More Like This: Approaches to Recommending Similar Items Using Subject Headings

Kevin Beswick
NCSU Libraries Fellow
code4lib 2014 conference, Raleigh, NC
Agenda

• What?
• Why?
• How?
• Evaluation. Are these approaches any good?
• Where are we going from here?
Recommendation Systems

• A system that presents a set of related items that would interest a particular user

• Collaborative filtering - look at user behavior
  • eg. full record page view data, circulation data, etc.

• Content-based filtering - look at properties of content itself
  • eg. call numbers, subject headings, etc.
Motivations

• Many popular web services offer this functionality
  • eg. Facebook, Netflix, Amazon, etc.
  • Users coming to expect it
  • Encourages use & makes it easier to use our service
  • Also…
bookBot

- Most of Hunt’s collection is stored in an ASRS
  - No physical browsing
  - Need to explore methods for serendipitous discovery
A Brief History of Browse @ NC State

- Virtual Browse team with members from many library departments

- Previous Projects:
  - “Browse Shelf” feature in library catalog
  - Virtual Browse kiosk @ Hunt Library
Advantages of Subject Heading Based Recommendation

- Vs. Call Number Browse
  - Can recommend more than items that are shelved next to each other
  - A lot of our e-books don’t have call numbers

- Vs. Collaborative Filtering
  - Hard to collect reliable circulation data for electronic resources
Four Algorithms/Approaches
Most Subject Headings

North Carolina. Department of Agriculture -- Bibliography -- Catalogs

Agriculture -- North Carolina -- Bibliography -- Catalogs

Administrative law -- North Carolina
First Subject Headings

Highest Weighted

1. North Carolina. Department of Agriculture -- Bibliography -- Catalogs
2. Agriculture -- North Carolina -- Bibliography -- Catalogs
3. Administrative law -- North Carolina

Lowest Weighted
Most Subject Terms

North Carolina. Department of Agriculture -- Bibliography -- Catalogs

Agriculture -- North Carolina -- Bibliography -- Catalogs

Administrative law -- North Carolina
Weighted Subject Terms

- North Carolina. Department of Agriculture
- Agriculture
- Administrative law
- North Carolina
- Bibliography
- Catalogs

- topical term
- geographic

+ general subdivision
+ chronological subdivision
Implementation

• Quick & simple implementation
  • Python / Flask - handle requests, provide testing interface
  • Solr / SolrMARC - handle the actual work
Python / Flask App

- Handles requests / responses
  - Accepts a bibliographic ID & algorithm type as input
  - Sends a different query to Solr depending on algorithm
    - Uses SolrPy library
  - Returns a list of recommendations in JSON
  - Also an HTML testing & evaluation interface
Solr / SolrMARC

- Indexed fields with SolrMARC:
  - Entire subject headings
  - Each subject heading term
  - Each topical, general, geographical, chronological, form subdivision
- Lean on Solr to do the heavy lifting in terms of returning the most related items
FindSimilar Results  Most Headings Algorithm

Results similar to: "Perception beyond Gestalt: progress in vision research"

<table>
<thead>
<tr>
<th>Subject Headings</th>
<th>Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gestalt psychology</td>
<td></td>
</tr>
<tr>
<td>Perception</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Title</th>
<th>Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>The psychology of perception; a philosophical examination of Gestalt</td>
<td>9.400139</td>
</tr>
<tr>
<td>theory and derivative theories of perception</td>
<td></td>
</tr>
<tr>
<td>The world in your head: a gestalt view of the mechanism of conscious</td>
<td>7.5581114</td>
</tr>
<tr>
<td>experience / Steven Lehar</td>
<td></td>
</tr>
<tr>
<td>The Legacy of Solomon Asch: essays in cognition and social psychology</td>
<td>7.5601114</td>
</tr>
<tr>
<td>/ edited by Irvin Rock</td>
<td></td>
</tr>
<tr>
<td>Indirect perception / edited by Irvin Rock; with a foreword by</td>
<td>7.5601114</td>
</tr>
<tr>
<td>Stephen E. Palmer</td>
<td></td>
</tr>
<tr>
<td>Reading and the psychology of perception</td>
<td>7.5601114</td>
</tr>
<tr>
<td>On perceived motion and figural organization [electronic resource]</td>
<td>7.5601114</td>
</tr>
<tr>
<td>/ Max Wertheimer; edited by Lothar Spillmann; with essays by Viktor</td>
<td></td>
</tr>
<tr>
<td>Sarris, Robert Sekuler, and Lothar Spillman; with contributions by</td>
<td></td>
</tr>
<tr>
<td>Michael Wertheimer ... [et al.]</td>
<td></td>
</tr>
</tbody>
</table>
How Well Do These Algorithms Perform?
Preliminary Observation

