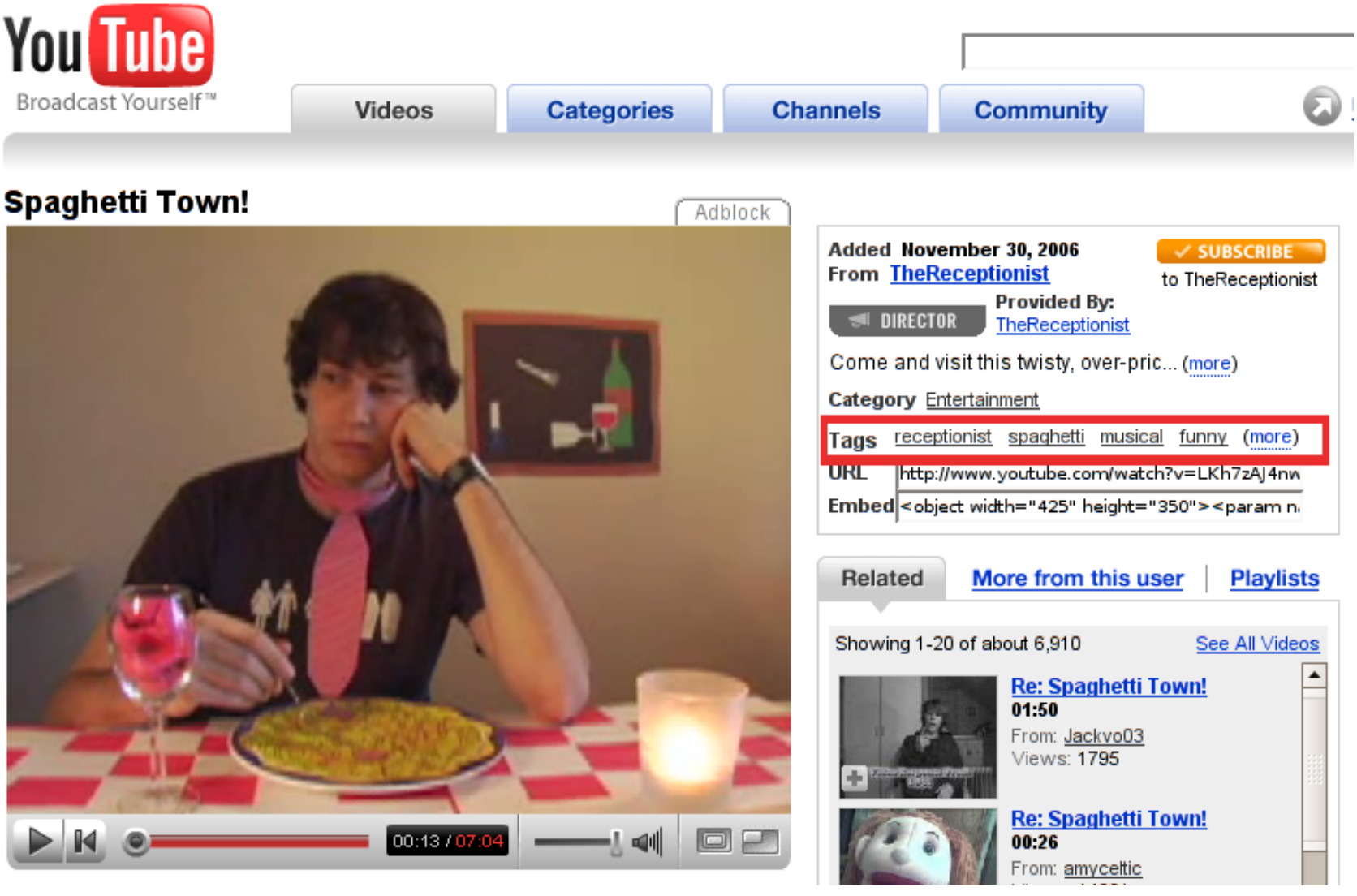
Can Social Bookmarking Improve Web Search?

