Automatically classifying data records using ANZSRC-FoR subject headings

Mingfang Wu
Australian Research Data Commons, Australia

2021-10-13
DCMI 2021 Virtual

#ARDC_AU
Outlines

- ANZSRC-FoR subject headings: A background about the use of subject metadata in the data catalogue: Research Data Australia

- Experiment: Automatic labelling/classification of data records with the schema - Approach & Result & Implication
There has been a growing number of data repositories & catalogues for publishing and sharing data.

re3data.org, the Registry of Research Data Repositories, had 23 repositories when it went online in 2012; the number quickly increased to over 1,200 data repositories from across the globe in three years and, by February 2020, the registry had more than 2450 repositories.
Subject metadata

● Dublin Core Definition:
  ○ A topic of the content of the resource.
  ○ Typically, subject will be expressed as keywords, key phrases, or classification codes that describe a topic of the resource. Recommended best practice is to select a value from a controlled vocabulary or formal classification scheme (e.g. Faceted Application of Subject Terminology).

● Benefit: Subject metadata is a powerful way of knowledge organisation and linkage of (distributed) resources for interoperability and discovery
  ○ Subject metadata is included in almost all metadata schemas. For schemas describing dataset, most of them include subject metadata as an optional not mandatory field.

● Cost: Manually labelling resources with subject metadata is not efficient and may introduce inconsistency and omission.
Research Data Australia - A National Data Catalogue

186K+ metadata records of datasets
60K+ research grants
100 Contributors

Browse by subjects
Disease gene prediction database

Deakin University

Dr Merridee Wouters (Aggregated by) Mr Martin Oti (Aggregated by)

Access the data

Cite
Save to MyRDA

Full description

This database includes gene predictions for disease phenotypes based on published Genome-Wide Association Data. May be used to choose primers for phenotype-specific resequencing of patient DNA.

For each prediction for following data is listed: phenotype, predicted gene, significant SNP, datasource, datasource reference.

Notes

The data was generated by a computer from clinical data, and some data from HuGE (http://hugenavigator.net/HuGENavigator/home.do) was used. The data is organised within a searchable

Subjects

Biological Sciences | Clinical Health (Organs, Diseases and Abnormal Conditions) | Genetics | Genetics Not Elsewhere Classified | Health | Inherited Diseases (incl. Gene Therapy) | database | genetic databases | genome-wide association study | humans | polymorphism | protein disease/genetics | single nucleotide | software |
Types of subject classification schemas

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) (2008 version)

<table>
<thead>
<tr>
<th>Division</th>
<th>Terms Coded</th>
</tr>
</thead>
<tbody>
<tr>
<td>Division 01 Mathematical Sciences</td>
<td></td>
</tr>
<tr>
<td>Division 02 Physical Sciences</td>
<td></td>
</tr>
<tr>
<td>Division 03 Chemical Sciences</td>
<td></td>
</tr>
<tr>
<td>Division 04 Earth Sciences</td>
<td></td>
</tr>
<tr>
<td>Division 05 Environmental Sciences</td>
<td></td>
</tr>
<tr>
<td>Division 06 Biological Sciences</td>
<td></td>
</tr>
<tr>
<td>Division 07 Agricultural and Veterinary Sciences</td>
<td></td>
</tr>
<tr>
<td>Division 08 Information and Computing Sciences</td>
<td></td>
</tr>
<tr>
<td>Division 09 Engineering</td>
<td></td>
</tr>
<tr>
<td>Division 10 Technology</td>
<td></td>
</tr>
<tr>
<td>Division 11 Medical and Health Sciences</td>
<td></td>
</tr>
<tr>
<td>Division 12 Built Environment and Design</td>
<td></td>
</tr>
<tr>
<td>Division 13 Education</td>
<td></td>
</tr>
<tr>
<td>Division 14 Economics</td>
<td></td>
</tr>
<tr>
<td>Division 15 Commerce, Management, Tourism and Service</td>
<td></td>
</tr>
<tr>
<td>Division 16 Studies in Human Society</td>
<td></td>
</tr>
<tr>
<td>Division 17 Psychology and Cognitive Sciences</td>
<td></td>
</tr>
<tr>
<td>Division 18 Law and Legal Studies</td>
<td></td>
</tr>
<tr>
<td>Division 19 Studies in Creative Arts and Writing</td>
<td></td>
</tr>
<tr>
<td>Division 20 Language, Communication, and Culture</td>
<td></td>
</tr>
<tr>
<td>Division 21 History and Archaeology</td>
<td></td>
</tr>
<tr>
<td>Division 22 Philosophy and Religious Studies</td>
<td></td>
</tr>
</tbody>
</table>

