Evaluating automatic subject indexing: A framework

A Calculus of Culture | Circumventing the Black Box of Culture Analytics, Nanning, 21-22 March 2017

Koraljka Golub, koraljka.golub@lnu.se

Linnæus University

Co-authors

- This work is directly based on joint work with the following researchers:
  - Dagobert Soergel, University of Buffalo, USA
  - George Buchanan, City University, London, UK
  - Douglas Tudhope, University of South Wales, UK
  - Marianne Lykke, University of Aalborg, Denmark
  - Debra Hiom, University of Bristol, UK

Golub, K., Soergel, D., Buchanan, G., Tudhope, D., Hiom, D., & Lykke, M. (2016). A framework for evaluating automatic indexing or classification in the context of retrieval. *Journal of the Association for Information Science and Technology,* 67(1), pp. 3-16.

Table of contents

- Part 1: Evaluation framework
  - Background
  - Suggested framework for evaluating automatic indexing

- Part 2: Application example
  - 3D DDC project
  - 3D DDC project: Methodology
  - 3D DDC project: Significance

Background

Introduction 1(2)

- 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
    - E.g., NLM´s Medical Text Indexer

MTI
NLM Medical Text Indexer
Providing Indexing Assistance
Since 2002
Biomedical Literature
MTI
MTIFL
MeSH Suggestions

Introduction 2(2)

Evaluation problem

- Research comparing automatic versus manual indexing is seriously flawed (Lancaster 2003, p. 334)
  - Out of context, laboratory conditions
  - Few reports on indexing tools in operating information systems

Suggested framework

- Based on a comprehensive literature review
- Three components of evaluating indexing quality:
  - **Directly** by an **evaluator** or comparison with a **gold standard**
  - **Directly** in an indexing **workflow**
  - **Indirectly** through analyzing **retrieval** performance

Terminology

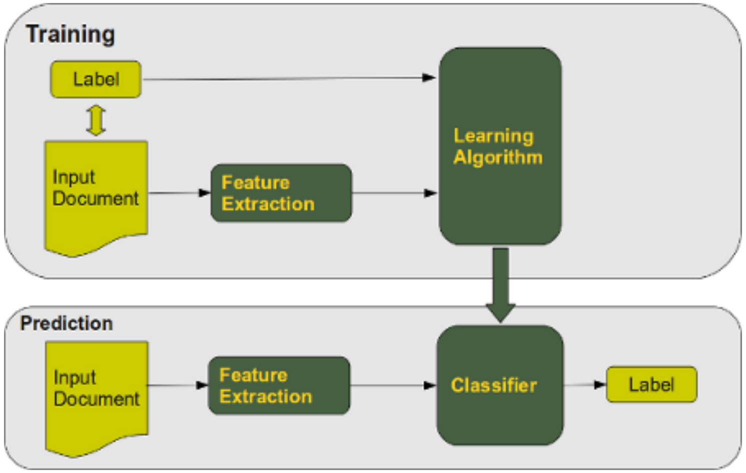
- Indexing: (un)controlled term assignment
  - Subject indexing: typically 3-20 subject index terms
    → to allow retrieval from various perspectives
  - Subject classification: typically 1 pre-combined class
    → mostly for browsing

- Automatic/automated indexing/classification
  - A variety of terms in literature, also prevalent:
    - Text categorization
    - Document clustering

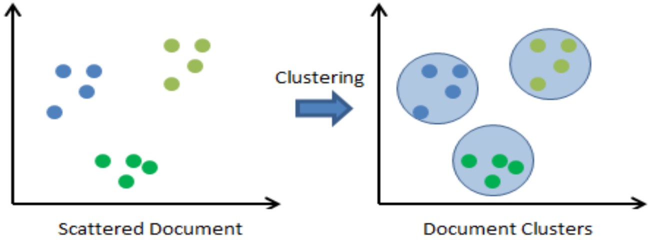
Automatic indexing

3 major approaches

- Text categorization



- Document clustering



- Document classification

Challenge A: relevance 1/3

- Purpose of indexing: making relevant documents retrievable

- Relevance
  - A complex phenomenon
    - Many possible document-query relationships

  - Subjective

  - Multidimensional and dynamic (Borlund 2003)