Paul Heymann, Georgia Koutrika, and Hector Garcia-Molina
Department of Computer Science
Stanford University

February 12th, 2008

Outline

YouTube


YouTube
Broadcast Yourself™
Videos
Categories
Channels
Community
Spaghetti Town!
Adblock
00:13 / 07:04
Added November 30, 2006
From TheReceptionist
SUBSCRIBE
to TheReceptionist
DIRECTOR
Provided By:
TheReceptionist
Come and visit this twisty, over-pric... (more)
Category Entertainment
Tags receptionist spaghetti musical funny (more)
URL http://www.youtube.com/watch?v=LKh7zAJ4nw
Embed <object width="425" height="350"><param n
Related
More from this user
Playlists
Showing 1-20 of about 6,910
See All Videos
Re: Spaghetti Town!
01:50
From: Jackvo03
Views: 1795
Re: Spaghetti Town!
00:26
From: amyceltic

Amazon

The Art of Computer Programming, Volumes 1-3 Boxed Set (Hardcover)

by Donald E. Knuth (Author)

(41 customer reviews)

List Price: $179.99

Price: **$179.99** & this item ships for **FREE with Super Saver Shipping**. Details

Availability: In Stock. Ships from and sold by **Amazon.com**. Gift-wrap available.

Want it delivered Friday, March 2? Order it in the next 14 hours and 29 minutes, and choose **One-Day Shipping** at checkout. See details

