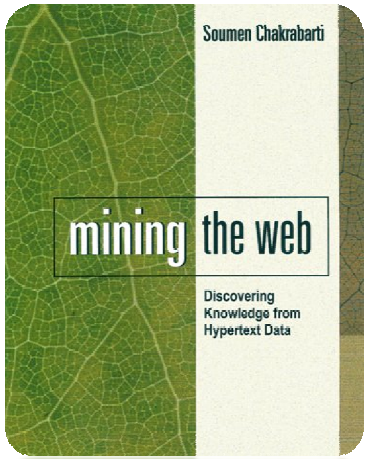
Data mining

Hypertext systems

Web databases

Example: Mining the Web

Controlled Vocabulary



Uncontrolled Vocabulary

Data mining

Hypertext systems

Web databases

data mining (7)

machine learning (3)

information retrieval (2)

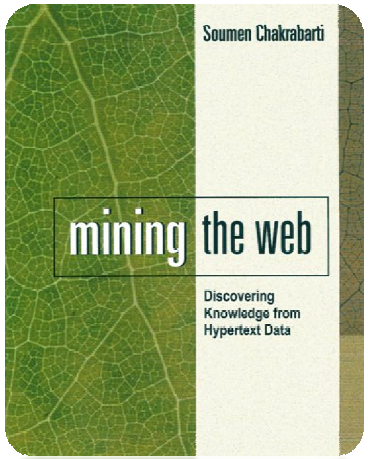
reference (2)

web (2)

... (17 others)

Example: Mining the Web

Controlled Vocabulary



Uncontrolled Vocabulary

Data mining

Hypertext systems

Web databases

equivalent

data mining (7)

machine learning (3)

information retrieval (2)

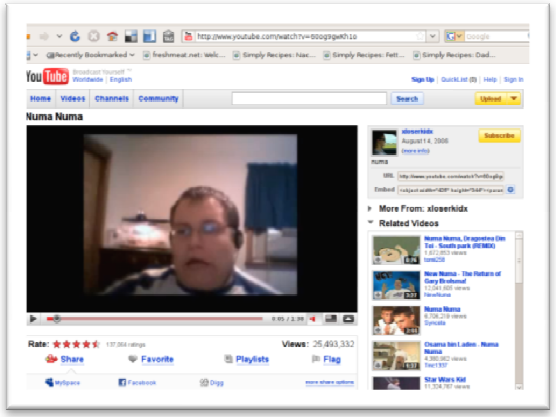
reference (2)

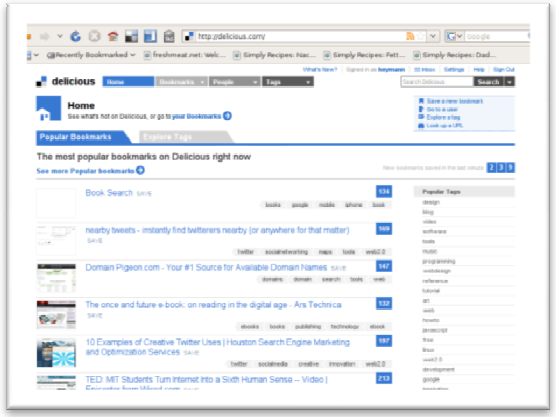
web (2)

... (17 others)

Background

- Tags are the emerging standard for organizing user generated content (e.g., Flickr, delicious, YouTube, …)





- Motivation: Better understand how tags organize data
- Books provide a ready-made gold standard for evaluating tags.

Big Questions

- How do tags compare to controlled vocabularies and other methods for browsing?
- Are traditional vocabularies any good? Are experts?
- Should websites use tags? Should offline systems like libraries use tags?

… just beginning to address these and similar questions!