• Most Headings & Most Terms algorithms looked to be producing decent recommendations a lot of the time

• First Headings algorithm - too few results in a lot of cases

• Weighted Terms algorithm
  • Weighting differs based on subject or user’s interests
  • We don’t want user input
Testing the Algorithms

• Manually test 50 titles on Most Headings & Most Terms algorithms
  • Is either reliable enough & worth implementing?
• 30 hand picked titles
  • representing different subject areas, item formats, lengths & amounts of subject headings
• 20 random titles
Testing the Algorithms

- Blind testing - algorithm unknown
- 10 recommended titles per item
- Rank result set out of 10, 1 point for each relevant work
- Qualitative comments for each result set
Results - Distribution of Scores

**Most Subject Terms Results**

<table>
<thead>
<tr>
<th>Score</th>
<th>Occurrences</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>5</td>
</tr>
<tr>
<td>1</td>
<td>5</td>
</tr>
<tr>
<td>2</td>
<td>4</td>
</tr>
<tr>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>4</td>
<td>2</td>
</tr>
<tr>
<td>5</td>
<td>2</td>
</tr>
<tr>
<td>6</td>
<td>2</td>
</tr>
<tr>
<td>7</td>
<td>1</td>
</tr>
<tr>
<td>8</td>
<td>3</td>
</tr>
<tr>
<td>9</td>
<td>1</td>
</tr>
<tr>
<td>10</td>
<td>16</td>
</tr>
</tbody>
</table>

**Most Headings Algorithm**

<table>
<thead>
<tr>
<th>Score</th>
<th>Occurrences</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>10</td>
</tr>
<tr>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>3</td>
<td>1</td>
</tr>
<tr>
<td>4</td>
<td>1</td>
</tr>
<tr>
<td>5</td>
<td>2</td>
</tr>
<tr>
<td>6</td>
<td>1</td>
</tr>
<tr>
<td>7</td>
<td>1</td>
</tr>
<tr>
<td>8</td>
<td>2</td>
</tr>
<tr>
<td>9</td>
<td>1</td>
</tr>
<tr>
<td>10</td>
<td>15</td>
</tr>
</tbody>
</table>
Results

• Most headings algorithm performs slightly better for shorter (less subdivisions) & fewer subject headings

• Terms algorithm performs significantly better for longer (more subdivisions) & higher numbers of subject headings

• Found that Gov. Docs & Fiction have interesting thematic recommendations that we can’t achieve with shelf browse
Observations

• Duplicate titles
  • Older vs. newer editions
  • Print vs. Electronic

• Format
  • Incorporate a higher weighting on format of recommended items
Observations

- Poorly assigned subject headings responsible for a lot of the poor recommendations
  - General vs. Specific recommendations
  - Automate review/assignment of subject headings to our collection?
Interface Considerations

• Inline “cover-flow” style presentation on full record page
  • Catches eye of user

• Title - “Similar Titles” or “Related Items” etc.

• 5 or so recommendations per title
Takeaways

• Overall, the algorithms perform decently for our collection, but could still be improved in a number of ways

• Your mileage may vary - all collections are different

  • Very dependent on quality & coverage of subject headings
Steps Forward

• Use either Most Terms algorithm by itself, or a hybrid of Most Terms & Most Headings

• Still under active development

  • Explore & implement fixes for issues discussed earlier to improve performance
Thank you!

• Kevin Beswick, NCSU Libraries Fellow (IT & Digital Library Initiatives)

• kdbeswic@ncsu.edu

• @kbeswick on Twitter