29 used & new available from $143.99

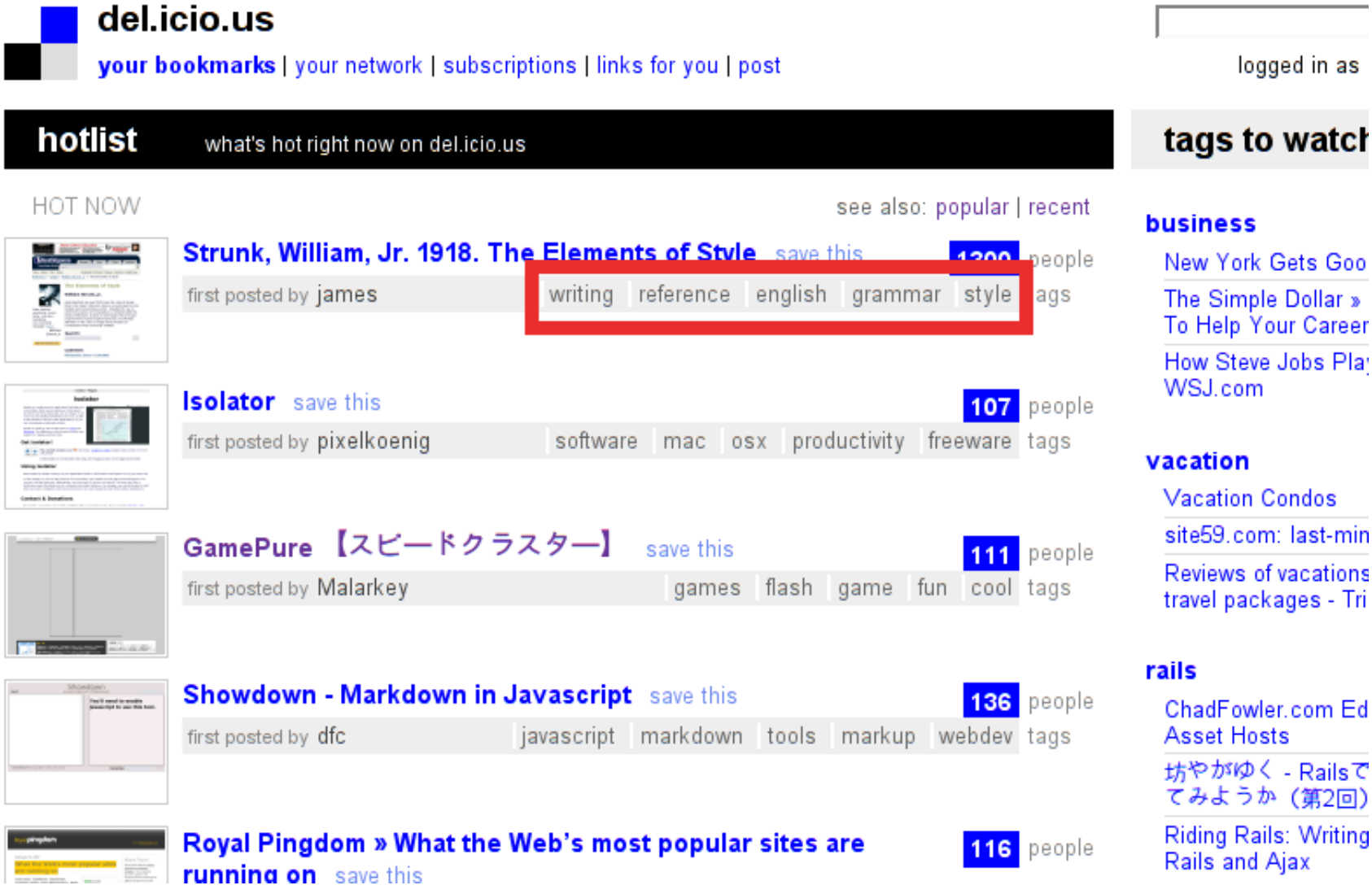
Customers tagged this product with

First tag: genius (M. Rubens "yoga instructor" on Nov 16, 2005)
Last tag: computer science

Sort by: Popularity GO!

computer science (4)	classic (1)	knuth (1)
algorithms (2)	comprehensive (1)	mathematics (1)
programming (2)	computer books (1)	max christmas (1)

del.icio.us


del.icio.us
your bookmarks | your network | subscriptions | links for you | post
logged in as
hotlist
what's hot right now on del.icio.us
tags to watch
HOT NOW
see also: popular | recent
Strunk, William, Jr. 1918. The Elements of Style
save this
people
first posted by james
writing reference english grammar style
tags
Isolator
save this
107
people
first posted by pixelkoenig
software mac osx productivity freeware
tags
GamePure 【スピードクラスター】
save this
111
people
first posted by Malarkey
games flash game fun cool
tags
Showdown - Markdown in Javascript
save this
136
people
first posted by dfc
javascript markdown tools markup webdev
tags
Royal Pingdom » What the Web's most popular sites are running on
save this
116
people
business
New York Gets Goo
The Simple Dollar »
To Help Your Career
How Steve Jobs Play
WSJ.com
vacation
Vacation Condos
site59.com: last-min
Reviews of vacations
travel packages - Tri
rails
ChadFowler.com Ed
Asset Hosts
坊やがゆく - Railsで
てみようか（第2回）
Riding Rails: Writing
Rails and Ajax

Outline

Can social bookmarking improve web search?

Subproblems

Are there “enough” URLs?

Are there “enough” tags?

Are the URLs *valuable*?

Are the tags *redundant*?

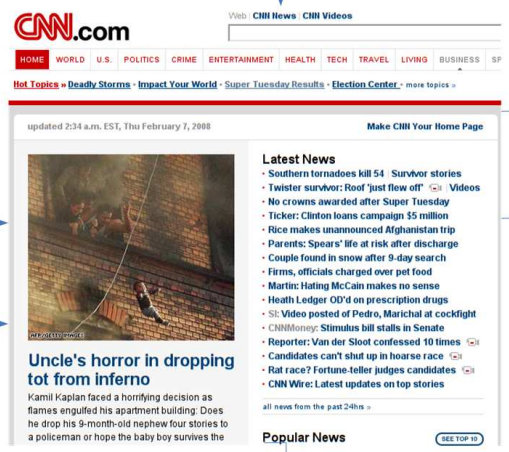
Tags versus Other Content

Back Link Text

… He is also a CNN Contributor, appearing on a variety of shows, including The Situation Room, Anderson Cooper 360, Lou Dobbs Tonight, and many others...

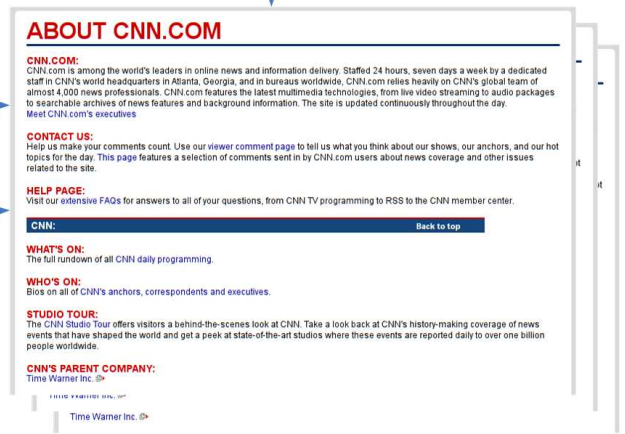
Page Text

CNN.com is among the world's leaders in online news and information delivery. Staffed 24 hours, seven days a week by a dedicated staff in CNN's world headquarters in Atlanta, Georgia, ...



