## Tagging Human Knowledge

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

February 4th, 2010

▲□▶▲□▶▲□▶▲□▶ □ のQ@

## Outline

#### Introduction

Library Research Methods

Our Approach

Our Work

Conclusion



## Talk Goals

- 1. Introduce library research methods
- 2. Explain what's missing on the web

(日) (日) (日) (日) (日) (日) (日)

3. Suggest how tags might help

## Outline

Introduction

#### Library Research Methods

**Our Approach** 

Our Work

Conclusion

◆□▶ ◆□▶ ◆ □▶ ◆ □▶ ● □ ● ● ●

Library Research Methods

# Orthogonal ways to find information.

▲□▶ ▲□▶ ▲□▶ ▲□▶ ▲□ ● のへぐ

Library Research Method Web Counterpart

(日) (日) (日) (日) (日) (日) (日)

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

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

Web Counterpart Web Search

・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・

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

Web Counterpart Web Search Wikipedia

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

Web Counterpart Web Search Forward Links

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

Web Counterpart Web Search Wikipedia Forward Links Back Links/Similar Pages

▲□▶ ▲□▶ ▲□▶ ▲□▶ □ ● のへで

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

Web Counterpart Web Search Wikipedia Forward Links Back Links/Similar Pages Curated Links

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

Web Counterpart Web Search Wikipedia Forward Links Back Links/Similar Pages Curated Links Social Search

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

Web Counterpart Web Search Wikipedia Forward Links Back Links/Similar Pages Curated Links Social Search Vertical Search

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

Web Counterpart Web Search Wikipedia Forward Links Back Links/Similar Pages Curated Links Social Search Vertical Search Directories? Tags? Tags?

・ロト・日本・モート ヨー うへの

# Classified Bookstacks Browsing (i.e., Taxonomy)



Ajax TK5105.8885.A52

イロト イポト イヨト イヨト

ъ

# Classified Bookstacks Browsing (i.e., Taxonomy)



Web authoring software World Wide Web Specific aspects of, TK5105.8762-8887 or services on. the Internet Wide area networks Computer networks Telecommunication Electrical engineering, Electronics. Nuclear engineering. Technology

Ajax

TK5105.8885.A52 Special, A-Z TK5105.8885.A-Z TK5105.8883-8885 TK5105.888-8885

> TK5105.87-8887 TK5105.5-9 TK5101.0-9 TK

Т

ヘロマ ヘ動 マイロマー

# 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.

ヘロト ヘポト ヘヨト ヘヨト

TitleAdding AjaxAuthorPowers, ShelleyTermAjax (Web site development ...)TermWeb site development



▲□▶▲□▶▲□▶▲□▶ □ のQ@

TitleAdding AjaxAuthorPowers, ShelleyTermAjax (Web site development ...)TermWeb site development



◆□▶ ◆□▶ ▲□▶ ▲□▶ ■ ののの

Web site development

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

TitleAdding AjaxAuthorPowers, ShelleyTermAjax (Web site development ...)TermWeb 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.



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

Introduction

Library Research Methods

#### Our Approach

Our Work

Conclusion

◆□▶ ◆□▶ ◆ □▶ ◆ □▶ ● □ ● ● ● ●

# 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

conomics Education excesses Egypt destrone element dopedia trasis engineering England English teats foton glish Literature engineered exercisered Environment a interrology work: Erotica Inchatology Espionage essay Essay rope to see European history evenuese Evolution extenses alism exploration f tables feerie fairies fairy tale Fairy 7 mily Fantasy fantaxy fiction farm farming fashion Favorit make author brook and areas Feminism service Fic Fict Finance Finished man First Edition on nong too ty Folklore folktales Food toot and drink footbel for sele foreign i and a France researce French reach return french because uses friends Friendship as funny resitive future teachers goal ten Gardening gentees gave away Gay gay fatton gay men gend Ogy General General Fiction orienter Geography terman theorem Germany Geschichte peschiede ift pin girls gibt plotestation God pode Gelf pane good Gothic govern design Graphic Novel Graphic Novels graphic Great anthology gross grief growing up Guide guidebool