Results Roadmap

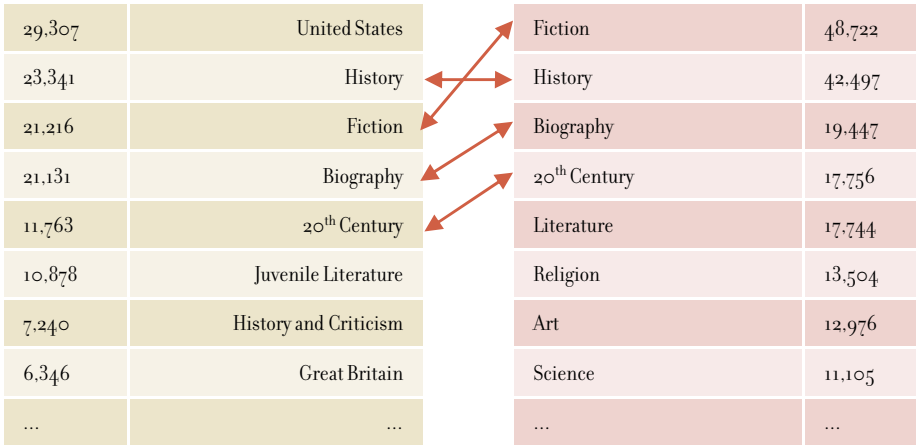
1. Vocabulary Term Equivalence

Results Roadmap

1. Vocabulary Term Equivalence

"data mining" ⟷ equivalent ⟷ "data mining"

2. Ranking Differences



29,307	United States	Fiction	48,722
23,341	History	History	42,497
21,216	Fiction	Biography	19,447
21,131	Biography	20th Century	17,756
11,763	20th Century	Literature	17,744
10,878	Juvenile Literature	Religion	13,504
7,240	History and Criticism	Art	12,976
6,346	Great Britain	Science	11,105
…	…	…	…

Results Roadmap

1. Vocabulary Term Equivalence

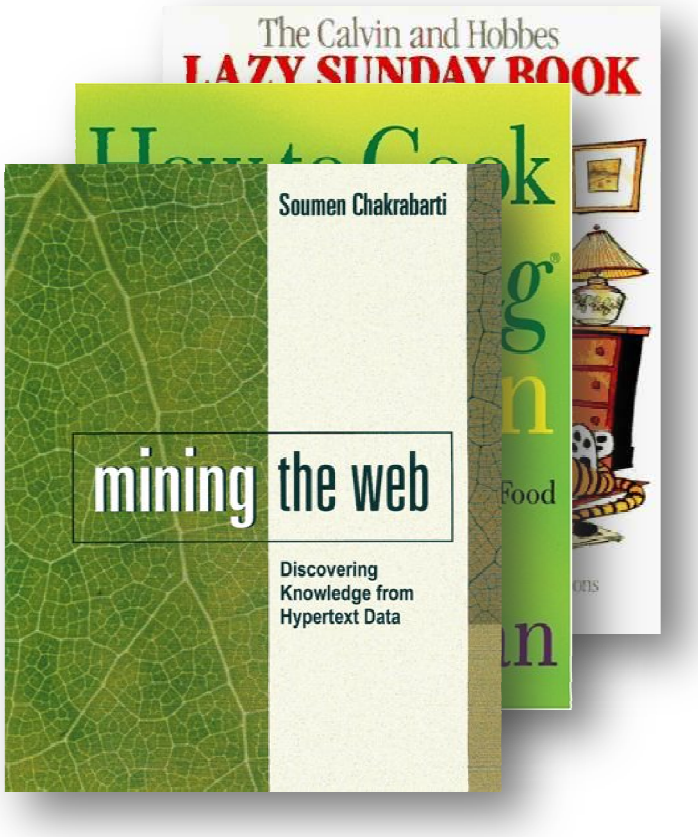
"data mining" ⟷ equivalent ⟷ "data mining"

2. Ranking Differences

29,307	United States	Fiction	48,722
23,341	History	History	42,497
21,216	Fiction	Biography	19,447
21,131	Biography	20th Century	17,756
11,763	20th Century	Literature	17,744
10,878	Juvenile Literature	Religion	13,504
7,240	History and Criticism	Art	12,976
6,346	Great Britain	Science	11,105
…	…	…	…

