Tagging Human Knowledge

Paul Heymann, Andreas Paepcke,
and Hector Garcia-Molina
Department of Computer Science
Stanford University

February 4th, 2010

Outline

Talk Goals

1. Introduce library research methods
2. Explain what's missing on the web
3. Suggest how tags might help

Outline

Library Research Methods

Orthogonal ways to find information.

Standard Library Research Methods (Mann 2005)

Library Research Method	Web Counterpart
Keyword/Boolean Searching	
Encyclopedias	
Citation Searching	
Related Record Searching	
Subject Bibliographies	
People Sources	
Type of Literature Searching	
Browsing Bookstacks	
Controlled Vocabulary Search	

Standard Library Research Methods (Mann 2005)

Library Research Method	**Web Counterpart**
Keyword/Boolean Searching	Web Search
Encyclopedias	
Citation Searching	
Related Record Searching	
Subject Bibliographies	
People Sources	
Type of Literature Searching	
Browsing Bookstacks	
Controlled Vocabulary Search	

Standard Library Research Methods (Mann 2005)

Library Research Method	Web Counterpart
Keyword/Boolean Searching	Web Search
Encyclopedias	Wikipedia
Citation Searching	
Related Record Searching	
Subject Bibliographies	
People Sources	
Type of Literature Searching	
Browsing Bookstacks	
Controlled Vocabulary Search	

Standard Library Research Methods (Mann 2005)

Library Research Method	Web Counterpart
Keyword/Boolean Searching	Web Search
Encyclopedias	Wikipedia
Citation Searching	Forward Links
Related Record Searching	
Subject Bibliographies	
People Sources	
Type of Literature Searching	
Browsing Bookstacks	
Controlled Vocabulary Search	

Standard Library Research Methods (Mann 2005)

Library Research Method	Web Counterpart
Keyword/Boolean Searching	Web Search
Encyclopedias	Wikipedia
Citation Searching	Forward Links
Related Record Searching	Back Links/Similar Pages
Subject Bibliographies	
People Sources	
Type of Literature Searching	
Browsing Bookstacks	
Controlled Vocabulary Search	

Standard Library Research Methods (Mann 2005)

Library Research Method	Web Counterpart
Keyword/Boolean Searching	Web Search
Encyclopedias	Wikipedia
Citation Searching	Forward Links
Related Record Searching	Back Links/Similar Pages
Subject Bibliographies	Curated Links
People Sources	
Type of Literature Searching	
Browsing Bookstacks	
Controlled Vocabulary Search	

Standard Library Research Methods (Mann 2005)

Library Research Method	Web Counterpart
Keyword/Boolean Searching	Web Search
Encyclopedias	Wikipedia
Citation Searching	Forward Links
Related Record Searching	Back Links/Similar Pages
Subject Bibliographies	Curated Links
People Sources	Social Search
Type of Literature Searching	
Browsing Bookstacks	
Controlled Vocabulary Search	

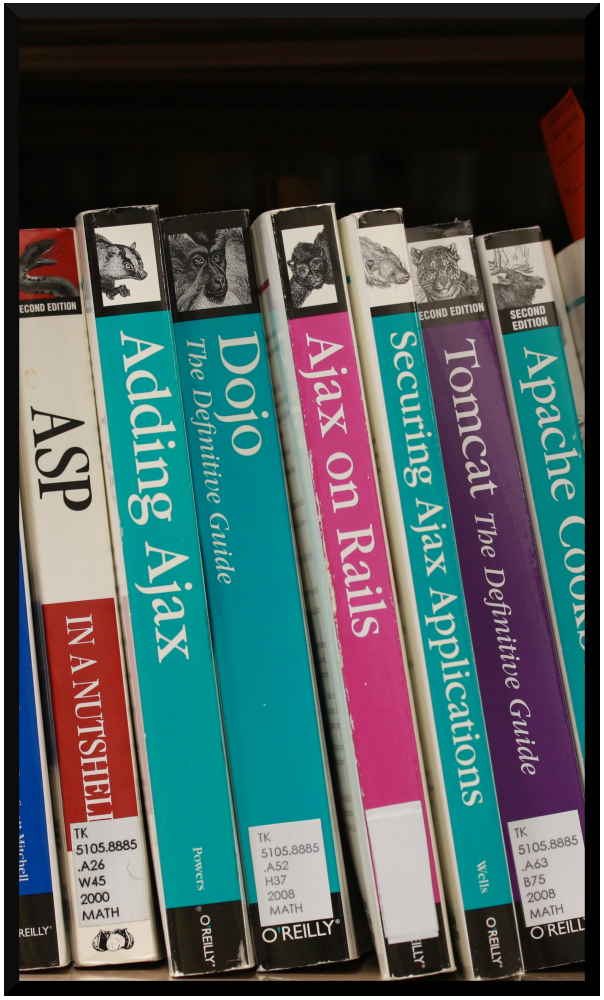
Standard Library Research Methods (Mann 2005)

