(Semi)-automated subject indexing of Swedish resources:

Evaluating (a combination of) cataloguers’, end users’ and automated index terms in retrieval

Koraljka Golub

International Conference on Dublin Core and Metadata Applications 2016 (DC-2016)
Purpose

- Improve searching and browsing for information in the Swedish language, esp. in interoperable information systems like LIBRIS, SwePub, Sondera

- Subject index terms from controlled vocabularies like SAO and DDC offer:
  - Uniformity of term format
  - Provide context
  - Browsing

- But, expensive while increase of digital documents → 2 possible solutions:
  - 1) (semi)-automated solutions
  - 2) author/end-user tagging
Your search on "växjö" resulted in 32270 hits

Archives

KFUM Växjö
KFUM Växjö
Fonds, 1945–1945

Sound & Image

Radio (11003) TV (8248)
Recorded sound (392) Film/video (167)
Multimedia (11) Texts (1)

19818 Hits

 [=Växjö, Föreningen Öppna Kanalen Växjö, 2015-08-03--2015-08-09]
TV, 2015-08-03--2015-08-09

 [=Växjö, Föreningen Öppna Kanalen Växjö, 2015-11-02--2015-11-08]
TV, 2015-11-02--2015-11-08

Morgon i P4 med Peje Johansson, Anneli Koskinen och Lars-Peter Hielle [Växjö, ...
Radio (Digitized), 2016-09-23

6547 Hits

Library

Book (4684) Article/chapter (907)
Journal etc. (373) Other (215)
Speech (136) Film/video (67)
Map (66) Image (36)
Multimedia (30) Poster (15)
Musical score (12) Music (3)
Manuscript (3)

Välkommen till Växjö
[Book, 1985]

Välkommen till Växjö
[Book, 1989]

Växjö : gatu- och kvartersnamn / Eva Selling
Selling, Eva, 1926-
<table>
<thead>
<tr>
<th>Term</th>
<th>Index</th>
</tr>
</thead>
<tbody>
<tr>
<td>Växjö</td>
<td>subject</td>
</tr>
<tr>
<td>Växjö 1600-talet</td>
<td></td>
</tr>
<tr>
<td>Växjö 1700-talet kartor</td>
<td></td>
</tr>
<tr>
<td>Växjö 1800-talet</td>
<td></td>
</tr>
<tr>
<td>Växjö 1800-talet kartor</td>
<td></td>
</tr>
<tr>
<td>Växjö 1920-talet</td>
<td></td>
</tr>
<tr>
<td>Växjö 1920-talet 1920-talet</td>
<td></td>
</tr>
<tr>
<td>Växjö 1990-talet</td>
<td></td>
</tr>
<tr>
<td>Växjö Als, se Växjö allmänna idrottsföreningen</td>
<td></td>
</tr>
<tr>
<td>Växjö–Alvesta järnväg</td>
<td></td>
</tr>
<tr>
<td>Växjö Arbetarekommun</td>
<td></td>
</tr>
<tr>
<td>Växjö Aringsås Bolmsö Drev Jät Söraby Öja Öjaby</td>
<td></td>
</tr>
<tr>
<td>Växjö Asa</td>
<td></td>
</tr>
<tr>
<td>Växjö Attsjö Skönstorps Nytorp</td>
<td></td>
</tr>
<tr>
<td>Växjö barn- och ungdomslitteratur</td>
<td></td>
</tr>
<tr>
<td>Växjö Berg</td>
<td>15</td>
</tr>
<tr>
<td>Växjö Berg Bjursjö</td>
<td></td>
</tr>
<tr>
<td>Växjö Berg Byggnader</td>
<td></td>
</tr>
<tr>
<td>Växjö Berg Hjulatorp</td>
<td></td>
</tr>
<tr>
<td>Växjö Berg Kyrkhemmet</td>
<td></td>
</tr>
<tr>
<td>Växjö Berg Källuhult</td>
<td></td>
</tr>
<tr>
<td>Växjö Berg lokalhistoria</td>
<td></td>
</tr>
<tr>
<td>Växjö Berg Matkull</td>
<td></td>
</tr>
<tr>
<td>Växjö Berg Vinverige</td>
<td></td>
</tr>
<tr>
<td>DDK-nummer</td>
<td>Rubrik</td>
</tr>
<tr>
<td>------------</td>
<td>---------------------------------------</td>
</tr>
<tr>
<td></td>
<td><strong>DDK:s huvudklasser</strong></td>
</tr>
<tr>
<td>500</td>
<td>Naturvetenskap</td>
</tr>
<tr>
<td>540</td>
<td>Kemi</td>
</tr>
<tr>
<td>541-547</td>
<td>Kemi</td>
</tr>
<tr>
<td>541</td>
<td>Fysikalisk kemi</td>
</tr>
<tr>
<td>542</td>
<td>Tekniker, procedurer, apparatur, utrustning, material</td>
</tr>
<tr>
<td>543</td>
<td>Analytisk kemi</td>
</tr>
<tr>
<td>546</td>
<td>Oorganisk kemi</td>
</tr>
<tr>
<td>547</td>
<td>Organisk kemi</td>
</tr>
</tbody>
</table>
Aims