Forward Link Text

CNN.com is among the world's leaders in online news and information delivery. Staffed 24 hours, seven days a week by a dedicated staff in CNN's world headquarters in Atlanta, Georgia, ...



Tags

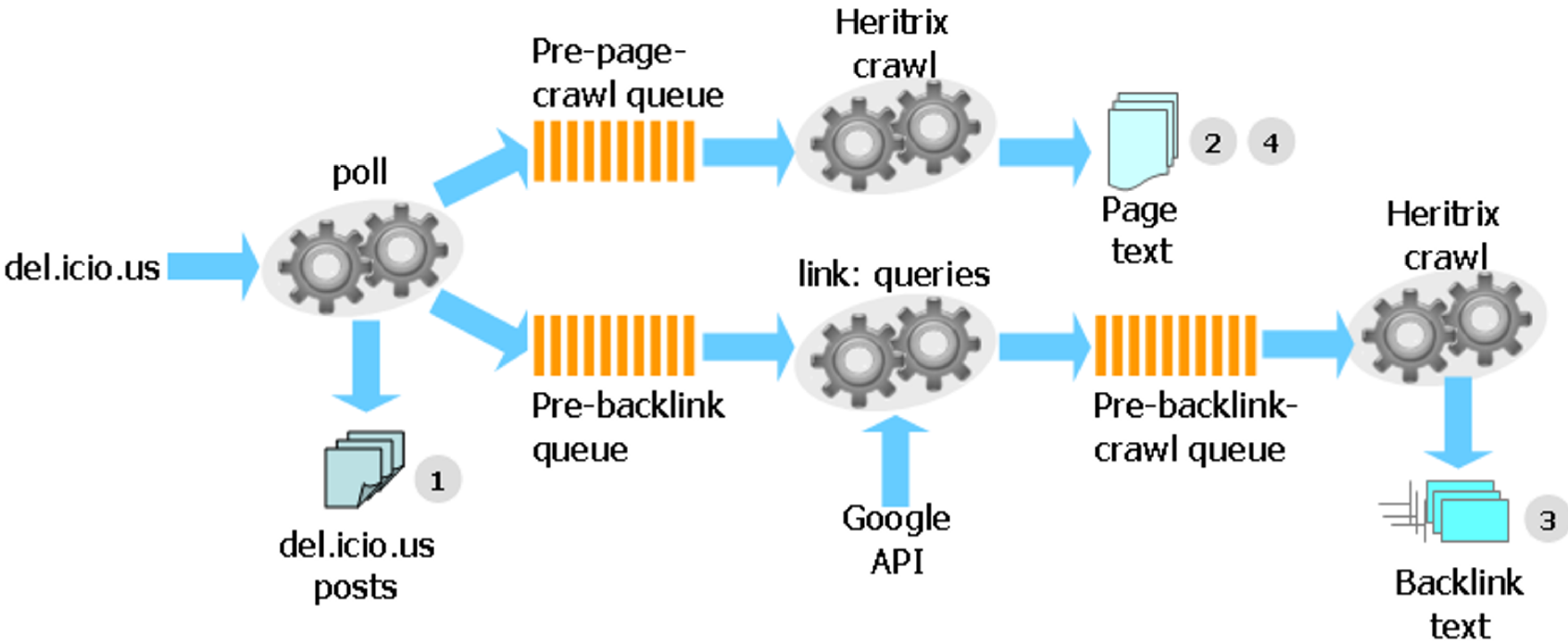
news cnn daily media

Outline

del.icio.us posts

Bookmarks/Posts	Triples
paul: news, uk → bbc.co.uk *08:33:25*	(paul, news, bbc.co.uk) (paul, uk, bbc.co.uk)
mary: recipes, food → food.com *08:33:23*	(mary, recipes, food.com) (mary, food, food.com)
dave: tv, cnn, news → cnn.com *08:33:21*	(dave, tv, cnn.com) (dave, cnn, cnn.com) (dave, news, cnn.com)