Library Research Method	Web Counterpart
Keyword/Boolean Searching	Web Search
Encyclopedias	Wikipedia
Citation Searching	Forward Links
Related Record Searching	Back Links/Similar Pages
Subject Bibliographies	Curated Links
People Sources	Social Search
Type of Literature Searching	Vertical Search
Browsing Bookstacks	
Controlled Vocabulary Search	

Standard Library Research Methods (Mann 2005)

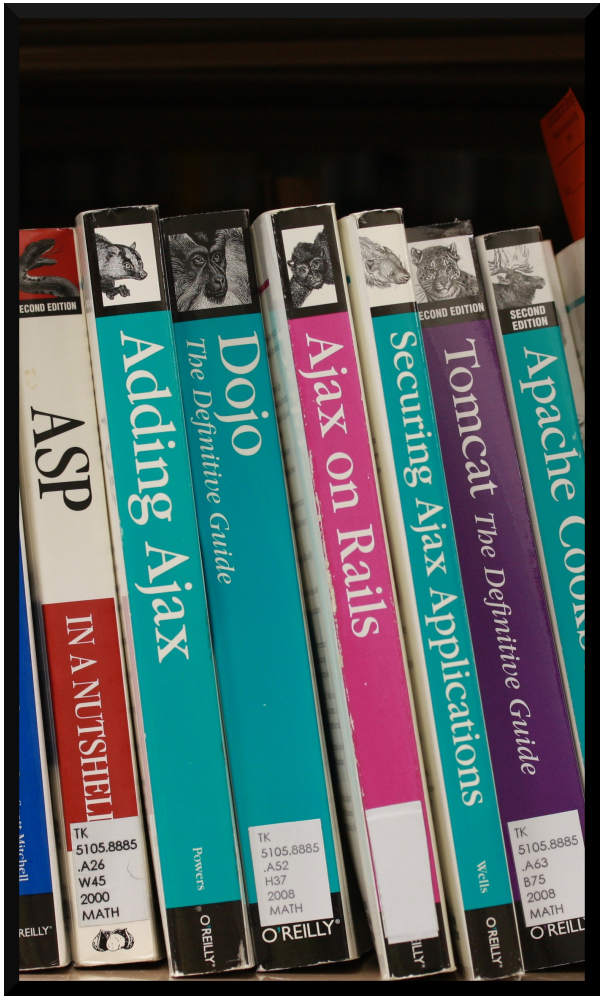
Library Research Method	**Web Counterpart**
Keyword/Boolean Searching	Web Search
Encyclopedias	Wikipedia
Citation Searching	Forward Links
Related Record Searching	Back Links/Similar Pages
Subject Bibliographies	Curated Links
People Sources	Social Search
Type of Literature Searching	Vertical Search
Browsing Bookstacks	~~Directories?~~ Tags?
Controlled Vocabulary Search	Tags?

