This is the published version of a paper presented at ISKO UK Conference 2017: Knowledge Organization: what's the story?, 11 – 12 September 2017, London.

Citation for the original published paper:


N.B. When citing this work, cite the original published paper.

Permanent link to this version:
http://urn.kb.se/resolve?urn=urn:nbn:se:lnu:diva-67925
Automatic subject classification for improving retrieval in Swedish repositories

Koraljka Golub¹, Jukka Tyrkkö¹, Andreas Kerren², Ilir Jusufi³, Anders Ardö³

¹Faculty of Arts and Humanities, Linnaeus University, ²Faculty of Technology, Linnaeus University, ³Faculty of Engineering, Lund University

PROBLEM
- (New) digital collections generally lack controlled subject index terms or classes, preventing successful retrieval in serious academic inquiry.
- In SwPub, the unified Swedish repository, the only control of subject is a shallow national categorization of scientific disciplines.
- Dewey Decimal Classification (DDC) introduced in Sweden in 2013 as a new ‘national standard’ replacing SAB (Sveriges Allmänna Biblioteksförening; Swedish library classification system).

WHAT AND WHY
- Assign DDC classes and captions to Swedish metadata records in local DIVA repository
- Objective: to allow for subject browsing and searching based on an established knowledge organization system

HOW
- Two string matching algorithms ([1], [2]), alone and in combination; machine learning in combination with visualisation
- Take advantage of hierarchies for disambiguation and relative index terms as synonyms (Swedish DDC 23)
  - Apply and evaluate visualisation for determining the right class
- Compare against search logs in relation to existing national category (required) and author keywords (optional)
- Evaluate against carefully crafted gold standard and in retrieval [3]
- Implement interactive visualisations as part of the user interface; evaluate.

Sample
- About 25,000 repository records at LNU (all research publications, no student theses)
- Less than half have an abstract (11,464) ⇒ challenge of little text
- Only 1,272 records in Swedish have an abstract and optional author keywords
- Quality control problems (e.g., Swedish-language data with English-language tag) ⇒ cleaning up required

PRELIMINARY RESULTS
Algorithm 2 assigns top three levels only, using the same term space but different matching approach. Algorithm 1 identifies terms and calculates scores; Algorithm 2 takes all terms for a class, indexes them using a search engine (Lucene), and compares against repository records as a search query.

The figure on the left presents pairs of classes assigned by algorithms 1 and 2 for each article. Some match, some are complementary and the value of combining them will be further investigated in the future.

FUTURE WORK
- Data clean-up
- English DDC needed with mappings to Swedish
- Challenge of little text
- Fine-tuning of DDC term extraction (notes, hierarchies)
- Visualization / ML
- Evaluation, including in retrieval based on (a combination of) different subject access points

Acknowledgments
Thanks are due to OCLC which provided the project with electronic DDC, Swedish version. We are very grateful to Rebeca Green and Sandi Jones for all their advice on how to best process and use the electronic DDC files. The project is financed by IEC and Linnaeus University.

DATA EXTRACTION FROM MARC XML
- Class number (153a) and caption (153)
- 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)
- Scope notes on usage for further disambiguation (680)
- Notes to classes that are not related but mistakenly considered to be so (353)

Example for 616.9101 Medicinski virologi

Results examples
1. Title: ADHD: Diagnostik och behandling, vårdens organisation och patientens delaktigheten systematisk litteraturöversikt
   National category: Medical and Health Sciences; Health Sciences
   Algorithm 1: 616.8589 ADHD (uppmärksamhetsstörning med hyperaktivitet)
   Algorithm 2: 616 Sjukdomar

2. Akademiska yrkestitelning i Växjö: Program, studenter och arbetsermärkningar
   National category: Sociology
   Algorithm 1: 808.0663 Skrivande för och om högre utbildning
   Algorithm 2: 378 Högre utbildung (tertium utbildning)

NARRATIVES
- The categories in SAB and DDC exhibit only partial matches.
- Disciplinary narratives differ in terms of granularity but also conceptualization of concepts:
  - In DDC VIROLOGY is a subcategory of DISEASES, in SAB a subcategory of GENETICS AND MICROBIOLOGY
- LEADERSHIP not a concept in SAB under samhällskunskap (civics) at all

References