Authors vs. Readers
A Comparative Study of Document Metadata and Content in the WWW

by Michael G. Noll

http://www.michael-noll.com/publications/
Overview

- Introduction
- Authors
- Readers
- Authors vs. Readers
- Summary & conclusion
Introduction

- how can we benefit from things like “Web 2.0”? = what do end users bring to the table?
- how much and which kind of user-supplied (meta)data is out there?
- what can we expect to do with it?
- how does it compare to “traditional” metadata?
Introduction
Introduction

Hi, I engineer documents!
Introduction
Introduction
Introduction
Information sources

- del.icio.us
  - bookmarks & tags
  - end users

- document source
  - HTML metadata
  - ICRA rating labels
  - actual document content

- PageRank
  - Google
  - authors & publishers
  - page popularity

- Authors vs. Readers

August 31, 2007
DMOZ100k06

- data set created from these information sources
- based on random sample of 100,000 web documents (2.1%) from the Open Directory
- idea: help researchers and allow comparison of results
- freely available:
  
  http://www.michael-noll.com/dmoz100k06/
DMOZ100k06

- overall statistics

<table>
<thead>
<tr>
<th>Total documents</th>
<th>97,578</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total bookmarks</td>
<td>180,246</td>
</tr>
<tr>
<td>Total (common) tags</td>
<td>25,311</td>
</tr>
<tr>
<td>Bookmarked documents</td>
<td>13,771</td>
</tr>
<tr>
<td>Tagged documents</td>
<td>4,992</td>
</tr>
</tbody>
</table>
DMOZ100k06

- overall statistics

<table>
<thead>
<tr>
<th></th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total documents</td>
<td>97,578</td>
</tr>
<tr>
<td>Total bookmarks</td>
<td>180,246</td>
</tr>
<tr>
<td>Total (common) tags</td>
<td>25,311</td>
</tr>
<tr>
<td></td>
<td>6,090 unique</td>
</tr>
<tr>
<td>Bookmarked documents</td>
<td>13,771</td>
</tr>
<tr>
<td></td>
<td>14.1%</td>
</tr>
<tr>
<td>Tagged documents</td>
<td>4,992</td>
</tr>
<tr>
<td></td>
<td>5.1%</td>
</tr>
</tbody>
</table>

- per document

<table>
<thead>
<tr>
<th></th>
<th>mean</th>
<th>std.dev.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bookmarks</td>
<td>1.85</td>
<td>47.68</td>
</tr>
<tr>
<td>Tags (common)</td>
<td>0.26</td>
<td>1.80</td>
</tr>
<tr>
<td>PageRank</td>
<td>3.13</td>
<td>1.66</td>
</tr>
</tbody>
</table>
DMOZ100k06

- PageRank distribution of docs in the data set
Authors
Authors: available metadata

Results

- document title >> everything else
- keywords > description
- forget about ICRA
- cf. Google study [Dec'05]

Interesting

- drop at PR0, peak at PR8
Readers
Readers: available metadata

Results

- bookmarks $\geq$ tags due to experiment setup
- popularity is king

Interesting

- search engines meet readers' taste, or "follow the mass" effect?
Readers: available metadata

Looking only at bookmarked/tagged documents

- main tagging window between PR5 and PR7
- tagging is shifted towards lower PR compared to bookmarking

Interesting

- no “long tail”
Readers: top tags and top bookmarks

Results

- most metadata concentrated on a small set of docs
- power law graph
- Zipf's law for tags, starting at #100
- distribution for docs (us) similar to findings for users (others)
Authors vs. Readers
Authors vs. Readers

Matching metadata of authors and readers

- authors: title, keywords, description, body
  + <combined>

- readers: tags
Authors vs. Readers

Matching metadata of authors and readers

- authors: title, keywords, description, body + <combined>
- readers: tags

Preprocessing improved performance

- e.g. 46% → 58% matches for <combined>
Authors vs. Readers

Results

- body >> HTML metadata
- parts of body can be sufficient
- relatively stable for body, decreasing for metadata with higher PR
Authors vs. Readers

- the more popular a document, the less does its metadata reflect the perception of users
- user annotations provide additional information not available in a document itself (+ authors)
- keywords (23%) > description (15%), opposite of what search engines prefer
Summary and conclusion

- will not repeat results from previous slides :-)
- tags provide additional information which is not contained in a document itself
  - good: helps information retrieval and classification
  - bad: concentration on a relatively small subset of docs, but techniques such as PEBL can help
- upcoming DMOZ100k06 corpus will have much more information: *all* user annotations + more
DMOZ100k06

"A large research data set about document metadata based on a random sample of 100,000 web documents from the Open Directory combined with data retrieved from search engines, Google, and iCMR."