3. Annotation Overlap

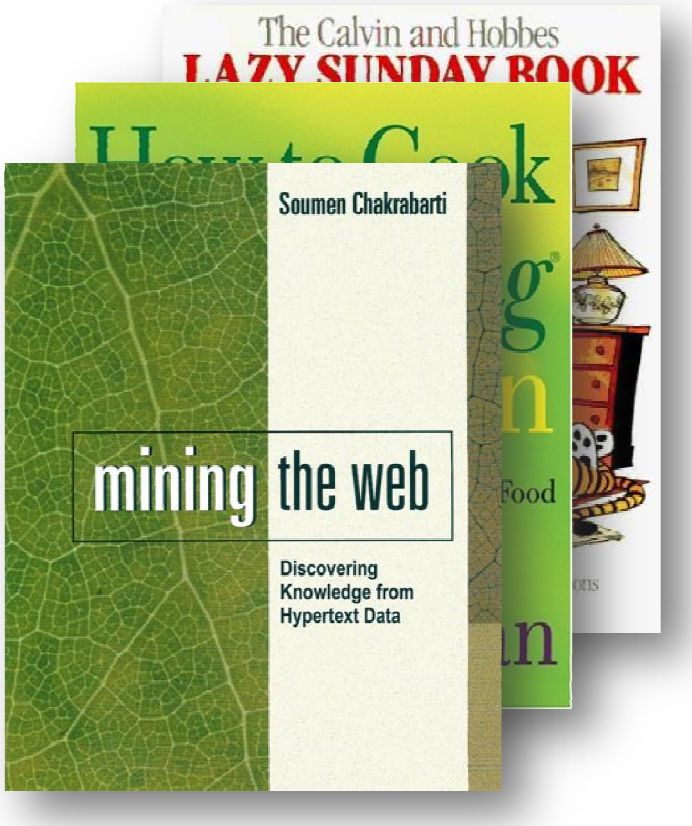
Dataset and Equivalence



Statistics

- 309,071 LibraryThing works
- 8,783 unique subject headings
- 47,957 unique tags

Dataset and Equivalence



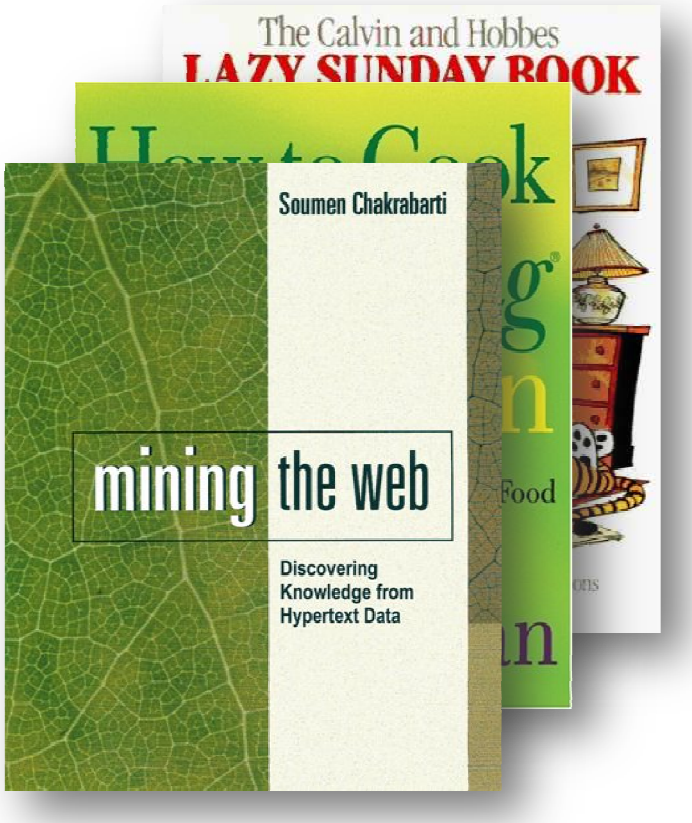
Statistics

- 309,071 LibraryThing works
- 8,783 unique subject headings
- 47,957 unique tags

Equivalence

- 48% of subject headings = tags
  (e.g., "data mining" and "data mining")

Dataset and Equivalence



Statistics

- 309,071 LibraryThing works
- 8,783 unique subject headings
- 47,957 unique tags

Equivalence

- 48% of subject headings = tags
  (e.g., "data mining" and "data mining")

(More headings approximately equal!)

Do Experts and Users Have Different Priorities?

Controlled

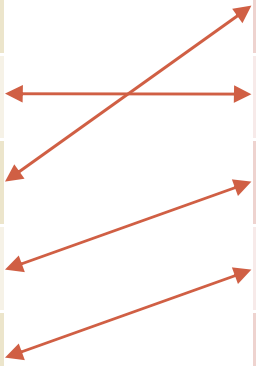
29,307	United States
23,341	History
21,216	Fiction
21,131	Biography
11,763	20th Century
10,878	Juvenile Literature
7,240	History and Criticism
6,346	Great Britain
…	…

Uncontrolled

Fiction	48,722
History	42,497
Biography	19,447
20th Century	17,756
Literature	17,744
Religion	13,504
Art	12,976
Science	11,105
…	…

Do Experts and Users Have Different Priorities?