Classified Bookstacks Browsing (i.e., Taxonomy)



Ajax TK5105.8885.A52

Classified Bookstacks Browsing (i.e., Taxonomy)



Ajax	TK5105.8885.A52
Special, A-Z	TK5105.8885.A-Z
Web authoring software	TK5105.8883-8885
World Wide Web	TK5105.888-8885
Specific aspects of, or services on, the Internet.	TK5105.8762-8887
Wide area networks	TK5105.87-8887
Computer networks	TK5105.5-9
Telecommunication	TK5101.0-9
Electrical engineering, Electronics, Nuclear engineering.	TK
Technology	T

Classified Bookstacks Browsing (i.e., Taxonomy)

Pros:

1. Serendipity!
2. Corpus overview!

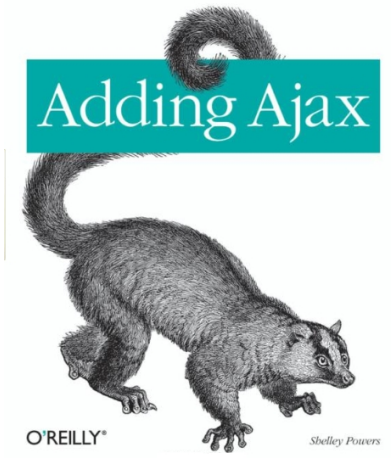
Cons:

1. Expensive
2. Hard to change

Taxonomies help us comprehend, browse whole collections.

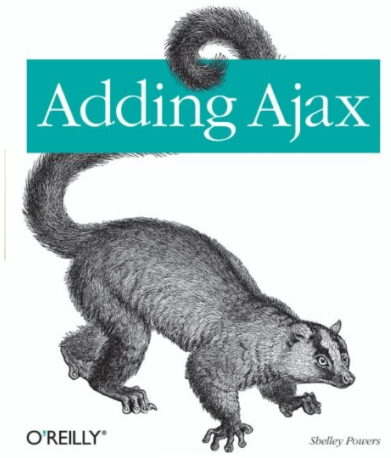
Controlled Vocabulary Searching

Title	Adding Ajax
Author	Powers, Shelley
Term	Ajax (Web site development ...)
Term	Web site development



Controlled Vocabulary Searching

Title	Adding Ajax
Author	Powers, Shelley
Term	Ajax (Web site development ...)
Term	Web site development

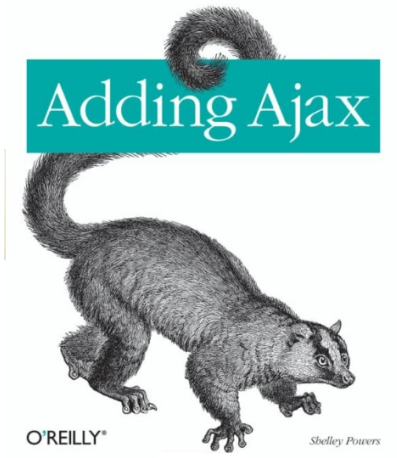


Web site development

UF	Development of Web sites
BT	Internet programming
NT	Ajax (...)
NT	Document Object Model (...)
NT	Mason (...)

Controlled Vocabulary Searching

Title	Adding Ajax
Author	Powers, Shelley
Term	Ajax (Web site development ...)
Term	Web site development

