The Current State of Automated Content Tagging: Dangers and Opportunities

DCMI Webinar

Joseph Busch
Agenda

- Automated tagging
- Pre-defined Boolean queries
- Case study
Taxonomy
Strategies
The business of organized information
Semantic STAFFING Experts placing experts
Generate complete and consistent metadata for all content in all systems, to be able to do analytics and identify emergent patterns.
How to generate complete and consistent metadata

Indexer Inconsistency: 70%

Automated Tools Consistency: 80%
Best case scenario: Automated suggestion + SMEs approve or improve
Cloud Computing
Automated tagging methods used by tools and applications

- Entity extraction
- Sentiment analysis
- Keyword extraction
- Summarization
- Predefined Boolean queries
- Trained categorizers
- Statistical categorizers
Agenda

- Automated tagging
- Pre-defined Boolean queries
- Case study
Boolean search: How hard is it do?

**Machine learning**
- Machine learning is opaque.
  - Benefit: No preparation. Content just gets processed.
  - Problem: Categories are generic, may be irrelevant, can be biased, and are difficult to change or tune.

**Boolean queries**
- Pre-defined categories (aka a taxonomy) plus Boolean queries to scope the context for categories are transparent.
  - Benefit: Relevant categories.
  - Problem: Requires work to set up, and specialized skills.
Boolean queries

- Basic operators
  - AND (conjunctive)
  - OR (disjunctive)
  - NOT (negation)

- Venn diagrams
Proximity operators

- Proximity search (specified distance). **Hint**: Proximity operators and syntax are not standardized.
  - NEAR
  - NOT NEAR
  - FOLLOWED BY
  - NOT FOLLOWED BY
  - SENTENCE
  - FAR
Query syntax

- Bounded phrase
  - Usually quotation marks, e.g.
    “health insurance”

- Truncation (right, left, internal)
  - Usually an asterisk, e.g.
    child*
    “pre-existing condition*”

- Nested statements
  - Parentheses (that must match up)
    (“health insurance” AND (children* OR “pre-existing condition*”))
How to create a Boolean query (1)

1) Brainstorm a list of 10 relevant words and phrases.

2) Use that list to identify 10 relevant items (articles, videos, websites, etc.)
   ▪ E.g., do a Google search, search Google Scholar, search the NYT (or any other newspaper that you subscribe to), search Library of Congress Chronicling America (1789-1963), etc.

3) Review 10 relevant items and write down the words and phrases that provide a context for the theme/topic/concept.
   ▪ Titles, headings, summaries, introductions (at the beginning) and conclusions (at the end) are good areas to focus on without having to read the whole thing.

4) Note any named entities (people, organizations, events, laws, etc.) that are closely associated with the theme/topic/concept.
   ▪ E.g., for gun violence Gabrielle Giffords, Michael Bloomberg, Doctors Against Gun Violence, March for our Lives, etc.
How to create a Boolean query (2)

5) Consolidate the terms.
   - Identify duplicates, synonyms, as well as any concepts that you want to combine even if they are not synonyms.
   - Re-label the term as needed to reflect the concept/category. Also consider and note any other relationships between terms. Prioritize the terms. Rank from 1-N, most relevant to least relevant.
   - **Hint**: Rank each term by higher, medium, lower relevance, then sort and rank from 1-N.

6) Write a query for each term.
   - Note that regular plurals (-s, -es, -ies) are usually (but not always) included automatically, but you always need to specify irregular plurals, e.g., “mice”.

7) Qualify the scope for each term.
   - Does the term require any qualification of the scope, e.g., by population, setting, geography, etc.?
   - Validate that the term is disjunctive, distinct, and requires no further qualification.

8) Combine the terms into a single nested query with an OR operator.
Agenda

- Automated tagging
- Pre-defined Boolean queries
- Case study
Case study

Childhood Obesity  Disease Prevention and Health Promotion  Health Care Quality  Health Coverage
## Testing process: Text collections

