Theme Creation for Digital Collections

Xia Lin, Jiexun Li, Xiaohua Zhou
Drexel University
Philadelphia, USA
Challenges in Concept Representation

- Representations of associative relationships
  - Concept relationships are not always in hierarchies.
  - There are many types of associative relationships
- Dynamic concept representation
Dynamic Concept Representation

- Concepts are only meaningful when
  - specific domains or document collections are identified.
  - specific context of use is understood
  - specific groups of users are identified
- The same concept might be looked at from different perspectives when the user’s information needs are different.
- How do we catch all of these in concept representation?
Looking for solutions …

Subject Headings → Concepts

Documents → Terms

Topics? → Themes?

Users → Information Needs
Getting Users Involved

- Approach 1
  - Semantic Mapping + Visualization
    - Computers extract associations from documents and terms
    - Users understand them through visual displays and visual interactions
  - My ISKO2006 Paper
    - “Dynamic Concept representation through a Visual Concept Explorer”
    - Example: the “Brain” maps
Getting Users Involved

- Approach 2
  - Semantic Mapping + User-initiated Theme creation
    - Collection-specific semantic mapping
    - Clustering of user search terms (from the search log)
    - User’s creation of themes and topics
    - User-computer collaboration of theme editing/updating/dynamic changing
Semantic Profiles

What indexing terms are mostly likely associated with a given free-text term?
- The mapping of indexing terms to free terms based on the statistic associations within a collection
- The mapping reflects the nature of the collection.

<table>
<thead>
<tr>
<th>Free-text Term: Health</th>
<th>IPL</th>
<th>IPL-Teens</th>
</tr>
</thead>
<tbody>
<tr>
<td>health</td>
<td>health</td>
<td></td>
</tr>
<tr>
<td>Mental health</td>
<td>fitness</td>
<td></td>
</tr>
<tr>
<td>health care</td>
<td>exercise</td>
<td></td>
</tr>
<tr>
<td>disease</td>
<td>nutrition</td>
<td></td>
</tr>
<tr>
<td>activism</td>
<td>stress</td>
<td></td>
</tr>
<tr>
<td>disorders</td>
<td>tic</td>
<td></td>
</tr>
<tr>
<td>public Health</td>
<td>panic</td>
<td></td>
</tr>
<tr>
<td>psychology</td>
<td>medicine</td>
<td></td>
</tr>
<tr>
<td>mental</td>
<td>attention</td>
<td></td>
</tr>
<tr>
<td>Illness</td>
<td>deficit</td>
<td></td>
</tr>
<tr>
<td>reproduction</td>
<td>ADHD</td>
<td></td>
</tr>
<tr>
<td>medicine</td>
<td>depression</td>
<td></td>
</tr>
</tbody>
</table>
What terms are mostly likely associated with a given indexing term?

<table>
<thead>
<tr>
<th>Indexing Term: Health</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>IPL</strong></td>
</tr>
<tr>
<td>Stress</td>
</tr>
<tr>
<td>depress</td>
</tr>
<tr>
<td>teen</td>
</tr>
<tr>
<td>anxiety</td>
</tr>
<tr>
<td>health</td>
</tr>
<tr>
<td>disorder</td>
</tr>
<tr>
<td>eat</td>
</tr>
<tr>
<td>Mental</td>
</tr>
<tr>
<td>......</td>
</tr>
</tbody>
</table>
DEMO:

- http://project.cis.drexel.edu/IPL/
# Semantic Mapping of IPL Subjects

Keywords are subject terms indexers assigned to each item for subject representations. Please look up similar keywords before creating a new one.

