Evaluating utility of subject headings in a data repository: 
A preliminary finding from a data search log and record classification

Presented by: 
Mingfang Wu, Australian Research Data Commons
mingfang.wu@ardc.edu.au

Contributors: 
Rowan Brownlee, Australian Research Data Commons
Ying-Hsang Liu, University of Southern Denmark
Jenny Xiuzhen Zhang, RMIT University, Australia

NKOS, 10 Sept. 2020
Outlines

- A background about the studied data catalogue: Research Data Australia
- Log analysis: the usage of subject headings
- Experiments on data record classification
- Future work
Research Data Australia - A National Data Catalogue

144K+ metadata records of dataset
60K+ research grants
99 Contributors

Types of subject vocabularies

Anzsrc-for: The Australian and New Zealand Standard Research Classification (ANZSRC, fields of research)

Global change master directory (GCMD) keywords

Australian Pictorial Thesaurus (apt)

Thesaurus of Psychological Index Terms (psychit)

Library of Congress Subject Headings (lcsh)
ANZSRC-for: The Australian and New Zealand Standard Research Classification - Fields of Research

- ANZSRC ensures that R&D statistics collected are useful to governments, educational institutions, international organisations, scientific, professional or business organisations, business enterprises, community groups and private individuals in Australia and New Zealand.

- ANZSRC-FoR include major fields and related sub-fields of research and emerging areas of study investigated by businesses, universities, tertiary institutions, national research institutions and other organisations.
Anzsrc-for: The Australian and New Zealand Standard Research Classification - Fields of Research

- 22 two digits
- 157 four digits
- 1238 six digits
- 1417 terms in three layers
Number of records per anzsrc-for two digits

04: Earth Sciences  
06: Biological Sciences  
21: History and Archaeology
All text strings (including subject headings) are indexed.
1. Advanced search
2. Facet filter
3. Facet search (vocabulary + keyword)
Log analysis: the usage of subject headings

- Transaction log: one year (2019) of activities recorded from the RDA catalogue
- About 2 million entries/activities, 63% from Australia
- About 496,739 sessions (with 30 minutes duration from the same IP address)
- 37,056 sessions have at least a search event (keyword search, advanced search, subject (facet) filter, subject search)
- 4668 (12.6%) of search sessions involved filters/search with the anzsrc-for subjects, only 45 (0.1%) with gcmd subject
Subject usages per anzsrc-for two digits code
Subject distribution among clicks and the collection
Log analysis: the usage of subject headings

- There is less bias in user’s behaviour of applying subject headings, compared to the content bias toward a few subject headings.

- However, this log shows low usage of subject headings

- Exploring causes
  - Further log analysis, e.g. correlation between subject usage and
    - query types
    - domain knowledge
    - search quality
  - Interface design
  - At the record level: only half of the indexed records have anzsrc-for codes
Machine learning for record classification

- Assign anzsrc-for code to unlabelled records automatically
  - Aim to improve search experience for both human and machine
  - Understand domain coverage of the collection

- Train models, three components are essential for the training:
  - Labels - anzsrc-for code
  - Classifier - four supervised machine learning methods:
    - multinomial logistic regression (MLR), multinomial naive bayes (MNB), K Nearest Neighbors (KNN), Support Vector Machine (SVM)
  - Data - (~78k) records with anzsrc-for code
    - Split into two sets: training set, test set

- Apply model(s)/best prediction to unlabelled records
Record classification with anzsrc-for code

- Use 77918 records that have an anzsrc-for code for training models
- Step by step: first the top two digits, then move down to four, six digits
- Four models: multinomial logistic regression (MLR), multinomial naive bayes (MNB), K Nearest Neighbors (KNN), Support Vector Machine (SVM)

<table>
<thead>
<tr>
<th>Model</th>
<th>Training Set Accuracy</th>
<th>Test Set Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Logistic Regression</td>
<td>0.769149</td>
<td>0.701299</td>
</tr>
<tr>
<td>SVM</td>
<td>0.696435</td>
<td>0.676324</td>
</tr>
<tr>
<td>Multinomial Naïve Bayes</td>
<td>0.702965</td>
<td>0.659341</td>
</tr>
<tr>
<td>KNN</td>
<td>0.906460</td>
<td>0.642358</td>
</tr>
</tbody>
</table>

Acknowledgement:
Adapted the code from Miguel Frenandez Zafra
Performance per category

Most correlated unigrams:

<table>
<thead>
<tr>
<th>Code</th>
<th>Top 5</th>
<th>Bottom 5</th>
</tr>
</thead>
<tbody>
<tr>
<td>04</td>
<td>earth, airborne, geophysical, mount, igs</td>
<td>al, unit, two, australia, region</td>
</tr>
<tr>
<td>15</td>
<td>study, financial, survey, university, dataset</td>
<td>given, number, received, document, expert</td>
</tr>
</tbody>
</table>

04: Earth Science
15: Commerce, Management, Tourism and Services
Examples of classification within two-digits code

Method: MLR
06: Biological Sciences (41505 records)
02: Physical Sciences (3533 records)

06: 17268 records (out of 41505) have both 0601 and 0604 labels
Discussion and future work

- User behaviour:
  - Evidence that subject headings are used
    - Why and why not
  - Low usage of subject headings from this log collection
    - Is this unique to this data catalogue and interface?
      Log analysis + survey and interview

- Collection characteristics:
  - Large proportion of records from the catalogue without a “standard” vocabulary for the subject headings a known issue
  - Those with subject headings are biased toward a few categories
    - Encourage underrepresented subject areas to publish and share data
  - Record classification works for some categories
    - Explore correlation, improvement
Thanks!