Realtime Web Crawling



Outline

Size and Growth

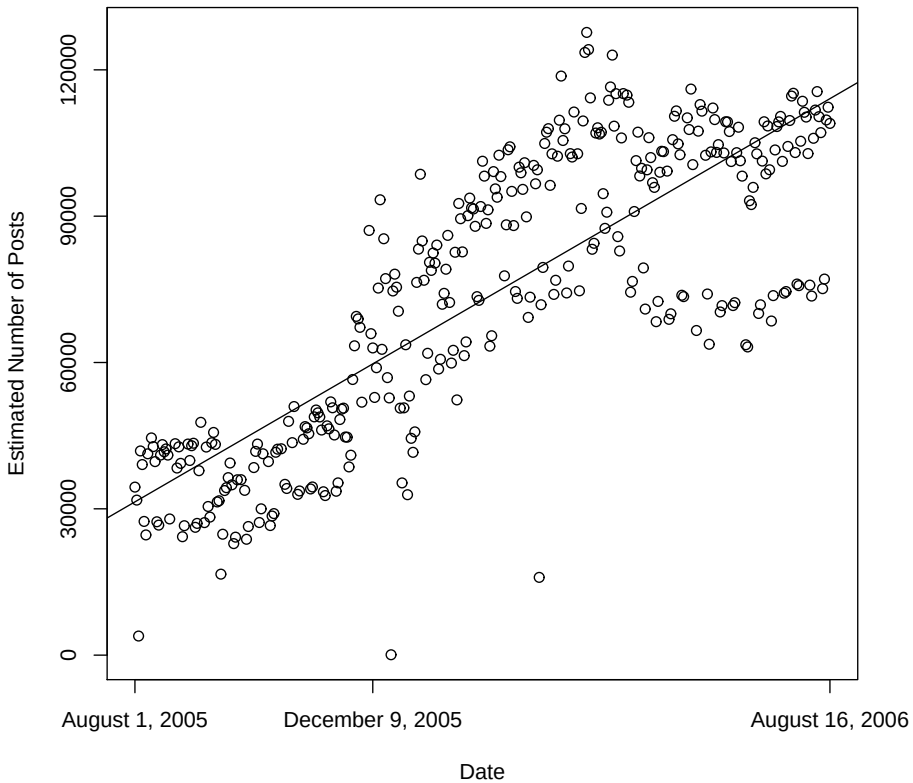
≈ 120 thousand (≈ 10^5) posts/day

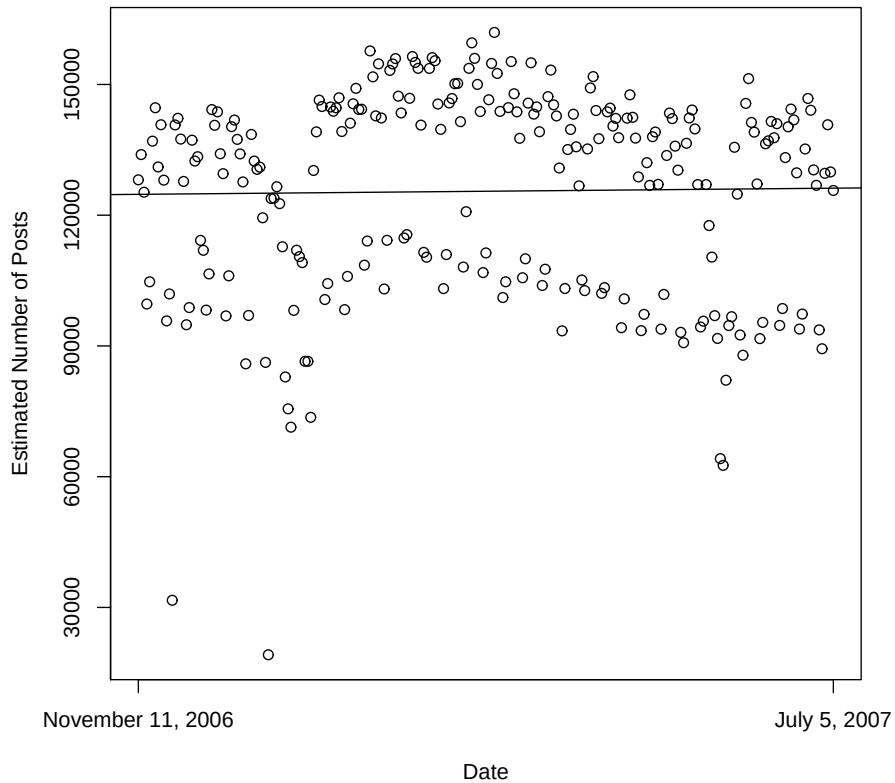
(versus ≈ 10^6 blog posts/day)