Challenge A: relevance 2/3

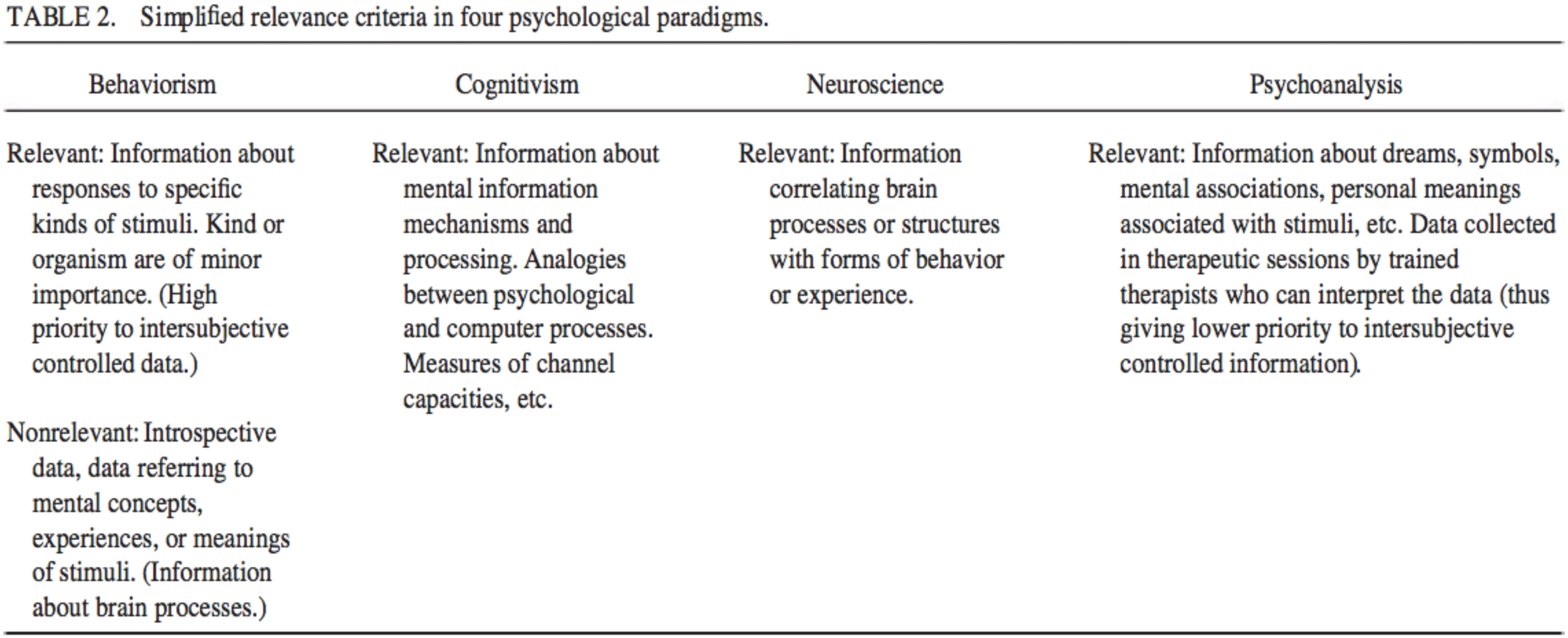


TABLE 2. Simplified relevance criteria in four psychological paradigms.

Behaviorism	Cognitivism	Neuroscience	Psychoanalysis
Relevant: Information about responses to specific kinds of stimuli. Kind or organism are of minor importance. (High priority to intersubjective controlled data.) Nonrelevant: Introspective data, data referring to mental concepts, experiences, or meanings of stimuli. (Information about brain processes.)	Relevant: Information about mental information mechanisms and processing. Analogies between psychological and computer processes. Measures of channel capacities, etc.	Relevant: Information correlating brain processes or structures with forms of behavior or experience.	Relevant: Information about dreams, symbols, mental associations, personal meanings associated with stimuli, etc. Data collected in therapeutic sessions by trained therapists who can interpret the data (thus giving lower priority to intersubjective controlled information).

Challenge A: relevance 3/3

- 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 1/3

- ISO 5963:1985
  - Document-oriented definition of subject indexing
  - Three steps
    1. Determining the subject content of a document
    2. A conceptual analysis to decide which aspects of the content should be represented
    3. Translation of those concepts or aspects into a controlled vocabulary

- Request-oriented indexing (user-oriented)
  - The indexer's task is to understand the document and then anticipate for what topics or uses this document would be relevant

Challenge B: indexing 2/3

- Aboutness
  - Dependent on factors like interest, task, purpose, knowledge, norms, opinions and attitudes
  - Social tagging offers potential end-user perspectives

- 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

- Inter-indexer and intra-indexer inconsistency
  - Worse with higher exhaustivity and specificity and bigger vocabularies

Challenge B: indexing 3/3

- Indexing can be consistently wrong as well as consistently good
  - High indexing consistency not always a sign of good indexing quality

- 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

Suggested framework for evaluating indexing

Overview

Triangulation of methods and exploration of multiple perspectives and contexts
3 complementary approaches:

1. Evaluating indexing quality **directly** through assessment by an **evaluator** or by comparison with a **gold standard**.