Controlled		Uncontrolled	
29,307	United States	Fiction	48,722
23,341	History	History	42,497
21,216	Fiction	Biography	19,447
21,131	Biography	20th Century	17,756
11,763	20th Century	Literature	17,744
10,878	Juvenile Literature	Religion	13,504
7,240	History and Criticism	Art	12,976
6,346	Great Britain	Science	11,105
…	…	…	…

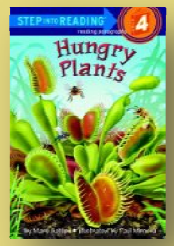


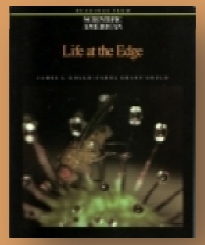
Kendall's Tau: 0.305 for shared terms

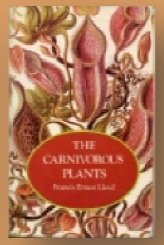
Overlap Example: Books Labeled with Library Term or Tag "Carnivorous Plants"

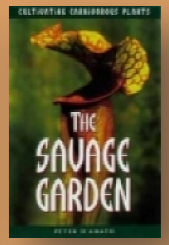
Library Term

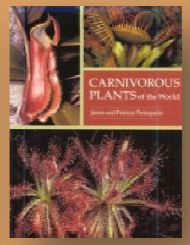
Tag

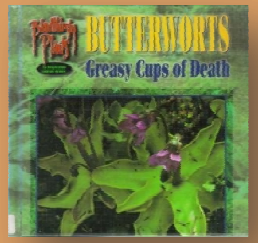


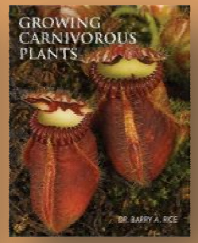


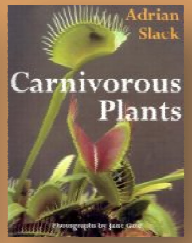


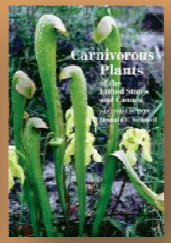


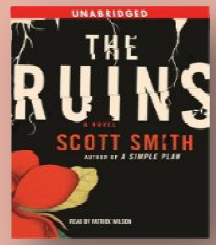








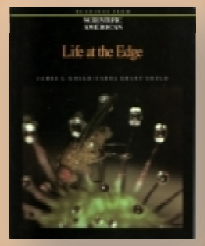


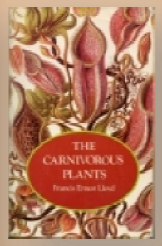


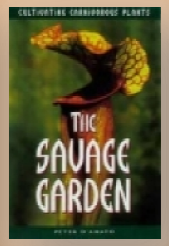
Overlap Example: Books Labeled with Library Term or Tag "Carnivorous Plants"