Tag info: fantasy Induire tetary, fartary, Martary, Fartary, Fartary, Gener, Fartary, Gener, Janlary, Phartary, Tantary,	LT tog doud   LT author dou
Yantanyi, Sanasty, Sanastan, Sanaten, Sanatan, Santan, Santan, Santanin, Kartanin, Kartanyi, Santanin, Santani, Santanyi, Santon, Santon, Kartan, Karan, Jama, Jama - Santanyi, parine fartumi, Jitemburu/Jantanyi Johat D.	Related tags (show runtum) advanture (british
Tag and its alianes used 1,303,676 times by 25,735 users.	children's children's children's
Most often tapped fantaxy	Heroture dessic
The Hobbit, or, There and Bodi Again by J. P. Tollows (sets) Heary Potter will the Philosopher's Bose by J. K. Powhy (sets) Havy Potter and the Half-Blood Prince by J. K. Powhy (sets) Havy Potter and the Chaft of the Phoenic by J. K. Powhy (sets) Havy Potter and the Chaft of the Phoenic by J. K. Powhy (sets)	classic discworld chapters indust fantasy fiction Harry Potter horse humor harrour industry Lord of
K. Rowing (1990) Harry Potter and the Prisoner of Atilaban by 3. K. Rowing (1930) Harry Potter and the Gobiet of Pre by 3. K. Powing (1941)	the Roge magic Mode Carth Investory Namia Rovel Low Protector (1980) Jointy
Harry Potter and the Deathly Hallows by J. K. Powling (2001) The Petrowship of the Pang by J. P. R. Tolkies (400) The Lord of the Pang by J. P. P. Tolkies (440)	romance settle SOBIOE fection SettleS Tobles vampines

▲□▶▲□▶▲□▶▲□▶ □ のQ@

#### Data (our focus)

- 1. Terms
- 2. Structure
- 3. Topics

# This Work

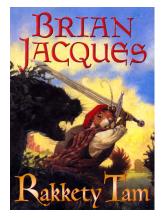
## 1. Analyzes books (not URLs!)

### 2. Compares tags to taxonomies, controlled vocabulary

◆□▶ ◆□▶ ▲□▶ ▲□▶ ■ ののの

- (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  $(2 \times 10^6 \text{ records})$ LCSH squirrels, fantasy, animals LCC PZ7.J15317 Rak 2004 DDC [Fic] LibraryThing  $(3 \times 10^5 \text{ works})$ user tags redwall, children's, anthropomorphic fantasy **Goodreads**  $(7 \times 10^3 \text{ ISBNs})$ user tags fantasy, redwall, young-adult Mechanical Turk  $(1 \times 10^4 \text{-tags})$ paid tags sword, champion, adventure

## Outline

Introduction

Library Research Methods

**Our Approach** 

Our Work

Conclusion

◆□ > ◆□ > ◆ □ > ◆ □ > ● □ ● ● ● ●

# 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 ...

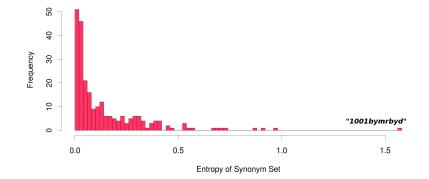
 0.11
 1001 books to read ...

0.07 1001bymrbyd

(entropy  $\approx$  1.5)

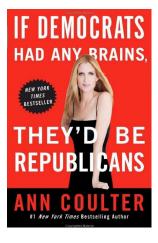
Key Idea: Calculate entropy of probability distribution assuming a user chooses a tag at random in proportion to frequency.

# Synonymy Entropy Distribution (Top Tags)



◆□▶ ◆□▶ ◆三▶ ◆三▶ ● 三 のへで

# 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)

◆□▶ ◆□▶ ▲□▶ ▲□▶ □ のQ@

# 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

animals

children's



Evaluation

## Perceived Tag Helpfulness

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

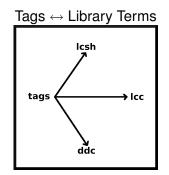
(ロ) (同) (三) (三) (三) (○) (○)

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

## Also In The Paper

System $\leftrightarrow$ System				
LibraryThing				
toread		to-read		
tbr	2009			
fantasy	fiction	currently-reading		
sf literature	wishlist	non-fiction		
religion	own			
art		sci-fi-fantasy		
goodreads*				
	90	oaroado		

Key Idea: Federation.



Key Idea: Similar topics.

▲□▶ ▲□▶ ▲ 三▶ ▲ 三▶ - 三 - のへぐ

## Outline

Introduction

Library Research Methods

**Our Approach** 

Our Work

Conclusion

◆□ ▶ ◆□ ▶ ◆ □ ▶ ◆ □ ▶ ● □ ● ● ● ●

# 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

◆□▶ ◆□▶ ▲□▶ ▲□▶ ■ ののの

- (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

(日) (日) (日) (日) (日) (日) (日)

- (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.