2. Evaluating indexing quality **directly** in the context of an **indexing workflow**.

3. Evaluating indexing quality **indirectly** through **retrieval** performance.

Evaluating directly through an evaluator or a gold standard

- 2 main approaches:
  1. Ask evaluators to assess index terms assigned
  2. Compare to a gold standard
     - Used a lot by text categorization community
       - Text collections for training and evaluation (e.g., Reuters)

- Problems of relevance and indexing characteristics

- The validity and reliability of results derived solely from a gold-standard evaluation remains unexamined

Evaluating directly through an evaluator or a gold standard: recommendations (1/3)

- Select 3 distinct subject areas that are well-covered by the document collection
  - For each subject area, select 20 documents at random
- 2 professional subject indexers assign index terms as they usually do (or use index terms that already exist)
- 2 subject experts assign index terms
- 2 end users who are not subject experts assign index terms

Evaluating directly through an evaluator or a gold standard: recommendations (2/3)

- Assign index terms using all indexing methods to be evaluated (for example, several automatic indexing systems to be evaluated and compared)

- Prepare document records that include all index terms assigned by any method in one integrated listing

- 2 senior professional subject indexers and preferably 2 end users examine all index terms, remove terms assigned erroneously, and add terms missed by all previous processes

Evaluating directly through an evaluator or a gold standard: recommendations (3/3)

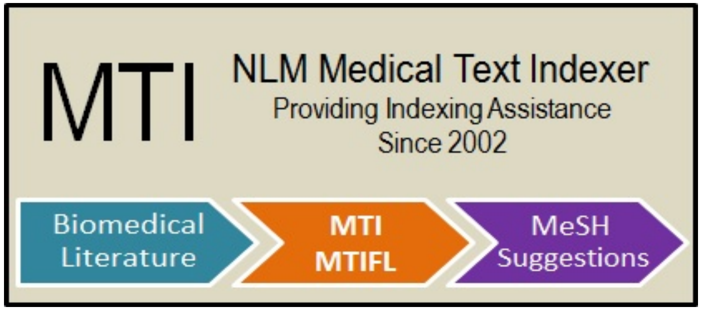
- Number of indexers, documents etc. must consider the context and available resources

- No studies how the numbers affect results

- Intuitively, less than 20 documents per subject area would make the results quite susceptible to random variation

Evaluating MAI tools in an indexing workflow

- Automatic indexing tools can be used for machine-aided indexing (MAI)
  - E.g., Medical Text Indexer


MTI
NLM Medical Text Indexer
Providing Indexing Assistance Since 2002

- Evaluating the quality of MAI tools should assess the value of providing human indexers with automatically generated index term suggestions

Evaluating in an indexing workflow: recommendations 1/2

- 4 phases

1. Collecting baseline data on unassisted manual indexing
2. A familiarization tutorial for indexers
3. An extended in-use study
   - Observe practicing subject indexers in different subject areas
   - Determine the indexers' assessments of the quality of the automatically generated subject term suggestions
   - Identify usability issues
   - Evaluate the impact of term suggestions on terms selected
4. A summative semi-structured interview

Evaluating in an indexing workflow: recommendations 2/2

- Such evaluation should consider:
  - The quality of the tool's suggestions
  - The usability of the tool in the indexing workflow
  - The indexers' understanding of their task
  - The indexers' experience with MAI
  - The resulting quality of the final indexing
  - Time saved

Evaluating indirectly through retrieval performance

- The major purpose of subject indexing is successful information retrieval
  - Assessing indexing quality by comparing retrieval results from the same collection using indexing from different sources
  - Emphasis on detailed analysis of how indexing contributes to retrieval successes or failures

- Soergel (1994): a logical analysis of effects of subject indexing on retrieval performance
  - Highly complex → need for real-like evaluation

Evaluating through retrieval: recommendations 1/3

- A test collection of ~10,000 documents
  - Drawn from an operational collection with available controlled terms
  - Covering several (three or more) subject areas

- Index some or all of these documents with all of the indexing methods to be tested

