Hybrid approach combining statistical and rule-based models for the automated indexing of bibliographic metadata in the area of planning and building construction

Dimitri Busch

Fraunhofer Information Center for Planning and Building IRB, Stuttgart, Germany

ECDA / LIS 2017,
September 27th, 2017

Fraunhofer IRB

Motivation

- ICONDA®Bibliographic (International Construction Database) is a bibliographic database which contains English-language documents (metadata entries) in the area of planning and building construction.
- The documents are indexed with descriptors from controlled vocabularies (FINDEX thesauri, an authority list).
- Until recently, indexing was done manually. The manual assignment of the descriptors was time-consuming and expensive.
- The presentation deals with an automated (semi-automatic) indexing system that was developed to solve the above problems.

Terminology used

- **Document:** metadata entry.
- **Descriptor:** prefered term, category, keyword.
- **Indexing**: assignment of descriptors to documents.
- **Categorizer:** computer program for indexing.

ICONDA: Sample document

Original Title	Advanced thermal insulation technologies in the built environment
Author	Livesey, Katie
Abstract	Reviews thermal insulation products, with a focus on advanced thermal insulation technologies such as aerogels, vacuum insulated panels, gas-filled panels and phase change materials.
Keyword	heat; insulation; efficiency; evaluation; insulating materials; materials; heat transmission; analysis
Publication year	2013
Language	English
Publication type	Journal article
Source	BRE information paper (2013), no.4/13, p.1-16

Controlled vocabularies

- **FINDEX Bau**; Type: faceted thesaurus; Author: Fraunhofer IRB; Subject: architecture and construction engineering; apx. 6500 terms; Relations: equivalence and hierarchy; bilingual: German und English.
- **FINDEX Raum**; Type: faceted thesaurus; Author: Fraunhofer IRB; Subjects: spatial planning, urban development, housing; apx. 2300 terms; Relations: equivalence, association and hierarchy; bilingual: German und English.
- **IRB Keyword List;** Type: authority list; Author: Fraunhofer IRB; Subjects: architecture, construction engineering, spatial planning etc.; apx. 56000 terms (all terms from FINDEX thesauri, other terms); Relations: not supported; bilingual: German und English.

Statistical categorizer

- The Fraunhofer IRB has developed a program for semi-automatic indexing of ICONDA documents.The program was put into application in the Feb. 2017.
- The program is based on the vector space (statistic) model.
- The descriptors are represented by profiles that are generated from training documents. [Pouliquen/Steinberger/Ignat 2003]
- In order to index a document, the program calculates similarities between the document and the descriptor profiles and proposes a ranked list of the descriptors to a human indexer.
- The indexer selects relevant descriptors and assigns the descriptors to the document.

Statistical categorizer: sample profile for the descriptor „Historic Building“

Term	Weight
historic	0.243
monument	0.205
baroque	0.198
palace	0.185
church	0.163
castle	0.152
restore	0.150
historical	0.148
…	…

Statistical categorizer: sample profile for the descriptor „Waste Prevention“

Term	Weight
waste	0.264
prevention	0.248
ludwigsburg	0.229
garbage	0.181
household	0.165
…	…

Statistical categorizer: Document representation

Sample Document

Title:

Ludwigsburg Residential Palace

Abstract:

The article deals with baroque architecture on the example of Ludwigsburg Residential Palace

Document representation

Term	Frequency
ludwigsburg	2
residential	2
palace	2
article	1
deal	1
baroque	1
architecture	1
example	1

Example of statistical indexing

Document

Title:

Ludwigsburg Residential Palace

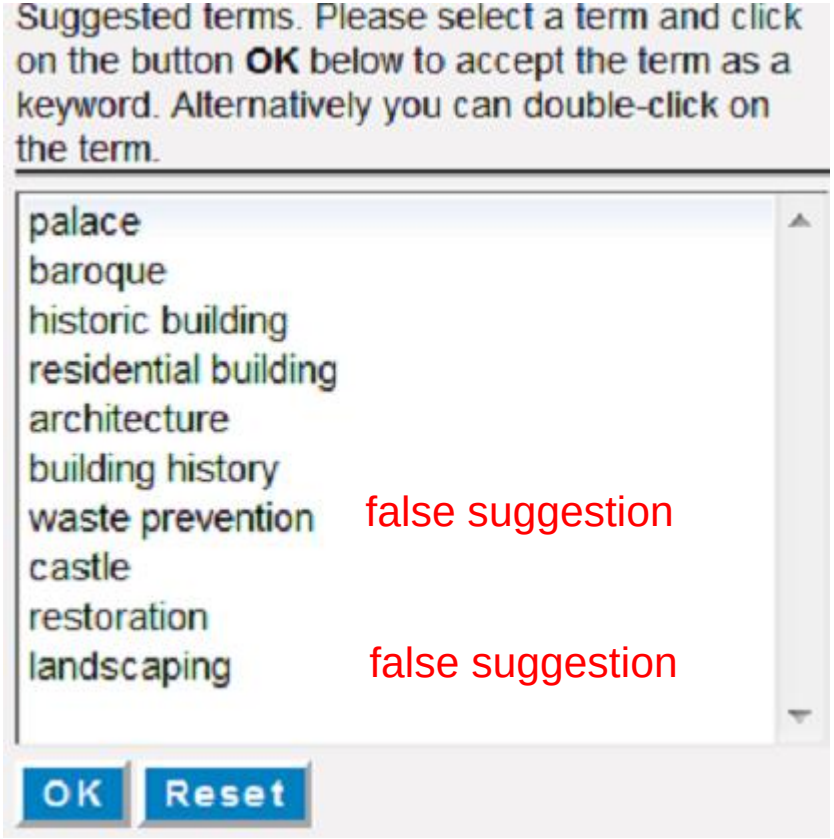
Abstract:

The article deals with baroque architecture on the example of Ludwigsburg Residential Palace

Assigned descriptors:

palace;baroque; historic building; architecture; building history

Suggested descriptors



Problems with statistical categorization

- The program makes credible suggestions, but it also makes false suggestions. Some evident descriptors which are clearly derivable from the document content, are not suggested.
- Solution: A rule-based categorizer can be used to complement and to correct the results of the statistical categorizer.

Indexing Rules

1. Sufficient rules

Example:

baroque->architectural style

If the document contains the word “baroque” then the descriptor “architectural style” will be assigned to the document

2. Necessary rules

Example:

waste or refuse <- waste prevention

The descriptor “waste prevention” can only be assigned if the document contains the word “waste” or the word “refuse”.

Ways to create rules

- Manual rule creation.
- Rule creation from thesaurus entries.
- Rule induction from training documents.

Example: Application of manually created rules to results of the statistical indexing

Document

Title:

Ludwigsburg Residential Palace

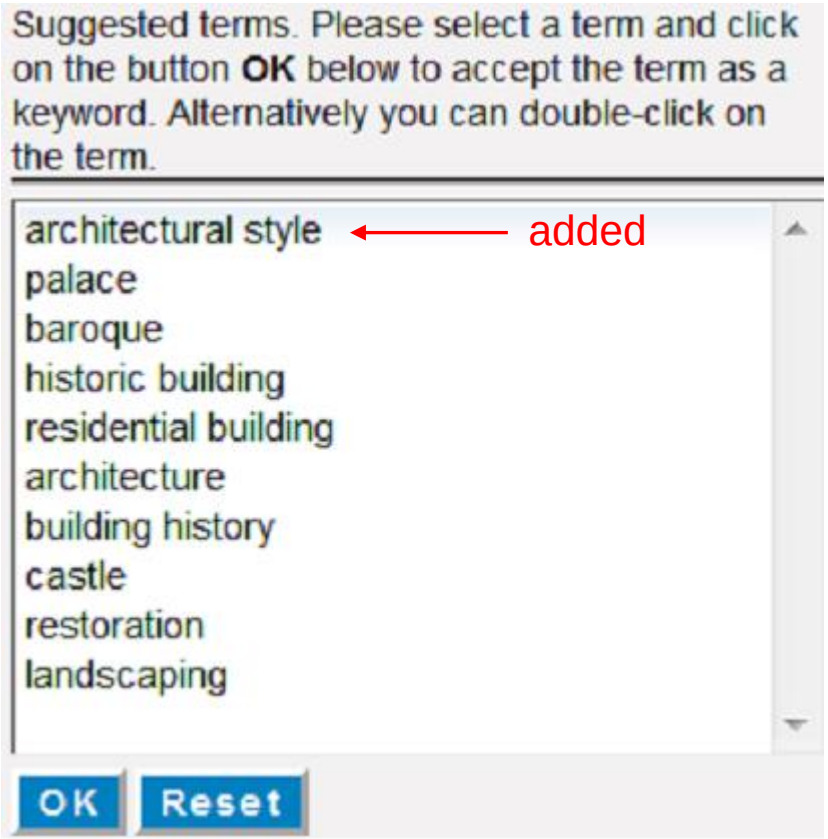
Abstract:

The article deals with baroque architecture on the example of Ludwigsburg Residential Palace

Assigned descriptors

architectural style; palace;baroque; historic building; architecture; building history

Suggested descriptors



„Waste Prevention“ was deleted.

