Evaluating automatic subject indexing: A framework

Keynote speech at 7th ISKO Italy Meeting
Bologna, 20 April 2015

Koraljka Golub
Linnaeus University, Sweden

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

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

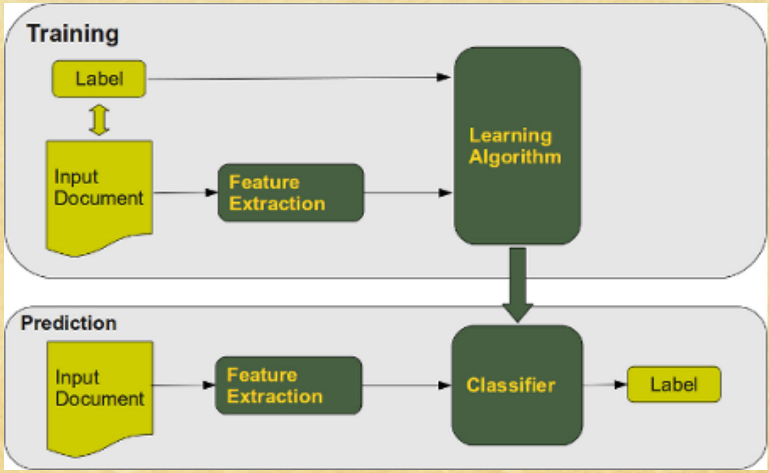
    → to allow retrieval from various perspectives
  - Subject classification: typically 1 precombined class

    → mostly for browsing

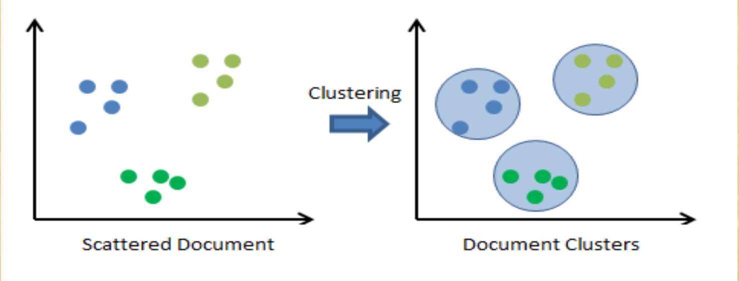
- Automatic/automated indexing/classification
  - A variety of terms in literature, also prevalent:
    - Text categorization
    - Document clustering