<table>
<thead>
<tr>
<th>User interfaces</th>
<th>API</th>
<th>Excel</th>
</tr>
</thead>
<tbody>
<tr>
<td>Test collection</td>
<td>90 Repository long form assets</td>
<td>400 WCMS short form assets</td>
</tr>
<tr>
<td>Content</td>
<td>Full text</td>
<td>Title &amp; summary only</td>
</tr>
<tr>
<td>Format</td>
<td>Clear text</td>
<td>Clear text, CSV</td>
</tr>
<tr>
<td>Topics</td>
<td>Childhood Obesity, Disease Prevention and Health Promotion, Health Care Quality, Health Coverage</td>
<td>Childhood Obesity, Disease Prevention and Health Promotion, Health Care Quality, Health Coverage</td>
</tr>
</tbody>
</table>
Sample pre-defined query: Health coverage

("health insurance coverage" OR "health coverage") OR ("healthcare reform" OR "health care reform") OR ("Better Care Reconciliation Act" OR BCRA) OR ("American Health Care Act" OR AHCA) OR ("Affordable Care Act" OR "ACA" OR Obamacare) OR ((Medicare OR Medicaid) NEAR/5 (spend* OR cover* OR expan*)) OR ("health insurance exchange" OR "HIE") OR ("health insurance" NEAR/5 marketplace*) OR ("federal* facilitated marketplace*" NEAR/10 "health insurance") OR ("federal* run marketplace*" NEAR/10 "health insurance") OR ((state NEAR/5 marketplace*) NEAR/10 "health insurance") OR ("small business marketplace*" NEAR/10 "health insurance") OR ("small-business marketplace*" NEAR/10 "health insurance") OR ("small business" NEAR/5 exchange*) NEAR/10 "health insurance") OR ("high-risk" OR "high risk") NEAR/10 "health insurance") OR (uninsured NEAR/5 (veteran* OR child* OR adult* OR people OR kid* OR citizen*)) OR ("pre-existing condition*" OR "preexisting condition*") NEAR/10 "health insurance") OR "health insurance rate*" OR (cost* OR rate* OR payment*) NEAR/10 "health insurance") OR ("health insurance" NEAR/10 "tax credit") OR (healthcare OR "health care") NEAR/5 spending) OR ((healthcare OR "health care") NEAR/5 utilization) OR ("high-deductible" OR "high deductible") NEAR/10 "health insurance") OR ("mental health" OR "substance abuse") NEAR/10 "health insurance") OR ("provider network*" NEAR/10 "health insurance") OR ("in-network" OR "out-of-network") NEAR/10 "health insurance") OR (PPO* OR HMO*) NEAR/5 (marketplace* OR plan* OR provider*)) OR ("health insurance" NEAR/10 (enroll* OR "re-enroll*" OR renew* OR "open-enrollment" OR "open enrollment") OR ((navigator* OR assistor* OR assister*) NEAR/10 ((("health insurance" OR Medicare OR Medicaid) NEAR/5 enroll*)) OR ("CHIP" OR "Children’s Health Insurance Program") OR ("individual mandate" NEAR/10 "health insurance") OR "employer-sponsored insurance" OR ((employer OR employee) NEAR/10 "health insurance"))
Overall trial results

Categorized to a Topic (Recall)

- Categorized: 67%
- Uncategorized: 33%

n=400

Categorized to the correct Topic (Precision)

- Correct: 89%
- Incorrect: 11%

n=270
Trial results for each Topic

<table>
<thead>
<tr>
<th>Topic</th>
<th>Incorrect</th>
<th>Correct</th>
<th>n</th>
</tr>
</thead>
<tbody>
<tr>
<td>Childhood Obesity</td>
<td>17%</td>
<td>83%</td>
<td>64</td>
</tr>
<tr>
<td>Disease Prevention and Health Promotion</td>
<td>20%</td>
<td>80%</td>
<td>65</td>
</tr>
<tr>
<td>Health Care Quality</td>
<td>4%</td>
<td>96%</td>
<td>69</td>
</tr>
<tr>
<td>Health Coverage</td>
<td>3%</td>
<td>97%</td>
<td>72</td>
</tr>
</tbody>
</table>
Automated tagging summary

- **Value proposition –**
  - Tag content consistently so it can be aggregated, analyzed and used by organizations.

- **Enabling technologies –**
  - Cloud services
  - Cognitive computing
  - Internet of Things

- **Tools and applications –**
  - Exist and are affordable.
Challenge

- Good implementation skills
  - Are hard to find
  - Training and expertise is needed
More information

Questions

Joseph Busch,
jbusch@taxonomystrategies.com
joseph@semanticstaffing.com
(m) 415-377-7912