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

+46480497707

A:34.25, Kalmar Nyckel, Kalmar

Anders Ardö

9.75

Contents

Purpose and aims

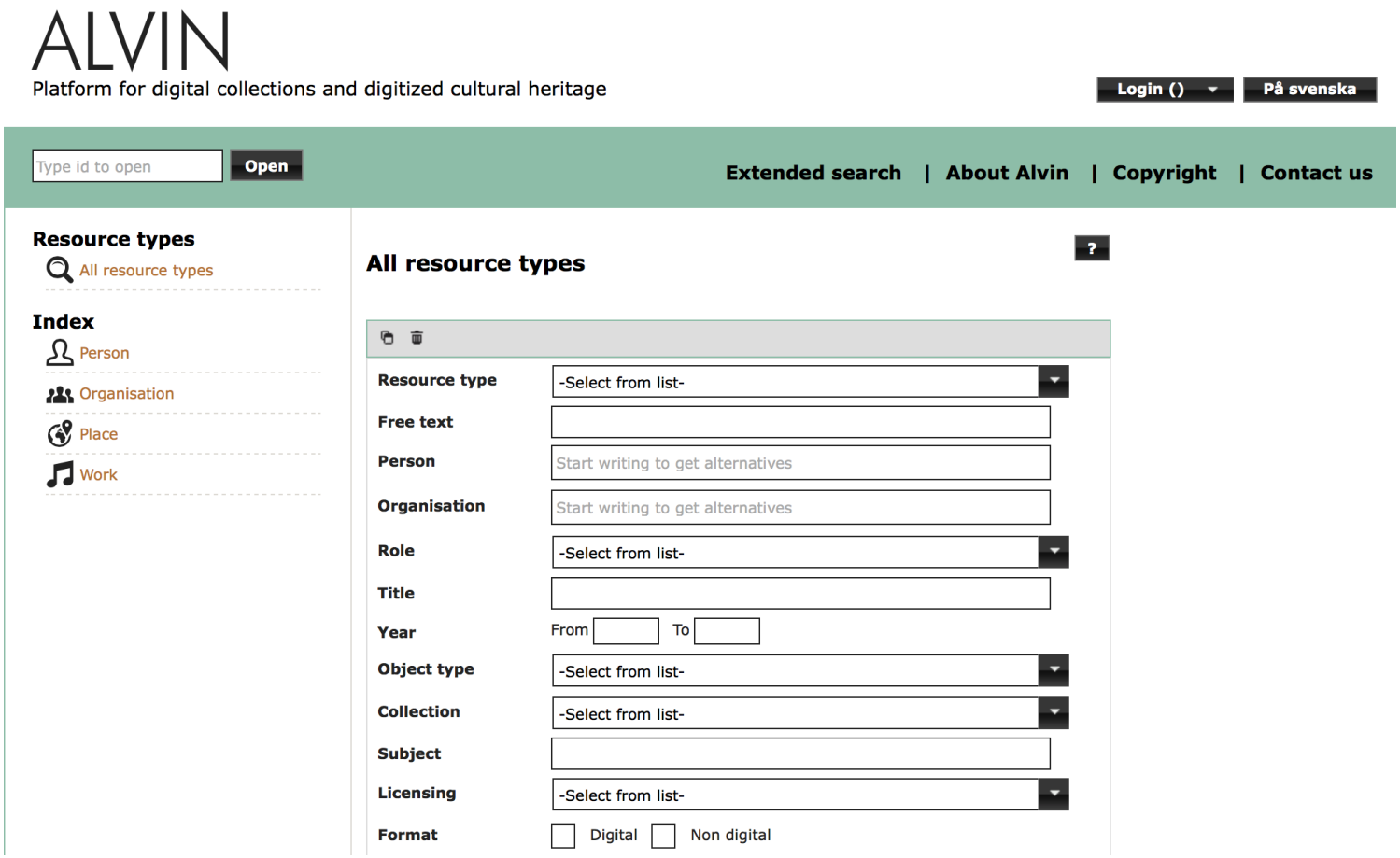
- To establish the value of automatically produced classes for Swedish digital collections

- Aims
  - Develop (and evaluate) automatic subject classification for Swedish textual resources from the Swedish union catalogue (LIBRIS)
    - http://libris.kb.se

