Automatic Classification Using DDC on the Swedish Union Catalogue

Koraljka Golub, Johan Hagelbäck, Anders Ardö

18th European NKOS Workshop, TPDL 2018 and DCMI 2018, Porto, Portugal
Johan Hagelbäck
SENIOR LECTURER

Department of Computer Science and Media Technology
Faculty of Technology

johan.hagelback@lnu.se
+46 8 049 77 07

A:34.25, Kalmar Nyckel, Kalmar

Anders Ardö
9.75
Contents

1. Purpose and aims

2. Method

3. Results

4. Future research
Purpose and aims

• To establish the value of automatically produced classes for Swedish digital collections

• Aims
  • Develop (and evaluate) automatic subject classification for Swedish textual resources from the Swedish union catalogue (LIBRIS)
    • http://libris.kb.se
  • Data set: 143,756 catalogue records containing DDC in LIBRIS
  • Using a machine learning approach
    • Multinomial Naïve Bayes (NB)
    • Support Vector Machine with linear kernel (SVM)
Rationale...

- Lack of subject classes and index terms from KOS in new digital collections
# All resource types

<table>
<thead>
<tr>
<th>Resource type</th>
<th>Free text</th>
<th>Person</th>
<th>Organisation</th>
<th>Role</th>
<th>Title</th>
<th>Year</th>
<th>Object type</th>
<th>Collection</th>
<th>Subject</th>
<th>Licensing</th>
<th>Format</th>
</tr>
</thead>
<tbody>
<tr>
<td>-Select from list-</td>
<td>Start writing to get alternatives</td>
<td>Start writing to get alternatives</td>
<td>-Select from list-</td>
<td>-Select from list-</td>
<td></td>
<td></td>
<td>-Select from list-</td>
<td>-Select from list-</td>
<td></td>
<td>-Select from list-</td>
<td></td>
</tr>
</tbody>
</table>
... Rationale

- DDC chosen as a new national standard in 2013

- LIBRIS has a large collection of resources with DDC assigned to Swedish resources to train on

- Explore automatic classification on Swedish DDC → interoperability, cross-search, multilingual, international...
Contents

1. Purpose and aims

2. Method

3. Results

4. Future research
DDC

- 23rd edition, MARCXML format
- 128 MB → relevant info extracted into MySQL database, total of 14,413 classes

- Class number (field 153, subfield a);
- Heading (field 153, subfield j);
- Relative index term (persons 700, corporates 710, meetings 711, uniform title 730, chronological 748, topical 750, geographic 751; with subfields);
- Notes for disambiguation: class elsewhere and see references (253 with subfields);
- Scope notes on usage for further disambiguation (680 with subfields); and,
- Notes to classes that are not related but mistakenly considered to be so (353 with subfields).
Data collection

- LIBRIS: 143,838 catalogue records in April 2018
  - Using OAIPMH protocol, MARCXML format
  - All LIBRIS records with 082 MARC field for DDC class
  - Relevant info extracted into MySQL:
    - Control number (MARC field 001), unique record identification number;
    - Dewey Decimal Classification number (MARC field 082, subfield a);
    - Title statement (MARC field 245, subfield a for main title and subfield b for subtitle); and,
    - Keywords (a group of MARC fields starting with 6*), where available -- 85.8% of records had at least one keyword.

- DDC classes truncated to 3-digit codes, to maximise training quality
Training problem: imbalance between classes

- The most frequent class is 839 (Other Germanic literatures) with 18,909 records
- In total 594 classes have less than 100 records (70 of those have only 1 single record)

→ A dataset called “major classes” containing only classes with at least 1,000 records:
  - 72,937 records spread over 29 classes
  (60,641 records spread over 29 classes when selecting records with keywords)
The different datasets generated from the raw LIBRIS data

<table>
<thead>
<tr>
<th>Dataset</th>
<th>ID</th>
<th>Records</th>
<th>Classes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Titles</td>
<td>T</td>
<td>143,838</td>
<td>816</td>
</tr>
<tr>
<td>Titles and keywords</td>
<td>T_KW</td>
<td>121,505</td>
<td>802</td>
</tr>
<tr>
<td>Keywords only</td>
<td>KW</td>
<td>121,505</td>
<td>802</td>
</tr>
<tr>
<td>Titles, major classes</td>
<td>T_MC</td>
<td>72,937</td>
<td>29</td>
</tr>
<tr>
<td>Titles and keywords, major classes</td>
<td>T_KW_MC</td>
<td>60,641</td>
<td>29</td>
</tr>
<tr>
<td>Keywords only, major classes</td>
<td>KW_MC</td>
<td>60,641</td>
<td>29</td>
</tr>
</tbody>
</table>
Classifiers

