Can Social Bookmarking Improve Web Search?

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

February 12th, 2008
Outline

Introduction

Problem Statement

Data Gathering Methodology

Analysis

Conclusions
Spaghetti Town!

Added November 30, 2006
From TheReceptionist
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 name...>

Related

Showing 1-20 of about 6,910

Re: Spaghetti Town!
01:50
From: Jacko03
Views: 1795

Re: Spaghetti Town!
00:26
From: amycetic
Views: 1521
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


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

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

Introduction

Problem Statement

Data Gathering Methodology

Analysis

Conclusions
Problem Statement

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
daily news media
Outline

Introduction

Problem Statement

Data Gathering Methodology

Analysis

Conclusions
del.icio.us posts

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)
Realtime Web Crawling
Outline

- Introduction
- Problem Statement
- Data Gathering Methodology
- Analysis
- Conclusions
Size and Growth

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

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

60–150 million posts

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

(versus \( \approx 10^9 – 10^{11} \) total URLs)
URL Indexing and Age

<table>
<thead>
<tr>
<th>Found Initially</th>
<th>57.5%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Indexed Within 4 Weeks</td>
<td>12.75%</td>
</tr>
<tr>
<td>Indexed Within 6 Months</td>
<td>12.75%</td>
</tr>
<tr>
<td>Never Indexed</td>
<td>17%</td>
</tr>
</tbody>
</table>

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

del.icio.us
URL Indexing and Age

<table>
<thead>
<tr>
<th>Found Initially</th>
<th>57.5%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Indexed Within 4 Weeks</td>
<td>12.75%</td>
</tr>
<tr>
<td>Indexed Within 6 Months</td>
<td>12.75%</td>
</tr>
<tr>
<td>Never Indexed</td>
<td>17%</td>
</tr>
</tbody>
</table>

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, ...
Outline

Introduction

Problem Statement

Data Gathering Methodology

Analysis

Conclusions
Conclusions

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

Questions?

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

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