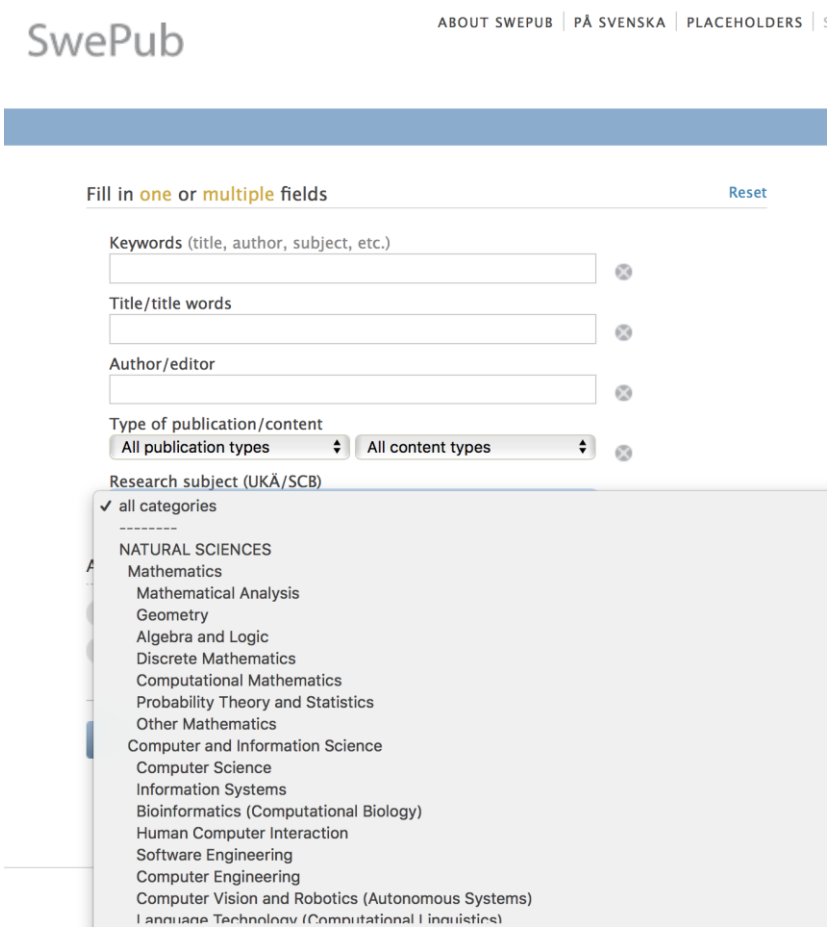
  - Data set: 143,756 catalogue records containing DDC in LIBRIS
  - Using a machine learning approach
    - Multinomial Naïve Bayes (NB)
    - Support Vector Machine with linear kernel (SVM)

Rationale…

- Lack of subject classes and index terms from KOS in new digital collections

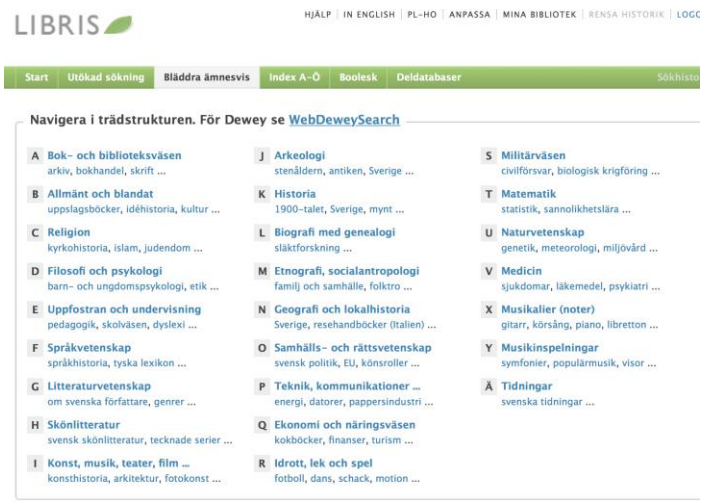

ALVIN
Platform for digital collections and digitized cultural heritage
Login ()
På svenska
Type id to open
Open
Extended search | About Alvin | Copyright | Contact us
Resource types
All resource types
Index
Person
Organisation
Place
Work
All resource types
?
Resource type
-Select from list-
Free text
Person
Start writing to get alternatives
Organisation
Start writing to get alternatives
Role
-Select from list-
Title
Year
From
To
Object type
-Select from list-
Collection
-Select from list-
Subject
Licensing
-Select from list-
Format
Digital
Non digital