- Pre-processing
  - Bag-of-words approach (stop-words retained) → over 130,000 unique words
  - Unigrams and 2-grams
  - TF-IDF scores

- Multinomial Naïve Bayes (NB) and Support Vector Machine with linear kernel (SVM) algorithms
  - Both have been used in text classification numerous times with good results
  - SVM typically better results than NB, but slower to train
  - NB can be trained incrementally, i.e. new training examples can be added without having to retrain the model with all training data
Evaluation measure

- Accuracy

- Amount of correctly classified examples

\[
\text{Accuracy} = \frac{\text{Correctly classified examples}}{\text{Total number of examples}} \times 100\%
\]
Matching against catalogue records

- The following fields were used as input to the machine learning models:
  - Title (field 245, subfield a)
  - Subtitle (field 245, subfield b)
  - Keywords (all fields starting with 6)

- The target label for each example is the DDC category (field 082, subfield a) formatted into the first three digits
  - (resulting in 816 unique DDC categories in the dataset)
Contents

1. Purpose and aims

2. Method

3. Results

4. Future research
Major results

- SVM better than NB on all classes
  - On test set, best result $81.4\%$ accuracy for classes with over 1,000 training examples, or $61.3\%$ accuracy for all classes
  - When using both titles and keywords, unigrams and 2-grams

- Features
  - Number of training examples significantly influences performance
  - Keywords better than titles, keywords + titles best
  - 2-grams slightly better on keywords and keywords + titles, but much longer training time
  - Stemming only marginally improves results
## NB

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Accuracy, unigrams</th>
<th>Accuracy, unigrams + 2-grams</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Training set</td>
<td>Test set</td>
</tr>
<tr>
<td>T</td>
<td>83.54%</td>
<td>34.89%</td>
</tr>
<tr>
<td>T_KW</td>
<td>90.01%</td>
<td>55.33%</td>
</tr>
<tr>
<td>KW</td>
<td>75.28%</td>
<td>59.15%</td>
</tr>
<tr>
<td>T_MC</td>
<td>90.83%</td>
<td>54.21%</td>
</tr>
<tr>
<td>T_KW_MC</td>
<td>95.42%</td>
<td>76.52%</td>
</tr>
<tr>
<td>KW_MC</td>
<td>86.94%</td>
<td>77.25%</td>
</tr>
</tbody>
</table>

## SVM

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Accuracy, unigrams</th>
<th>Accuracy, unigrams + 2-grams</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Training set</td>
<td>Test set</td>
</tr>
<tr>
<td>T</td>
<td>93.74%</td>
<td>40.91%</td>
</tr>
<tr>
<td>T_KW</td>
<td>97.50%</td>
<td>65.25%</td>
</tr>
<tr>
<td>KW</td>
<td>83.09%</td>
<td>64.02%</td>
</tr>
<tr>
<td>T_MC</td>
<td>93.95%</td>
<td>57.99%</td>
</tr>
<tr>
<td>T_KW_MC</td>
<td>97.89%</td>
<td>80.75%</td>
</tr>
<tr>
<td>KW_MC</td>
<td>90.58%</td>
<td>79.56%</td>
</tr>
</tbody>
</table>
Contents

1. Purpose and aims

2. Method

3. Results

4. Future research
Try improve algorithm performance…

• Take advantage of DDC
  
  • Class number (field 153, subfield a);
  • Heading (field 153, subfield j);
  • Relative index term (persons 700, corporates 710, meetings 711, uniform title 730, chronological 748, topical 750, geographic 751; with subfields);
  • Notes for disambiguation: class elsewhere and see references (253 with subfields);
  • Scope notes on usage for further disambiguation (680 with subfields); and,
  • Notes to classes that are not related but mistakenly considered to be so (353 with subfields).

• Establish how these contribute to classification accuracy

• Evaluate ensemble learners combining different types of algorithms
  • String matching in the lack of training examples
...Try improve algorithm performance

- A major issue is the imbalance between the different DDC categories
  - One approach to combat this could be to try a two-level hierarchical classification model:
    - First, classify an example into one of the 10 main categories (first digit in the DDC class)
    - Second, classify the example into one of the (up to 100) subcategories in the main category (second and third digit in the DDC class)

- A more modern approach to text classification using word embeddings and deep learning could also be evaluated
  - The major advantage of word embeddings is understanding of context (not just evaluating word by word without any relation between two words), but since context is of limited importance in DDC classification it is likely that this approach will not be more accurate than NB/SVM
Evaluation

• Test for all levels of classes

• Test with algorithms outputting more than one class

• Include misses in evaluation using measures like F-measure combining precision and recall

• Evaluate in the context of retrieval in real IR tasks
Thank you for your attention!

• Questions?
• Feedback?
• Collaborative ideas?

• Contact: koraljka.golub@lnu.se