**Mapping:**

<table>
<thead>
<tr>
<th>Subject Terms</th>
<th>Related Subject Terms</th>
</tr>
</thead>
<tbody>
<tr>
<td>Organizing the Web</td>
<td></td>
</tr>
<tr>
<td>Standards</td>
<td></td>
</tr>
<tr>
<td>Online Catalogs</td>
<td></td>
</tr>
<tr>
<td>Canada</td>
<td></td>
</tr>
<tr>
<td>Cataloging</td>
<td></td>
</tr>
<tr>
<td>HTML</td>
<td></td>
</tr>
</tbody>
</table>

**Keywords**

<table>
<thead>
<tr>
<th>Keywords</th>
<th>Related Keywords</th>
</tr>
</thead>
<tbody>
<tr>
<td>help</td>
<td></td>
</tr>
<tr>
<td>Icons</td>
<td></td>
</tr>
<tr>
<td>testing</td>
<td></td>
</tr>
<tr>
<td>Tests</td>
<td></td>
</tr>
<tr>
<td>Cataloging</td>
<td></td>
</tr>
<tr>
<td>World Wide Web</td>
<td></td>
</tr>
<tr>
<td>Internet</td>
<td></td>
</tr>
</tbody>
</table>

Click a term on the left-hand side to retrieve related keywords on the right.
Semantic Mapping of IPL Subjects

Keywords are subject terms indexes assigned to each item for subject representations. Please look up similar keywords before creating a new one.

Mapping: intellectual

Subject Terms
- Intellectual Property
- Patents
- Copyright
- Non-U.S. & International
- Censorship
- Issues
- Law & Law Enforcement Blogs
- Biography
- Other Sports
- Non-fiction

Related Subject Terms
- Copyright
- Intellectual Property
- Patents
- State Parties
- Law & Law Enforcement Blogs
- United States History
- United States
- Political Parties (United States)
- State Digital Libraries in the U.S.
- United States Travel & Tourism

Keywords
- intellectual property
- copyright
- Patents
- Witchcraft
- Poets
- Censorship
- Inventors
- Wicca
- Law
- German

Related Keywords
- copyright
- intellectual property
- Patents
- United States
- Turabian
- Modern Language Association
- MLA
- Chicago
- Supreme Court
- Legal

Click a term on the left-hand side to retrieve related keywords on the right.
Theme creation

- An attempt to bring together three types of semantics to create dynamic representation of concepts and topics
  - semantic analysis of collections.
  - User’s search patterns and search vocabulary
  - Controlled vocabularies or indexing terms
Current Research - 1

- Develop a method of concept representation that are
  - Dynamic
    - Over collections, users, time, etc.
  - Context-sensitive
    - Same themes may be defined different for different collections and different usages.
  - Semi-automatic
    - Integrates both manual and automatic processes.
  - Collaborative
    - Allows users/indexers to make changes.
Wiki interface for theme creation

American Literature

Themes
The collection page for all themes of IPL
24 articles in English

Literature
Themes under this subject

All Subjects > Literature >> American Literature

American Literature
Description:
The study of American literature and American authors.
Show more themes

Topic 1: Author Names

Keywords: American Poets, Minority American Authors, American Authors, American Playwrights, Female American Authors

Related Resource

IPL URL: Voices from the Gap, Today in Literature, The Whitman Archives, Literary Kicks, Mark Twain

Other URL: American Literature Anthology, American Authors, PAL: Perspectives in American Literature, American Authors on the Web
Test the theme creation process through IPL

- A group of students work together to create themes for IPL
  - Analyze three months of search logs together
  - Each describes a particular “theme” he or she “reads” from the search log.
  - Use the semantic mapping tool to create and enhance the representation.
Current research - 3

- Develop an IPL metadata creation tool with semantic supports.
  - Utilize the semantic mapping and theme creation process
    - Suggest associative lists for keywords and subject terms
  - Integrate multiple subject-representation objects
  - Support collaboration
Future Research

- Dynamic concept representation
  - It needs more research!
- Use of Coding schemes
  - SKOS, Topic Maps, and others
- Evaluation
  - “Aboutness” is very difficult to evaluate.
  - “Usefulness” might be the criteria!