Sondera
Three national sources. One view
Help | På Svenska | Taggar | Home
Archive Library Sound & Image
Search simultaneously in NAD, LIBRIS and SMDB to get an overview of how a person or an event is documented in TV, radio, writing and original documents.
SEARCH
Simple search
Keywords
Title
Creator (Author, archival creator, composer ...)
Year
SEARCH
Bladet
Mordet på Olof Palme

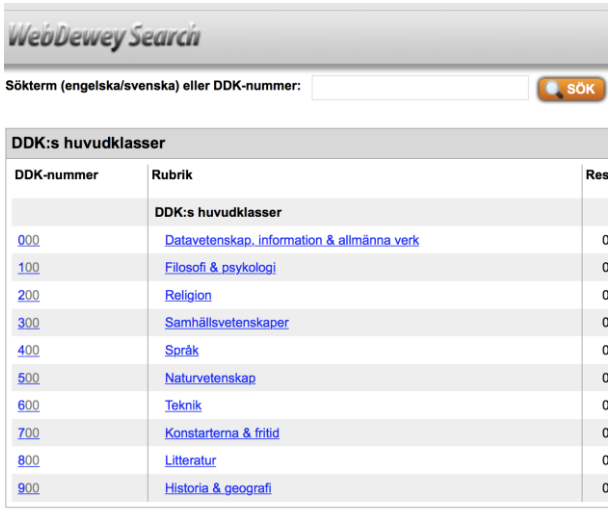

SwePub
ABOUT SWEPUB | PÅ SVENSKA | PLACEHOLDERS
Fill in one or multiple fields
Reset
Keywords (title, author, subject, etc.)
Title/title words
Author/editor
Type of publication/content
All publication types
All content types
Research subject (UKÄ/SCB)
all categories
NATURAL SCIENCES
Mathematics
Mathematical Analysis
Geometry
Algebra and Logic
Discrete Mathematics
Computational Mathematics
Probability Theory and Statistics
Other Mathematics
Computer and Information Science
Computer Science
Information Systems
Bioinformatics (Computational Biology)
Human Computer Interaction
Software Engineering
Computer Engineering
Computer Vision and Robotics (Autonomous Systems)

… Rationale

- DDC chosen as a new national standard in 2013



SAB → DDC



- 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…

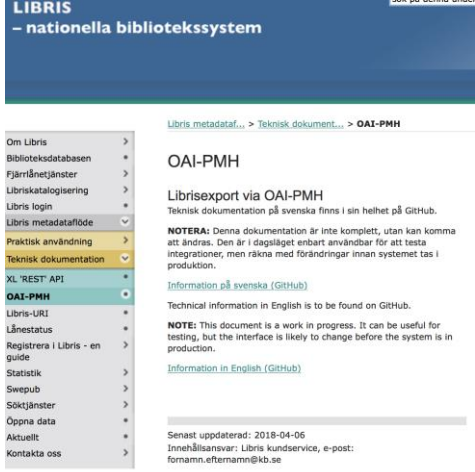
Contents

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

Dataset	ID	Records	Classes
Titles	T	143,838	816
Titles and keywords	T_KW	121,505	802
Keywords only	KW	121,505	802
Titles, major classes	T_MC	72,937	29
Titles and keywords, major classes	T_KW_MC	60,641	29
Keywords only, major classes	KW_MC	60,641	29

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

$$\text{Accuracy} = \frac{\text{Correctly classified examples}}{\text{Total number of examples}} \quad \%$$

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

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

Dataset	Accuracy, unigrams		Accuracy, unigrams + 2-grams	
	Training set	Test set	Training set	Test set
T	83.54%	34.89%	95.82%	34.15%
T_KW	90.01%	55.33%	98.14%	55.45%
KW	75.28%	59.15%	84.95%	58.11%
T_MC	90.83%	54.21%	98.63%	50.51%
T_KW_MC	95.42%	76.52%	99.66%	75.96%
KW_MC	86.94%	77.25%	94.24%	77.09%

SVM

Dataset	Accuracy, unigrams		Accuracy, unigrams + 2-grams	
	Training set	Test set	Training set	Test set
T	93.74%	40.91%	99.59%	40.45%
T_KW	97.50%	65.25%	99.90%	66.13%
KW	83.09%	64.02%	92.38%	64.09%
T_MC	93.95%	57.99%	99.62%	57.80%
T_KW_MC	97.89%	80.75%	99.93%	81.37%
KW_MC	90.58%	79.56%	96.30%	80.38%

Contents

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