60–150 million posts

12–75 million (≈ 10^7–10^8) unique URLs

(versus ≈ 10^9–10^{11} total URLs)

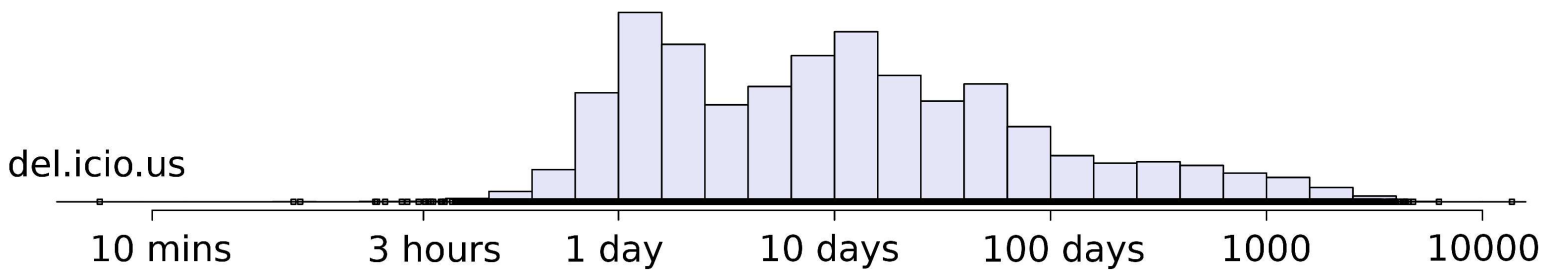




URL Indexing and Age

Found Initially	57.5%
Indexed Within 4 Weeks	12.75%
Indexed Within 6 Months	12.75%
Never Indexed	17%

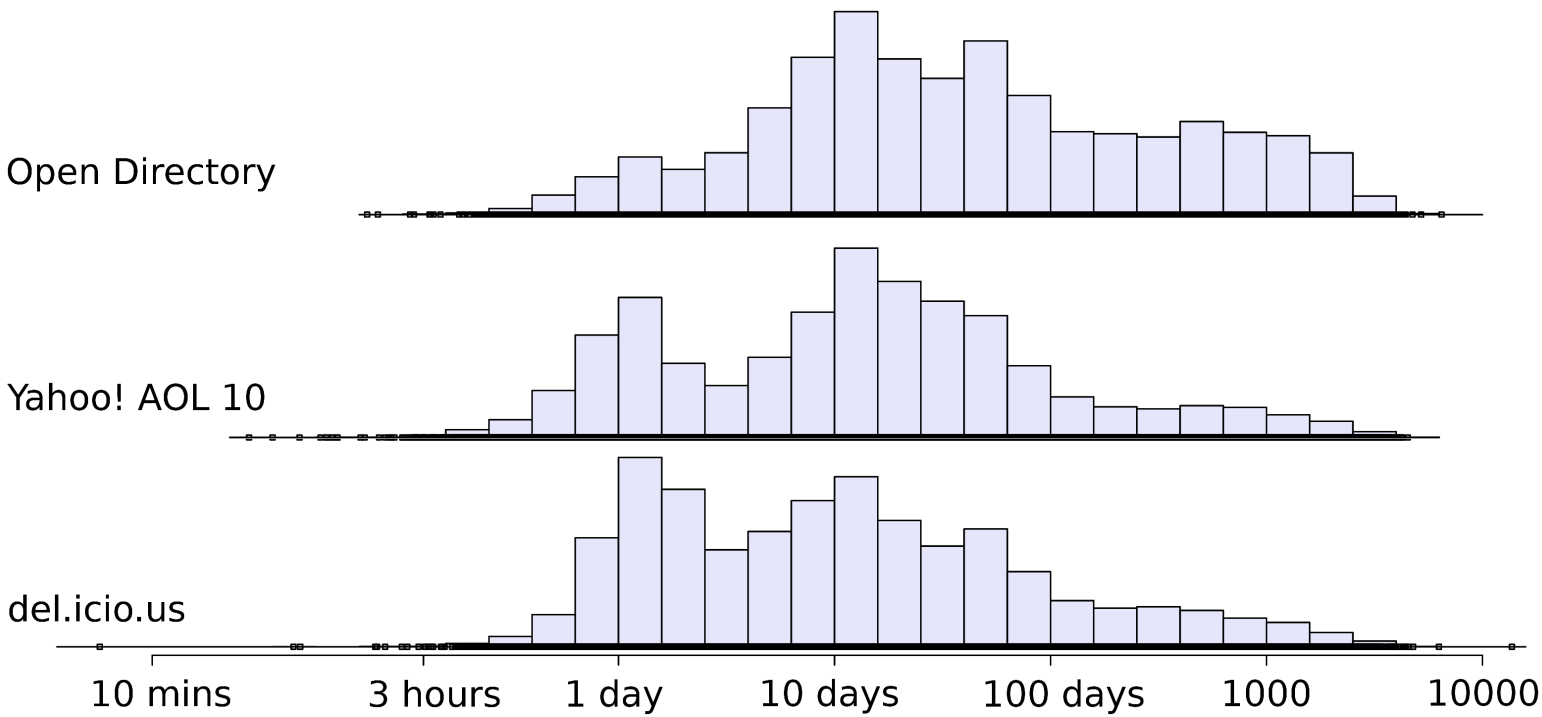
Of the 57.5% found initially, modification time at time of post:



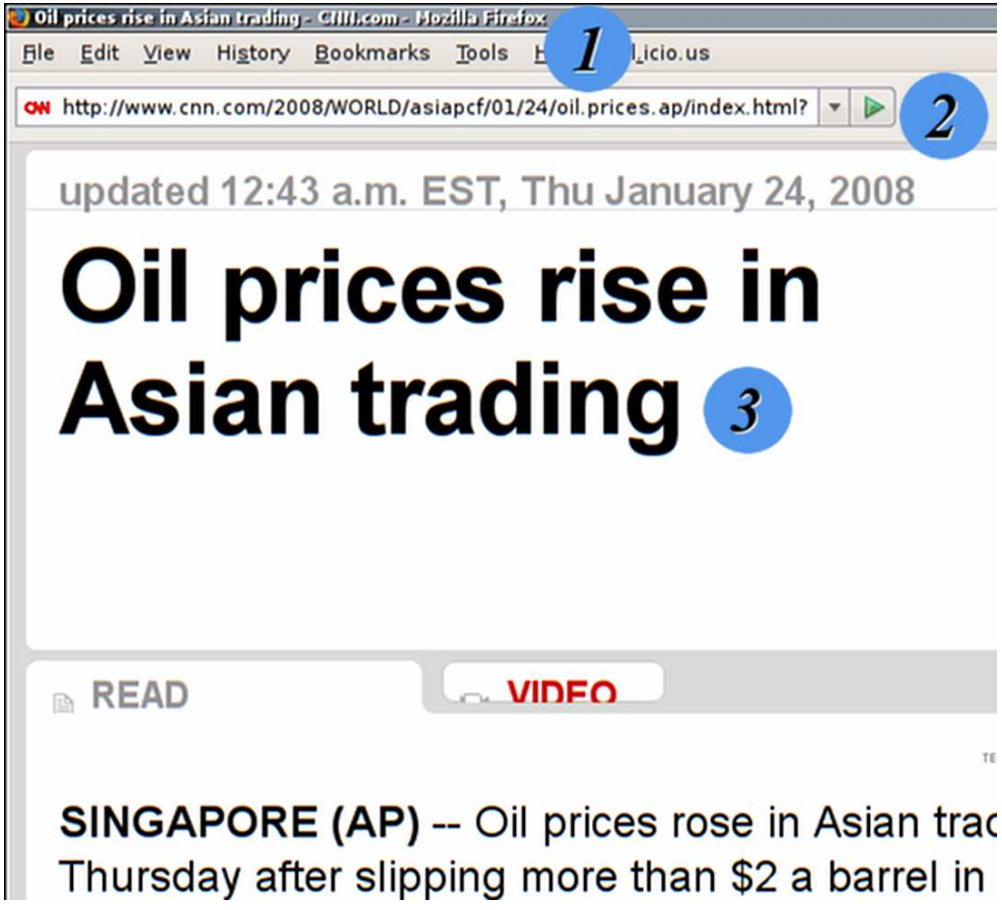
URL Indexing and Age

Found Initially	57.5%
Indexed Within 4 Weeks	12.75%
Indexed Within 6 Months	12.75%
Never Indexed	17%

