Can Social Bookmarking Improve Web Search?

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

What Do Customers Ultimately Buy After Viewing This Item?

82% buy the item featured on this page:
$76.00

7% buy
Pattern Recognition and Machine Learning (Information Science and Statistics) ★★★★★ (48)
$57.36

4% buy
Causality: Models, Reasoning and Inference ★★★★★ (2)
$39.20

3% buy
$71.96

Explore similar items

Tags Customers Associate with This Product
Click on a tag to find related items, discussions, and people.

- machine learning (6)
- graphical models (7)
- bayesian networks (6)
- computer science (7)
- algorithms (1)
- artificial intelligence (1)
- data mining (1)
- statistics (1)

Agree with these tags?
See all 11 tags...

Sell a Digital Version of This Book in the Kindle Store

If you are a publisher or author and hold the digital rights to a book, you can sell a digital version of it in our Kindle Store. Learn more

Rate This Item to Improve Your Recommendations

I own it ★★★★★ Rate this item
Del.icio.us

User is labeling data for us!
Problem Statement

Can social bookmarking improve web search?
Terminology

**Bookmarks/Posts**

- paul: news, uk → bbc.co.uk
  - 08:33:25
- mary: recipes, food → food.com
  - 08:33:23
- dave: tv, cnn, news → cnn.com
  - 08:33:21

**Triples**

- (paul, news, bbc.co.uk)
- (paul, uk, bbc.co.uk)
- (mary, recipes, food.com)
- (mary, food, food.com)
- (dave, tv, cnn.com)
- (dave, cnn, cnn.com)
- (dave, news, cnn.com)
Every 20-40 seconds delicious is polled to see the most recent posts.
Every 20-40 seconds delicious is polled to see the most recent posts.
Every two hours the pages in the queue are crawled.
Every 20-40 seconds delicious is polled to see the most recent posts.
Every two hours the pages in the queue are crawled.
Gathering data from backlinks.
Datasets

• Dataset Crawl: Breath-first search from tag “web”
  – 22,588,354 posts
  – 1,371,941 URLs

• Dataset Recent: The data are gathered from the delicious recent feed.
  – 11,613,913 posts
  – 3,004,998 URLs

• Dataset Month: Recent feed of delicious for May 25th, 2007 + crawl of outlinks
Positive Factors

• **Result 1:** page posted to del.icio.us are often recently modified.

• **Conclusion:** del.icio.us users post interesting pages that are actively updated or have been recently created.
Positive Factors

• **Result 2:** Approximately 25% of URLs posted by users are new, unindexed pages.

• **Conclusion:** del.icio.us can serve as a (small) data source for new web pages and to help crawl ordering.
Positive Factors

• **Result 3**: Roughly 9% of results for search queries are URLs present in del.icio.us. 

**Conclusion**: del.icio.us URLs are disproportionately common in search results compared to their coverage.
Negative Factors

\[ \approx 120 \text{ thousand } (\approx 10^5) \text{ posts/day} \]

(versus \( \approx 10^6 \) blog posts/day)

60–150 million posts

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

(versus \( \approx 10^9–10^{11} \) total URLs)
Tags

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

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
Conclusion

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

• Social bookmarking is a large phenomenon, but not nearly as large as the web.

• Despite this, relevant URLs are well represented, and popular tags overlap with popular queries.