1. Find out to what degree it is possible to apply automated subject indexing based on:
   - Controlled indexing languages like the Dewey Decimal Classification (DDC) and Swedish Subject Headings (SAO)
   - Derived indexing of keywords from the resource itself

2. Determine the value of automatically assigned index terms, in combination and comparison with end-user and cataloguers index terms in the process of information retrieval by end users
BACKGROUND
End-user indexing

- Author keywords or social tags in Web 2.0 services also provided by library catalogues
- Cheaper, provide additional perspectives like new scientific terms
- But, no control of word forms, homonymy, polisemity or synonymy

“…the cost savings made in the provision of low-quality indexing are cancelled out by the high costs incurred by searchers who fail either to find everything that they want (low recall) or, often more frustratingly, to avoid everything that they do not want (low precision)…”

(Furner 2010, 1861)
Offering users to choose from KOS
Offering users to choose from KOS

- The importance of controlled vocabulary suggestions for indexing and retrieval:
  - To help produce ideas of which tags to use
  - To make it easier to find focus for the tagging
  - To ensure consistency
  - To increase the number of access points in retrieval

- However, the value and usefulness of the suggestions proved to be dependent on the quality of the suggestions, both as to conceptual relevance to the user and as to appropriateness of the terminology

(Golub, Lykke, Tudhope 2014)
Automatic indexing…

3 major approaches

– Text categorization

– Document clustering

– String matching
…Automatic indexing…

Automatic indexing beneficial
- Address the scale and sustainability
- Enrich bibliographic records
- Establish more connections across resources

Reported success of automated tools
- Entirely replace manual indexing to machine-aided indexing (MAI)
  - MAI example: NLM’s Medical Text Indexer
Automatic indexing

Evaluation problem (Golub et al. 2016)

- Research comparing automatic versus manual indexing is flawed (Lancaster 2003, p. 334)

  • Out of context, laboratory conditions
  • Few reports on indexing tools in operating information systems
Challenge A: relevance 1/2

- Purpose of indexing: making relevant documents retrievable
- Relevance
  - A complex phenomenon
    - Many possible document-query relationships
    - E.g., for children/scientists, query/information need/task...
  - Subjective
  - Multidimensional and dynamic (Borlund 2003)
Challenge A: relevance 2/2

In practice, evaluation of IR is based on pre-existing relevance assessments

– Initiated by Cranfield tests
– A gold standard
  • A test collection consisting of a set of documents
  • A set of ‘topics’
  • A set of relevance assessments

– “In spite of the dynamic and multidimensional nature of relevance, in practice evaluation of information retrieval systems has been reduced to comparison against the gold standard—a set of pre-existing relevance judgments which are taken out of context. An early study on retrieval conducted by Gull in 1956 powerfully influenced the selection of a method for obtaining relevance judgments. Gull reported that two groups of judges could not agree on relevance judgments. Since then it has become common practice to not use more than a single judge or a single object for establishing a gold standard.”

(Saracevic 2008, 774)
Challenge B: indexing

Aboutness
   - Dependent on factors like interest, task, purpose, knowledge…

Exhaustivity and specificity of indexing
   - Related to indexing policies at hand
   - A subject correctly assigned in a high-exhaustivity system may be erroneous in a low-exhaustivity system

Terms assigned automatically but not manually might be wrong or they might be right but missed by manual indexing
   → Not good to use just the existing classes as the gold standard
Methodology …

• Based on Golub et al. (2016)
• Data collection: a subset of SMDB, SND, SwePub

• A comparison of assigned terms against a carefully crafted ‘gold standard’ and in the context of actual information retrieval

  – The ‘gold standard’ developed through input of
    • Professional catalogue librarians
    • End users who are experts in the subject at hand,
    • End users who are inexperienced in the subject
    • Several automated subject indexing software applications

  – Information retrieval will involve end users conducting actual searching on the indexed collection of resources and marking how relevant each retrieved resource is
… Methodology

• Automated subject index terms derived from several algorithms will be compared against the ‘gold standard’ and in the retrieval test

• The analysis will also include looking at what caused the retrieval of the document at hand: a cataloguer’s term, subject expert’s term, inexperienced user’s term or an automated term

• Also log analysis and questionnaires to help contextualize the results
Automated indexers

- To be built/adjusted for DDC, SAO and the Swedish language
- String matching
- Machine learning
- Commercial: Data Harmony (rule-based)
SIGNIFICANCE
Significance…

The value of the professional, the automated and the end-user ways of creating subject index terms will be determined
→ Allowing for informed decisions on ensuring high quality subject access points as part of the Swedish library and information infrastructure

– Cheap assignment of controlled subject terms useful at various stages of the metadata creation workflow:
  • By an author creating original index terms at the time of deposit;
  • By a reader annotating (for colleagues/world or for recommendation for inclusion in a collection);
  • By a cataloguer
...Significance

– Provision of (semi)-automated solutions for assigning DDC will enable:

  • Hierarchical browsing by subject
  • Retrievability of Swedish resources in multilingual systems
  • Integration of Swedish resources into the Semantic Web (as DDC is available as Linked Data)

– The resulting empirically tested comprehensive methodology framework will be of interest to other researchers
Partners

Linnaeus University
Lund University
University of South Wales

University of Aalborg
University of Buffalo
Charles Sturt University

Swedish National Library
Linnaeus University Library
Access Innovations
OCLC
References