Automatic indexing

- 3 major approaches
  - Text categorization



  - Document clustering



  - String matching

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

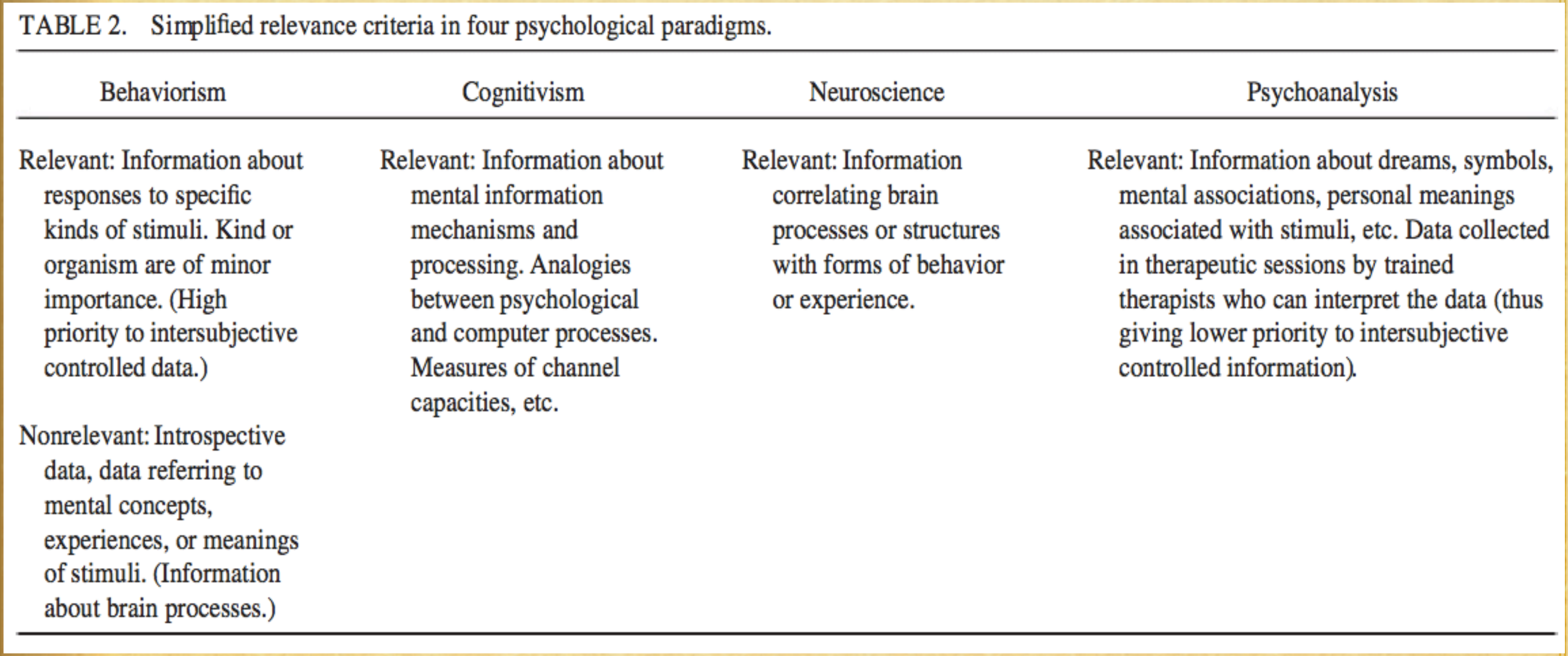


TABLE 2. Simplified relevance criteria in four psychological paradigms.

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.)	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).
Nonrelevant: Introspective data, data referring to mental concepts, experiences, or meanings of stimuli. (Information about brain processes.)			

- the relevance criteria of, for example, behaviorism, cognitivism, psychoanalysis, and neuro-science are very different even when they work on the same problem (e.g., schizophrenia) (Hjørland 2002, p. 263)

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
        - Determining the subject content of a document
        - A conceptual analysis to decide which aspects of the content should be represented
        - 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)

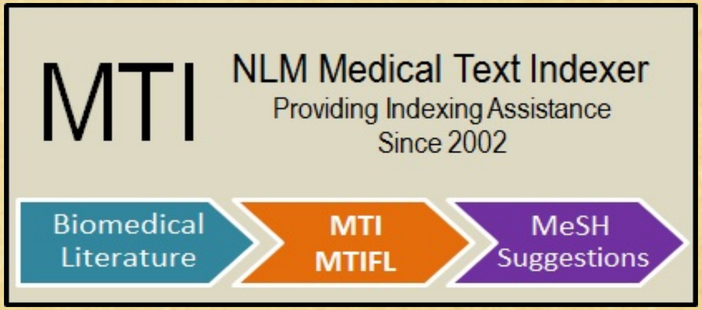
- Assign index terms using all indexing methods to be evaluated (for example, several automatic indexing systems to be evaluated and compared)
- Prepare document records that include all index terms assigned by any method in one integrated listing
- 2 senior professional subject indexers and preferably 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



- 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

Source and funding

- Golub, K., Soergel, D., Buchanan, G., Tudhope, D., Hiom, D., & Lykke, M. (2015). A framework for evaluating automatic indexing or classification in the context of retrieval. *Under revision for Journal of the Association for Information Science and Technology*
- Resulting from a JISC UK project EASTER
  - Evaluating Automated Subject Tools for Enhancing Retrieval
  - JISC Information Environment Programme 2009-2011
  - http://www.ukoln.ac.uk/projects/easter/

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.