- For each of the subject areas, choose a number of users
  - Ideally, equal numbers of end users, subject experts, and information professionals

Evaluating through retrieval: recommendations 2/3

- Users conduct searches on several topics
  - Some topics chosen by the user and some assigned
  - 1 topic: an extensive search for an essay or so requiring an extensive list of documents (likely to benefit from the index terms)
  - 1 topic: a factual search for information (may be less dependent on index terms)

- Users assess the relevance of each document found
  - Scale from 0 to 4, not relevant to highly relevant
  - Instruct the users how to assess relevance in order to increase inter-rater consistency

Evaluating through retrieval: recommendations 3/3

- Compute retrieval performance metrics for each individual indexing source and for selected combinations of indexing sources at different degrees of relevance

- Perform log analysis, observe several people how they perform their tasks, get feedback from the assessors through questionnaires and interviews
  - Consider also the effect of the user's query formulation

- Perform a detailed analysis of retrieval failures and retrieval successes, focusing on cases where indexing methods differ with respect to retrieving a relevant or irrelevant document

Conclusion

- Potential of automatic subject indexing
- Some claims of high success of automatic tools, but big evaluation challenge
- Proposed framework comprising 3 aspects: direct evaluation, direct evaluation in an indexing workflow, indirect evaluation through retrieval
  - Needs to be informed by empirical evidence

3D DDC project

(Semi)-automated subject indexing of Swedish resources based on Dewey Decimal Classification:

Evaluating (a combination of)

cataloguers', end users' and automatic index terms

Purpose

- Improve search for information in the Swedish language, esp. in information systems like libraries and SwePub
- Subject searching is a common yet the most challenging type of searching in information retrieval systems
- Subject index terms from controlled vocabularies like SAO and DDC offer:
  - Uniformity of term format
  - Relationships among terms
  - Browsing e.g., http://www.udcsummary.info/
- But, expensive while increase of digital documents (research data, repositories, SMDB) → (semi)-automated solutions and authors/end-users

