Complementary or Discrete Contexts in Online Indexing:

A Comparison of User, Creator and Intermediary Keywords

CAIS/ACSI 2006, York University, June 1-3
Social Bookmarking Tools

- share links, etc.
- users encouraged to add keywords

Examples:
- citeulike - journals, etc
  - http://www.citeulike.org/
- del.icio.us - websites, etc.
  - http://del.icio.us/
- connotea - journals, etc
  - http://www.connotea.org/
Everyone's library

Some recent papers posted to CiteULike - all mixed together. You can also see just your library.

- **Bacterial species and speciation.**
  by Cohan FM
  posted to microbial-species speciation by imeppley as ★★★ on 2006-03-23 21:58:24

- **Some Lattice-based Scientific Problems Expressed in Haskell**
  JFP (to appear)
  by Carpenter DB, Glaser H
  posted to functional-programming haskell lattice parallel scientific by dgibert as ★★★ on 2006-03-23 21:51:34

- **Mitochondrial genome variation and the origin of modern humans.**
  by Inman M, Kaessmann H, Pääbo S, Gyllensten U
  posted to complete divergence mt origins phylogeny by Archaeogenetics as ★★★ on 2006-03-23 21:51:33

- **Genetic variation: molecular mechanisms and impact on microbial evolution.**
  by Aber W
  posted to evolution horizontal-gene-transfer by imeppley as ★★★ on 2006-03-23 21:48:03

- Affinities, Benefits, and Costs: The ABCS of Good Scholars Gone Public

Advertisements
3 distinct groups involved in indexing
- author
- user
- intermediary (librarian, etc)

Each group has a different relation to the documents being indexed.

Can these differences be seen in the use of index terms by the three groups?
Research Questions

• Are user categories different from subject headings (descriptors) or author keywords?
• Are there differences in context, type or some other semantic relationship?
Methodology

- data collection from citeulike
  - journal articles from LIS related journals
  - journals indexed by INSPEC or Library Lit
  - 165 articles retrieved
- journal articles potentially have all three types of index terms
  - user tags from citeulike
  - author keywords from article
  - descriptors from online database (e.g. Library Lit, INSPEC)
Methodology (cont.)

- compare all terms
  - similarities/differences
  - non traditional categories
- descriptive statistics
  - usage patterns
  - term matches (synonymity, etc)
- term comparison
  - Voorbij (1998)
    - compared index terms and title words
Voorbij's Thesaural Relations

- 1. SAME
- 2. SYNONYM (Used For)
- 3. BT (Broader Term)
- 4. NT (Narrower Term)
- 5. RT (Related Term)
- 6. Related (Related but not in the thesaurus)
- 7. Not related (unrelated to other terms, may still be a good fit for the article)
General Results

- Spelling variations
  - UK vs US English spelling
- Punctuation
  - information_seeking vs information-seeking
- Sesquipedalians
  - informationseekingbehaviour
- Acronyms and abbreviations
  - ir for information retrieval
- Language
  - mostly English
Statistics

- 529 tags, 775 author keywords, 727 intermediary descriptors (from 165 articles)
- 92% of articles had 1-3 authors (max. 9)
- 83% of articles had 1-2 taggers (max. 13)
- 60% of articles had 1-3 tags, 4-6 author keywords, 3-5 intermediary descriptors
- positive correlation between number of taggers and number of tags
<table>
<thead>
<tr>
<th>Type</th>
<th># of Articles</th>
<th>Percent</th>
</tr>
</thead>
<tbody>
<tr>
<td>Same</td>
<td>103</td>
<td>62.4</td>
</tr>
<tr>
<td>Synonym</td>
<td>47</td>
<td>28.5</td>
</tr>
<tr>
<td>NT/BT</td>
<td>58</td>
<td>35.2</td>
</tr>
<tr>
<td>RT</td>
<td>79</td>
<td>47.9</td>
</tr>
<tr>
<td>Related</td>
<td>133</td>
<td>80.6</td>
</tr>
<tr>
<td>Not Related</td>
<td>157</td>
<td>95</td>
</tr>
</tbody>
</table>
Term Comparisons

- Same:
  - user: copyright, openaccess, romeo
  - author: universities, publishing, academic staff, copyright, ownership
  - intermediary: copy protection, copyright, electronic publishing, meta data
Narrower Term/Broader Term

- user: information, information-systems, philosophy, systems
- author: information systems, philosophy
- intermediary: management information systems, organisational aspects, philosophical aspects, statistical analysis
Related Term

• user: citation, internet, no-tag, research, web, writing
• author: scientific and technical information, primary literature, scholarly publishing, hypermedia links, link analysis, bibliographic citations
• intermediary: citation analysis, electronic publishing, information resources, Internet
Related Tags

- related but not strictly RT
- user: information-seeking-behaviour, information寻求行为
- author: libraries, information research, information retrieval, behaviour
- intermediary: libraries, information research, information retrieval, behaviour, task analysis
Related Tags (cont.)

- **user**: relevance
- **author**: user criteria, relevance judgments, information seeking
- **intermediary**: cognitive systems, educational computing, geophysics computing, human factors, information retrieval systems, meteorology, multimedia systems, relevance feedback, user interfaces
Unrelated Tags

- average of 3.5 per article, not all from tags

- Categories:
  - Time and Task Management
  - Geographic Tags
  - Specifics
  - Generalities
  - Emergent Vocabulary
  - Non-English tags
  - Other
Time and Task Management

- Examples:
  - todo
  - new
  - print, toprint
  - maybe
  - lis510
- Not related to the article—entirely between the user and the document
- project or task based
- toread is encoded in stars in article post
Geographic Tags

• Examples:
  - regions: Asia
  - countries: Turkey, Great Britain, Israel
  - cities: Sydney, Berkeley

• Country and region names

• only 4 examples in user or author terms, all others from descriptors
Specifcs

- **Names**
  - Bates, Aczel

- **Specific Systems**
  - Romeo, Pubmed, Citeseer

- **Specific Groups Studied**
  - medical students, library science students, college students

- **Methodologies**
  - survey, interview, pubmed-mining
Generalities

- Extremely general terms that could apply to almost any article in a field
- e.g.: computers, information, library/libraries
- As expected since tagging system lacks a thesaurus
Other Categories

- Emergent Vocabulary
  - folksonomy, tagging
- Non-English tags
  - etsint_prosesssit
  - Relevansvurdering
- no-tag
  - system created default tag assigned when the user assigns no tags
Conclusions

- There are differences between user, author and intermediary conceptions of the documents
  - Thesaural relations are rare in tags except Same and RT
  - Time and task management, geographic information, emergent vocabulary
  - Users and authors do not standardise their vocabulary
Acknowledgments

• Social Sciences and Humanities Research Council
• Faculty of Information and Media Studies, University of Western Ontario
Complementary or Discrete Contexts in Online Indexing:

A Comparison of User, Creator and Intermediary Keywords

CAIS/ACSI 2006, York University, June 1-3