- **22 terms coded with two digits**
- **157 terms coded with four digits**
- **1238 terms coded with six digits**

1417 terms in three layers

0201 Astronomical and Space Sciences
0202 Atomic, Molecular, Nuclear, Particle and Plasma Physics
0203 Classical Physics
0204 Condensed Matter Physics
0205 Optical Physics
0206 Quantum Physics
0299 Other Physical Sciences

- **020101 Astrobiology**
- **020102 Astronomical and Space Instrumentation**
- **020103 Cosmology and Extragalactic Astronomy**
- **020104 Galactic Astronomy**
- **020105 General Relativity and Gravitational Waves**
- **020106 High Energy Astrophysics; Cosmic Rays**
- **020107 Mesospheric, Ionospheric and Magnetospheric Physics**
- **020108 Planetary Science (excl. Extraterrestrial Geology)**
- **020109 Space and Solar Physics**
- **020110 Stellar Astronomy and Planetary Systems**
- **020199 Astronomical and Space Sciences not elsewhere classified**

**Exclusions:**
- a) String theory is included in Group 0206 Quantum Physics.
- b) Tropospheric and stratospheric physics are included in Group 0401 Atmospheric Sciences.
- c) Extraterrestrial geology is included in Group 0403 Geology.
- d) Satellite and space vehicle design and testing is included in Group 0901 Aerospace Engineering.
- e) Remote sensing is included in Group 0909 Geomatic Engineering.
- f) Communications technologies using satellites are included in Group 1005 Communications Technologies.
Number of records per anzsrc-for two digits

04: Earth Sciences  
06: Biological Sciences  
21: History and Archaeology
Machine learning for classifying/labelling subject metadata

- 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:
  - Classifier - four supervised machine learning methods:
    - multinomial logistic regression (MLR), multinomial naive bayes (MNB), K Nearest Neighbors (KNN), Support Vector Machine (SVM)
  - Labels - top layer 22 labels from the ANZSRC-FoR code
  - Data - (~78k) records with anzsrc-for code
    - Extract title & description from each metadata record
    - Split into two sets: training set, test set

- Apply model(s)/best prediction to test set
### Result

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>MLR</td>
<td>0.76</td>
<td>0.70</td>
</tr>
<tr>
<td>SVM</td>
<td>0.70</td>
<td>0.67</td>
</tr>
<tr>
<td>KNN</td>
<td>0.92</td>
<td>0.66</td>
</tr>
<tr>
<td>MNB</td>
<td>0.70</td>
<td>0.66</td>
</tr>
</tbody>
</table>
Performance per label/category

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

Most correlated unigrams:

- 04: Earth Science
- 15: Commerce, Management, Tourism and Services
What we have learnt

- Large proportion of records from the catalogue don’t the subject metadata a known issue
- Those with subject metadata are biased toward a few categories
  - Encourage underrepresented subject areas to publish and share data
- Automatic classification works for some categories
  - Explore correlation and improvement, and other machine learning methods
Discussion

- There does not exist a classifier that fits for all text classification tasks or all collections. This leads to a generalisation issue, it is hard for organisations who want to improve subject metadata but have no resources to train classification models for their collection.
- Traditional ML methods require a good amount of training dataset with quality annotation. Some new technologies (e.g. deep learning and transfer learning) may provide a way, for example, wording embedding technique by learning word association from other similar big collections.
- Practically there is a question about how to use ML outputs: MLs suggest labels based on probability, human expertise may be necessary to look at a range of suggestions and make a decision, this may lead to improve ML intelligence and classification accuracy - we help the ML community to help ourselves.
Acknowledgement

Ying-Hsang Liu, Oslo Metropolitan University, Norway
Rowan Brownlee, Australian Research Data Commons
Jenny Xiuzhen Zhang, RMIT University, Australia
Thank you
Contact: mingfang.wu@ardc.edu.au

www.ardc.edu.au

LinkedIn: australian-research-data-commons
Twitter: ARDC_AU