Automated conversion of thesaurus terms into indexing rules

The most terms in the FINDEX Bau/ Raum can be converted into rules. Only descriptors and non-descriptors are used. Training data are not required.

Example:

Entry in FINDEX Bau

Descriptor	Relation	Non-descriptor (synonym)
refuse disposal	BF (use for)	waste disposal

Rules generated:

refuse and disposal->refuse disposal

waste and disposal->refuse disposal

Apx. 8000 sufficcient rules can be generated from terms in FINDEX thesauri. Terms in the IRB Keyword List can be also converted into rules in a similar way. The rules can be manually corrected with an editing program.

Induction of necessary rules

- Application of the statistical categorizer to its training set.
- Detection of descriptors that was incorrecty proposed and creation of a new separate training set for each such descriptor. The new training set contains true positive und false positive examples (documents) for the descriptor.
- Rule induction for the descriptors: IREP/ RIPPER Algorithm. [Cohen 1995]
- Examples of rules:

  ultrasonic<-ultrasound

  ultrasound<-ultrasound

Evaluation of a prototype hybrid categorizer with necessary rules that were induced automatically

- 21 descriptors from FINDEX Bau.
- Primary training set: 1365 documents.
- Test set: apx. 450 documents.

Evaluation with Rank=3

Evaluation measure	Statistical indexing	Hybrid indexing
precision	apx. 0.30	apx. 0.46
recall	apx. 0.67	apx. 0.50
f-score	apx. 0.42	apx. 0.48

Evaluation of hybrid categorization with induced rules

Evaluation with Rank=5

Evaluation measure	Statistical indexing	Hybrid indexing
precision	apx. 0.21	apx. 0.35
recall	apx. 0.79	apx. 0.47
f-score	apx. 0.33	apx. 0.40

Evaluation with Rank=21

Evaluation measure	Statistical indexing	Hybrid indexing
precision	apx. 0.09	apx. 0.13
recall	apx. 1.0	apx. 0.55
f-score	apx. 0.17	apx. 0.21

The Precision and the F-Score increase if the hybrid approach is used.

Implementation

- Java
- Microsoft SQL Server
- JRC JEX API [Steinberger/Ebrahim/Turchi 2012]
- WEKA / JRip

Conclusion

- The statistical categorizer is quite applicable for the interactive semi-automatic indexing of bibliographic metadata.
- A rule-based categorizer can be used to complement and to correct the results of the statistical categorizer.
- The hybrid approach is particularly useful when a descriptor can not be successfully trained by the statistical categorizer.
- Similar approaches can be probably used for the indexing of German-language documents.

References

- Cohen, William (1995). Fast Effective Rule Induction. In Proceedings of the Twelfth International Conference on International Conference on Machine Learning (ICML'95), Tahoe City, July 9-12, 1995. San Francisco: Morgan Kaufmann, p.115-123
- Heß, Andreas, Philipp Dopichaj and Christian Maaß (2008). Multi-Value Classification of Very Short Texts. In 31st Annual German Conference on Artificial Intelligence (KI 2008), Kaiserslautern, September 26-28, 2008. Berlin: Springer, p.70 ff.
- Pouliquen, Bruno, Ralf Steinberger and Camelia Ignat (2003). Automatic Annotation of Multilingual Text Collections with a Conceptual Thesaurus. In Proceedings of the Workshop in Ontologies and Information Extraction (EUROLAN2003), Bucharest, July 28- August 8, 2003, p.9-28
- Steinberger, Ralf, Mohamed Ebrahim and Marco Turchi (2012). JRC EuroVoc Indexer JEX – A freely available multi-label categorisation tool. In Proceedings of the 8th international conference on Language Resources and Evaluation (LREC'2012), Istambul, May 21-27, 2012, p.798-805
- Villena-Román, Julio, Sonia Collada-Pérez, Sara Lana-Serrano et al. (2011). Hybrid Approach Combining Machine Learning and a Rule-Based Expert System for Text Categorization. In: Proceedings of the Twenty-Fourth International Florida Artificial Intelligence Research Society Conference. Palm Beach, May 18-20, 2011, p.323 ff.

Thank your for listening