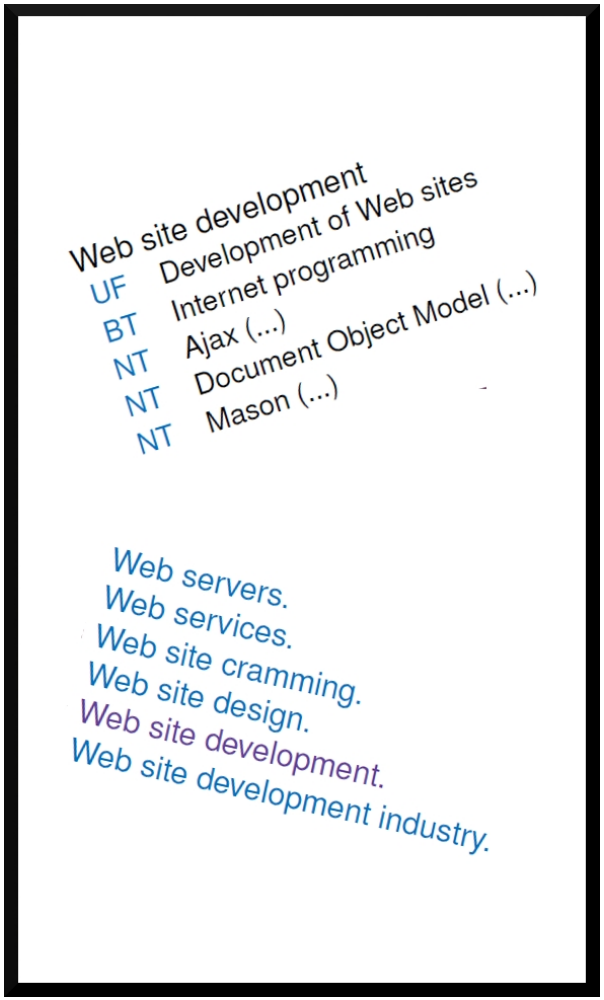


Web site development

UF	Development of Web sites
BT	Internet programming
NT	Ajax (...)
NT	Document Object Model (...)
NT	Mason (...)

Web servers.
Web services.
Web site cramming.
Web site design.
Web site development.
Web site development industry.

Controlled Vocabulary Searching



Pros:

1. Expand from item (by topic)
2. Expand from topic

Cons:

1. Taxonomists apply terms
2. Terms hard to find

Controlled vocabularies help us expand from a single item, document.

Outline

Research Question

Can tags provide some of what has been lost by not having a taxonomy or controlled vocabulary terms for the web?

Two Aspects of Tagging

Interface

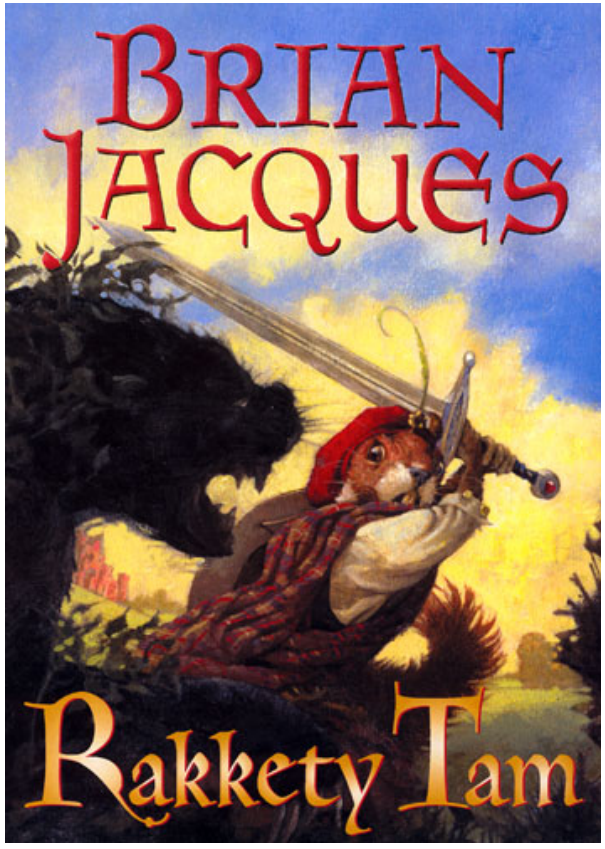
Data (our focus)

1. Terms
2. Structure
3. Topics

This Work

1. Analyzes books (not URLs!)
2. Compares tags to taxonomies, controlled vocabulary
   (i) Synonymy
   (ii) Paid labelers
   (iii) Tag types
   (iv) User preferences
   (v) Topic overlap
   (vi) Information integration
3. Tagging fares well in these comparisons

Data



Library of Congress (2×10^6 records)

LCSH	squirrels, fantasy, animals
LCC	PZ7.J15317 Rak 2004
DDC	[Fic]