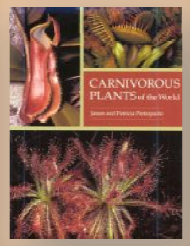
Library Term

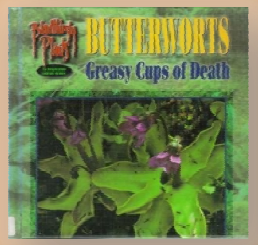
Tag

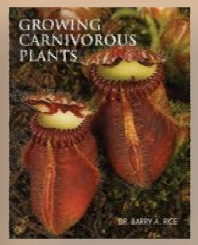


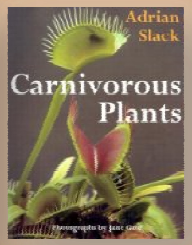


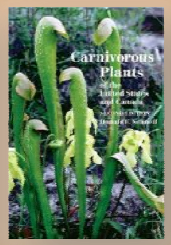


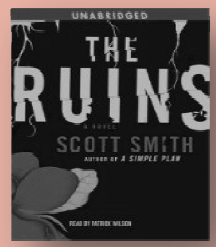






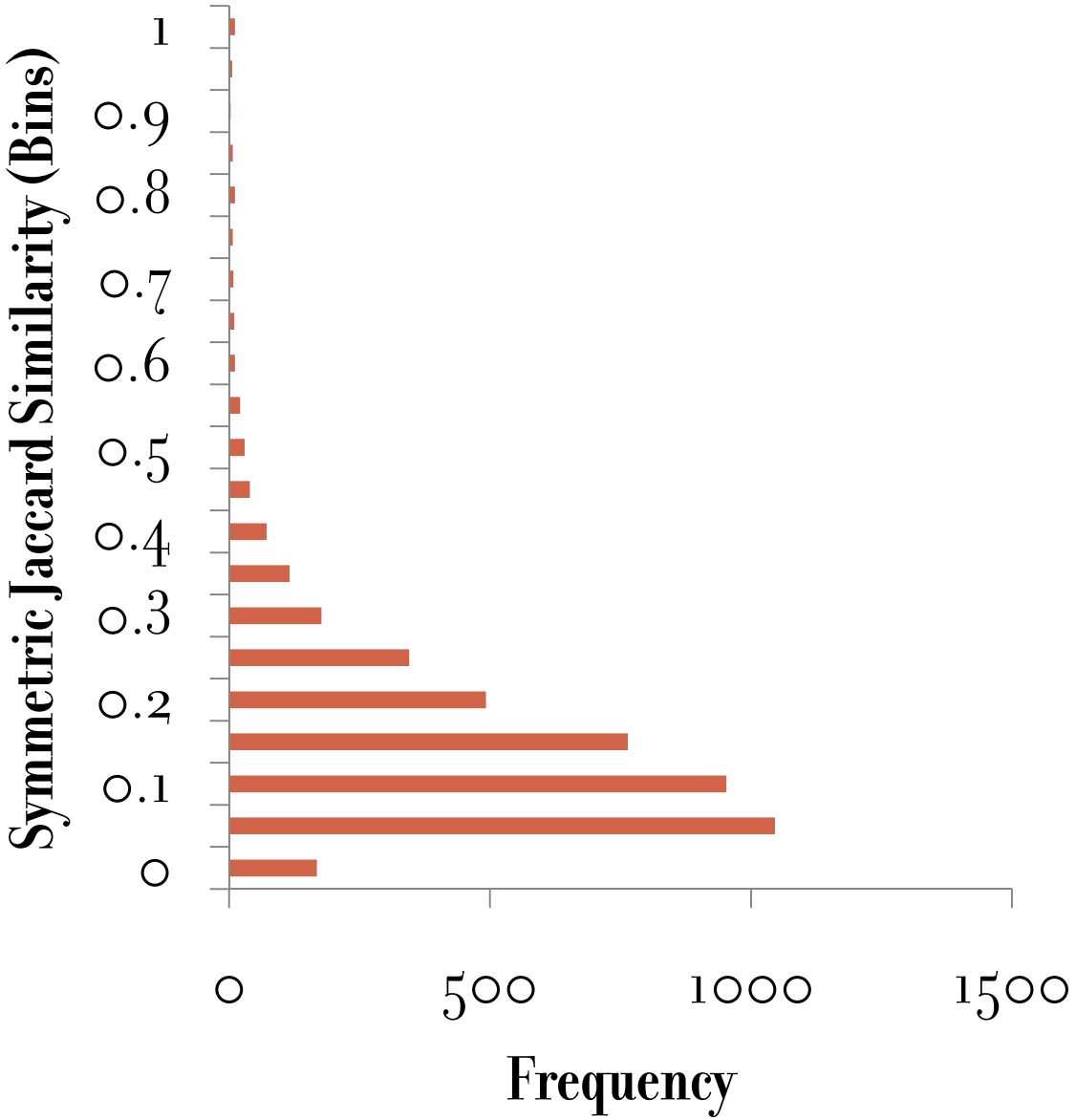




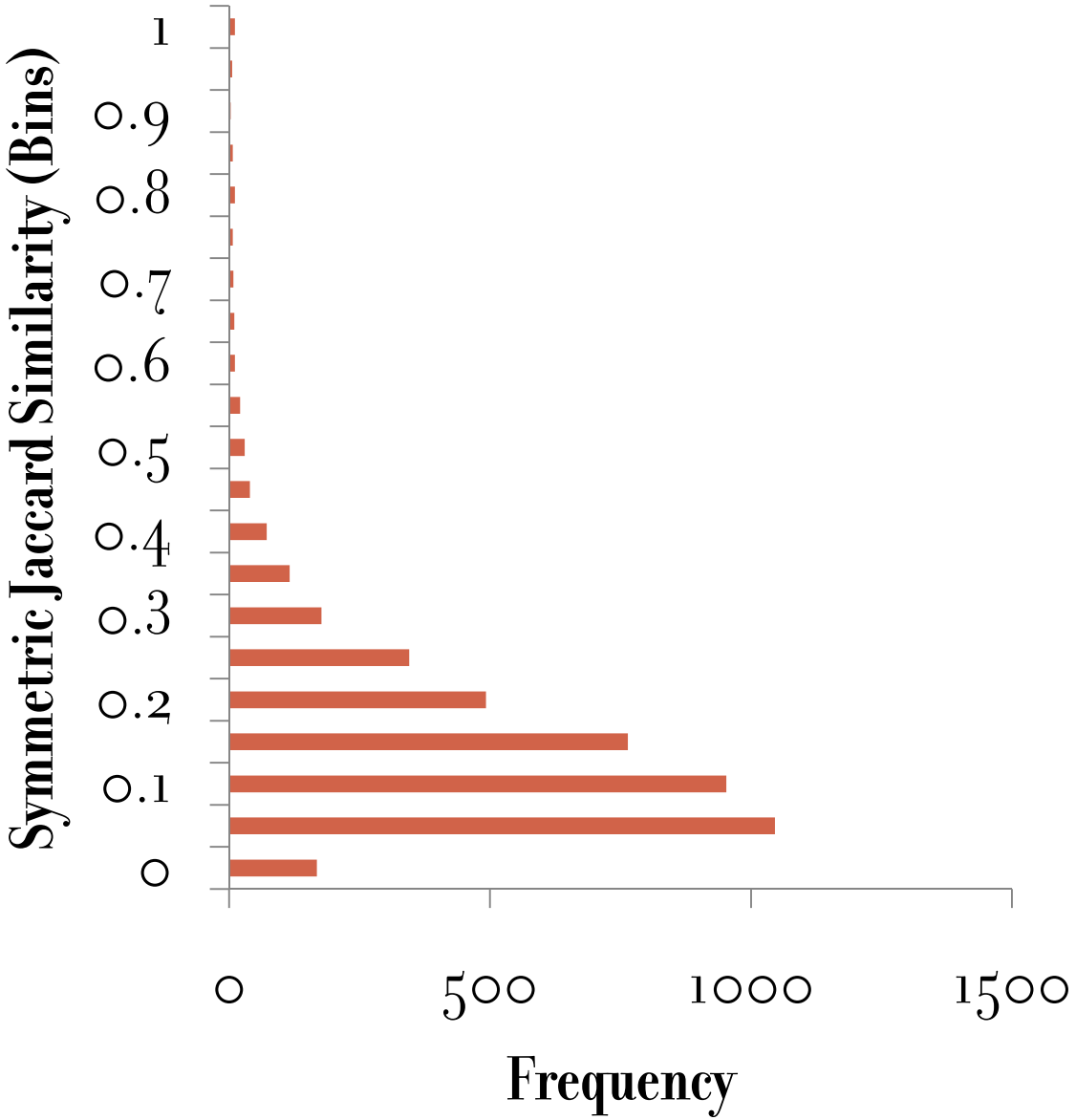


$$J_{sym} = \frac{O(t_i) \cap O(l_j)}{O(t_i) \cup O(l_j)} = \frac{8}{11}$$

Overlap Distributions



Overlap Distributions



Tag/Library Term	Jaccard
humorous stories	0.0
feminist fiction	0.0

Low Overlap Examples

Tag/Library Term	Jaccard
circus	0.13
electronics	0.13

Avg Overlap Examples

Tag/Library Term	Jaccard
origami	0.76
serging	0.77

High Overlap Examples

Vocabulary Problem Revisited

"People use a surprisingly great variety of words to refer to the same thing." --- Furnas et al., 1987

activities	whatsup
calendar	sparetime
events	funtime
cityevents	+ 100s of others

Vocabulary Problem Revisited

"People use a surprisingly great variety of words to refer to the same thing." --- Furnas et al., 1987

activities	whatsup
calendar	sparetime
events	funtime
cityevents	+ 100s of others

Corrollary : "Given the same words, people will annotate a surprising variety of objects." (2009)

Four Main Results

1. Experts and users use surprisingly similar terms.
   a) Approx. 50% syntactically the same.
   b) Add'l approx 10% semantically similar (by WikiESA).
2. More popular expert terms are more likely to be tags.
3. Experts terms and users tags have similar ranks.
4. But, experts and users differ in same term annotation.

Conclusions

1. Taggers tend to use a similar language to catalogers, suggesting that tags are meaningful.
2. At least for popular items in unspecialized collections (very common on the web), tagging appears to be very reasonable and effective.
3. Taggers apply terms differently from experts, suggesting that experts alone may not best label collections.

Questions?

http://infoblog.stanford.edu/