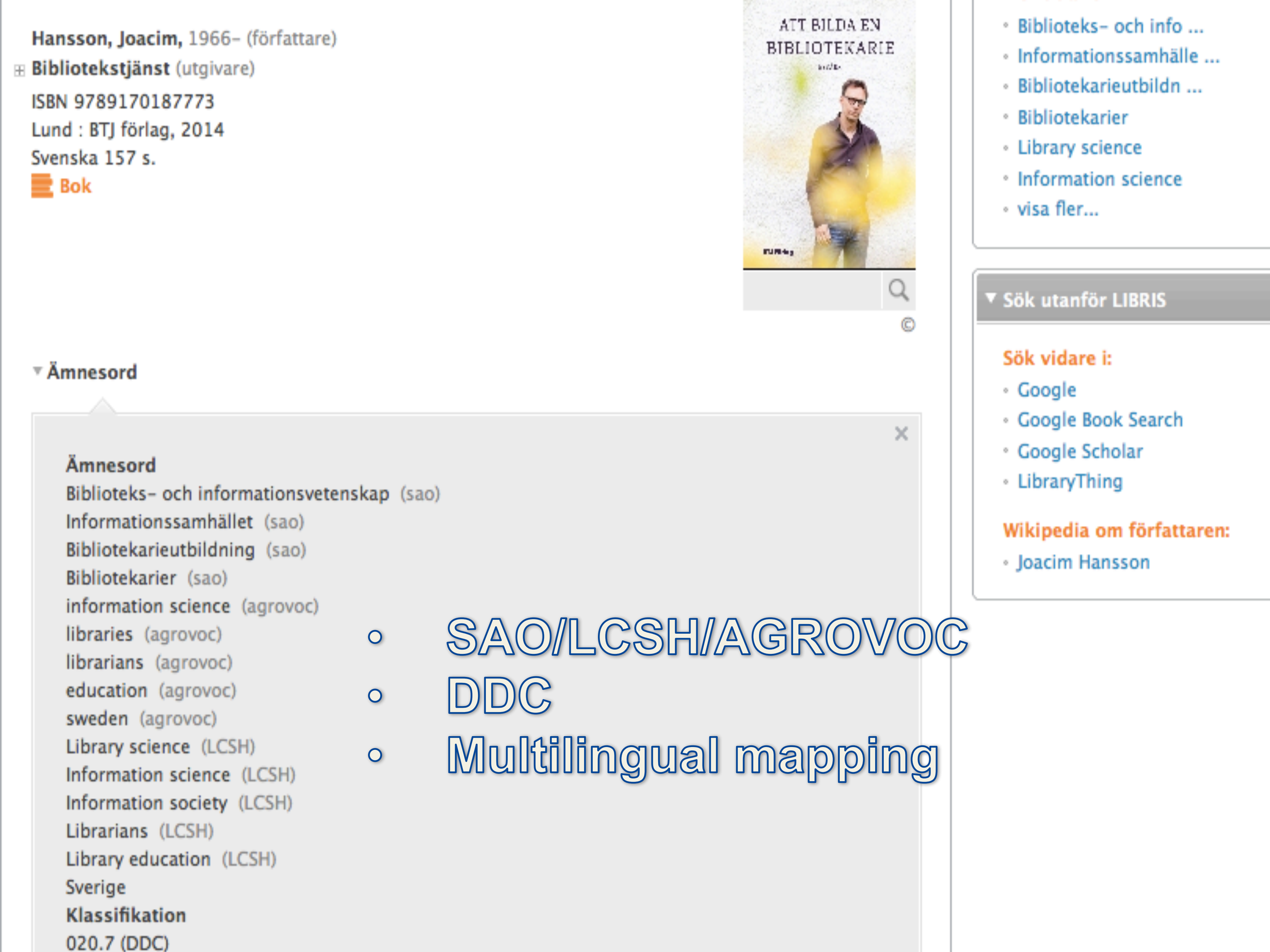
e.g., searching for ”bank” in AAT

Aims

- 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

- 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



Hansson, Joacim, 1966- (författare)

⊞ **Bibliotekstjänst** (utgivare)

ISBN 9789170187773

Lund : BTJ förlag, 2014

Svenska 157 s.

Bok

©

▾ **Ämnesord**

×

Ämnesord

Biblioteks- och informationsvetenskap (sao)

Informationssamhället (sao)

Bibliotekarieutbildning (sao)

Bibliotekarier (sao)

information science (agrovoc)

libraries (agrovoc)

librarians (agrovoc)

education (agrovoc)

sweden (agrovoc)

Library science (LCSH)

Information science (LCSH)

Information society (LCSH)

Librarians (LCSH)

Library education (LCSH)

Sverige

Klassifikation

020.7 (DDC)

- SAO/LCSH/AGROVOC
- DDC
- Multilingual mapping

- Biblioteks- och info ...
- Informationssamhälle ...
- Bibliotekarieutbildn ...
- Bibliotekarier
- Library science
- Information science
- visa fler...

▾ **Sök utanför LIBRIS**

Sök vidare i:

- Google
- Google Book Search
- Google Scholar
- LibraryThing

Wikipedia om författaren:

- Joacim Hansson

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, polisemy 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)

End-user indexing from KOS

- EnTag project http://www.ukoln.ac.uk/projects/enhanced-tagging/
- Findings:
  - 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
  - – 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, K., Lykke, M. & Tudhope, D. (2014). Enhancing social tagging with automated keywords from the Dewey Decimal Classification. Journal of Documentation, 70(5), 801-828.

Methodology

Methodology …

- Data collection: a subset of 111,000 Swedish-language documents from SwePub, SMDB, SND
- Development of a demonstrator information retrieval system

- A comprehensive evaluation approach involving 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, as well as 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 applications 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

- Domain analysis of two selected major areas within the DDC (humanities and applied sciences)

Automated indexers

- To be built/adjusted for DDC and Swedish
  - String matching
  - Machine learning
  - Keyword extractor (KEA)
  - 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 used in over 135 countries and available as Linked Data)

- The resulting empirically tested comprehensive methodology framework may be of interest to (inter)national researchers

Partners

- LNU
- LNUB
- KB
- SND
- Lund University
- University of South Wales
- University of Buffalo

- University of Aalborg
- Charles Sturt University
- Data Harmony Inc.
- OCLC

References

Borlund, P. (2003). The concept of relevance in IR. *Journal of the American Society for Information Science and Technology 54(10)*, 913-925.

Hjørland, B. (2002). Epistemology and the socio-cognitive perspective in information science. *Journal of the American Society for Information Science and Technology 53(4),* 257-270.

Lancaster, F. W. (2003). Indexing and abstracting in theory and practice. 3rd ed. Champaign: University of Illinois.

Saracevic, T. (2008). Effects of inconsistent relevance judgments on information retrieval test results: A historical perspective. *Library Trends 56(4)*, 763-783.

Soergel, D. (1994). Indexing and retrieval performance: The logical evidence. *Journal of the American Society for Information Science 45(8),* 589-599.