LibraryThing (3×10^5 works)

user tags	redwall, children's, anthropomorphic fantasy

Goodreads (7×10^3 ISBNs)

user tags	fantasy, redwall, young-adult

Mechanical Turk (1×10^4 $-tags)

paid tags	sword, champion, adventure

Outline

Synonymy Examples

P(t)	tag
0.99	homeschool
< 0.01	homeschooling
< 0.01	home school
< 0.01	home_school

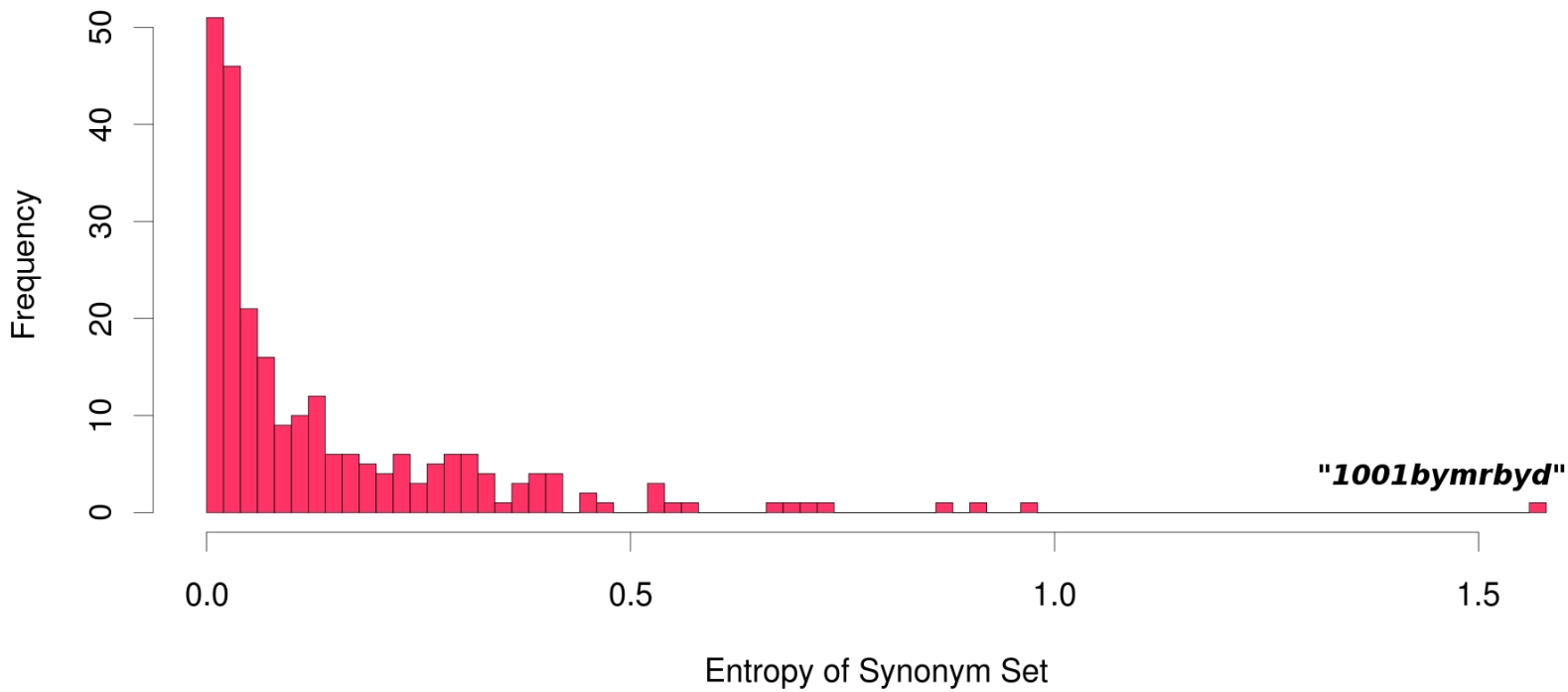
(entropy < 0.1)