Of the 57.5% found initially, modification time at time of post:



Tagging Caveats ("The Tagging 6")



1. **Title** (16%)
   Examples: "oil", "prices"
2. **Whole Domain** (20%)
   Examples: "news", "cnn"
3. **Page Text** (50%)
   Example: "singapore"
4. **Extended Text** (80%)
   Example: "inflation"
5. **Irrelevant** (7%)
   Example: "stanford"
6. **Subjective** (<5%)
   Example: "funny"

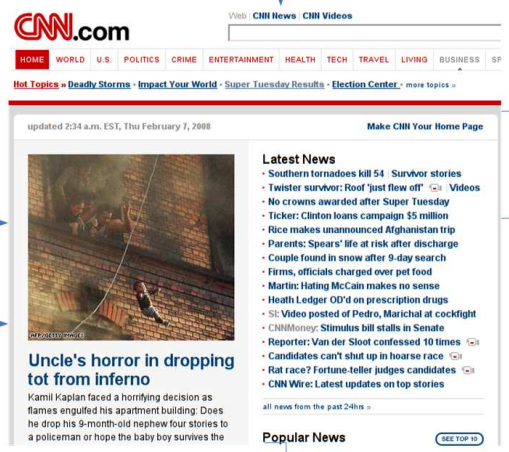
Tags versus Other Content

Back Link Text

… He is also a CNN Contributor, appearing on a variety of shows, including The Situation Room, Anderson Cooper 360, Lou Dobbs Tonight, and many others...

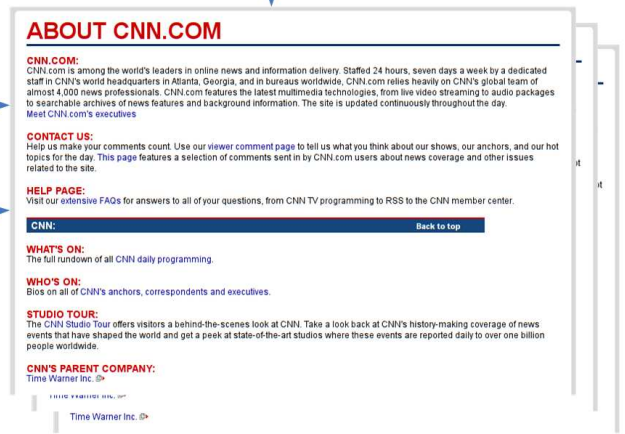
Page Text

CNN.com is among the world's leaders in online news and information delivery. Staffed 24 hours, seven days a week by a dedicated staff in CNN's world headquarters in Atlanta, Georgia, ...



Forward Link Text

CNN.com is among the world's leaders in online news and information delivery. Staffed 24 hours, seven days a week by a dedicated staff in CNN's world headquarters in Atlanta, Georgia, ...



Tags

news cnn daily media

Outline

Conclusions

1. Social bookmarking URLs are new and recent, though many tags may be redundant (given title, text, domains).

Conclusions

1. Social bookmarking URLs are new and recent, though many tags may be redundant (given title, text, domains).
2. Social bookmarking is a large phenomenon, but not nearly as large as the web.

Conclusions

1. Social bookmarking URLs are new and recent, though many tags may be redundant (given title, text, domains).
2. Social bookmarking is a large phenomenon, but not nearly as large as the web.
3. Despite this, relevant URLs are well represented, and popular tags overlap with popular queries.

Conclusions

1. Social bookmarking URLs are new and recent, though many tags may be redundant (given title, text, domains).
2. Social bookmarking is a large phenomenon, but not nearly as large as the web.
3. Despite this, relevant URLs are well represented, and popular tags overlap with popular queries.

Questions?
Check out the full paper at
http://dbpubs.stanford.edu/
or in the proceedings!