P(t)	tag
0.55	1001bymrbfd
0.26	1001 books you must ...
0.11	1001 books to read ...
0.07	1001bymrbyd

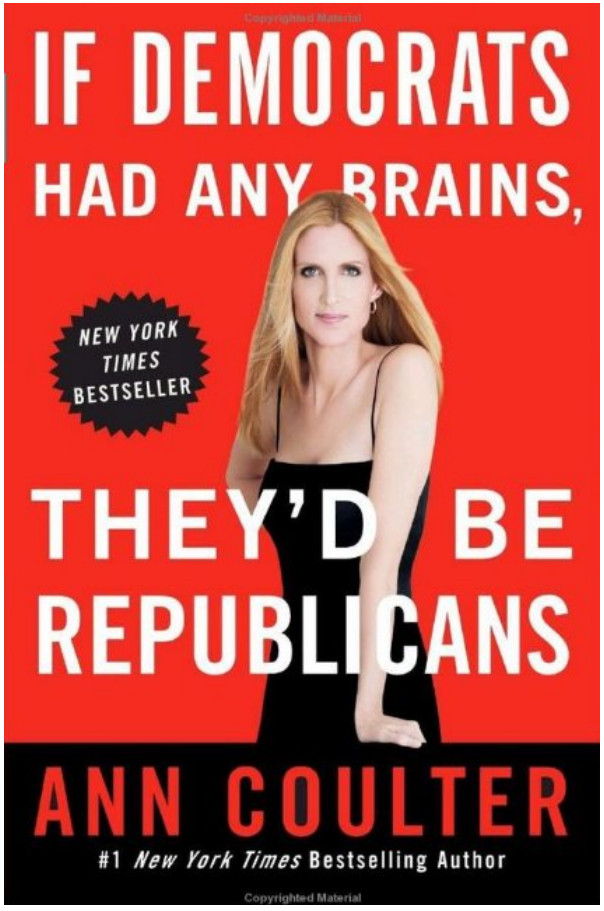
(entropy ≈ 1.5)

Key Idea: Calculate entropy of probability distribution assuming a user chooses a tag at random in proportion to frequency.

Synonymy Entropy Distribution (Top Tags)



Tag Quality?!



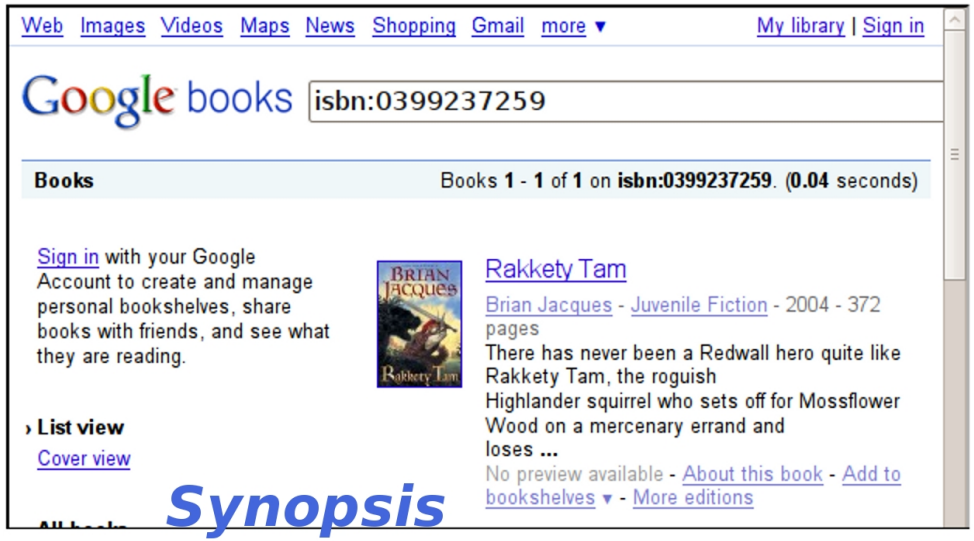
horrible (180),
why america is hated (152),
humor (128),
intelligent (122),
honest (109),
comedy (103),
truth (102),
accurate (96),
wingnut welfare (87),
patriotic (85),
patriot (55),
keeping america stupid (20),
ann coulter (19),
delusional (19),
evil (16),
stupid (16),
conservative (15)

Tag Type Distribution

	LT%	GR%
Objective, Content of Book	60.55	57.10
Personal or Related to Owner	6.15	22.30
Acronym	3.75	1.80
Unintelligible or Junk	3.65	1.00
Physical (e.g., “Hardcover”)	3.55	1.00
Opinion (e.g., “Excellent”)	1.80	2.30
None of the Above	0.20	0.20
No Annotator Majority	20.35	14.30
Total	100	100

Key Idea: Most tags describe content objectively.

Perceived Tag Helpfulness



Tag/Keyword	Not at all helpful 1	2	3	4	5	6	Extremely Helpful 7		Don't Understand / Other
sword	○	○	○	○	○	○	○		○
fantasy	○	○	○	○	○	○	○		○
anthropomorphic fantasy	○	○	○	○	○	○	○		○
squirrels	○	○	○	○	○	○	○		○
champion	○	○	○	○	○	○	○		○
redwall	○	○	○	○	○	○	○		○
adventure	○	○	○	○	○	○	○		○
animals	○	○	○	○	○	○	○		○
children's	○	○	○	○	○	○	○		○

Evaluation

Perceived Tag Helpfulness

	μ
$-tags	4.93
Rare User Tags	4.23
Moderate User Tags	5.80
Common User Tags	5.27
LCSH Main Topics	5.13

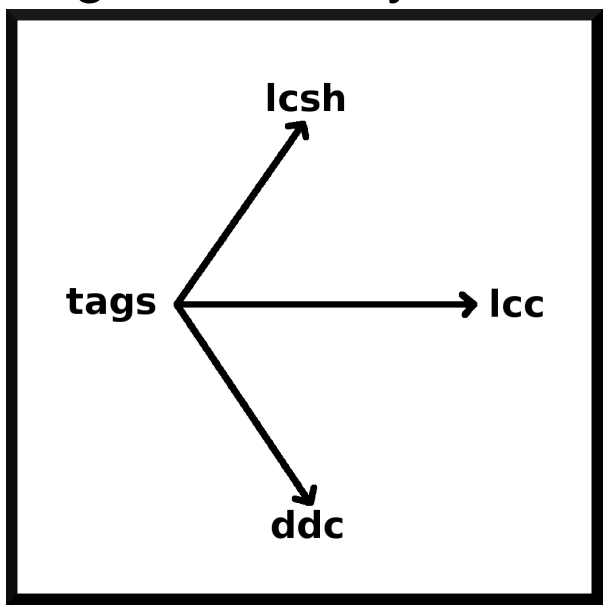
Key Idea 1: Paid taggers can supplement regular users.
Key Idea 2: Medium frequency tags are most valuable.

Also In The Paper

System ↔ System

Key Idea: Federation.

Tags ↔ Library Terms



Key Idea: Similar topics.

Outline

Conclusion

1. Library methods can inform web thinking
2. We lack some web counterparts
3. Tagging may be able to help
   (a) *Interface:* Tag cloud, browsable
   (b) *Data:* Little “problematic” synonymy
   (c) *Data:* Good tag types
   (d) *Data:* Terms perceived helpful
   (e) *Data:* Paid tagging
   (f) *Data:* Good topics
   (g) *Data:* Federation

Conclusion

1. Library methods can inform web thinking
2. We lack some web counterparts
3. Tagging may be able to help
   (a) *Interface:* Tag cloud, browsable
   (b) *Data:* Little “problematic” synonymy
   (c) *Data:* Good tag types
   (d) *Data:* Terms perceived helpful
   (e) *Data:* Paid tagging
   (f) *Data:* Good topics
   (g) *Data:* Federation

Questions?
Visit http://heymann.stanford.edu/ or
http://ilpubs.stanford.edu/ for more.