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

Abstract ICONDA® Bibliographic (International Construction Database) is a bibliographic database, which contains English-language documents in the area of planning and building construction. The documents are indexed with descriptors from controlled vocabularies (FINDEX thesauri, an authority list). The manual assignment of the descriptors is time-consuming and expensive. To solve this problem, an automated indexing system was developed. The indexing system combines a statistical classifier that is based on the vector space model with a rule-based classifier. In the statistical classifier, descriptor profiles are automatically trained from already indexed documents. The results provided by the statistical classifier will be improved with the rule based classifier that filters incorrect and adds missing descriptors. The rules can be created manually or automatically from already indexed documents. The hybrid approach is particularly useful when a descriptor cannot be successfully trained by the statistical classifier. In this case, the system can be easily fine-tuned by adding specific rules for the descriptor.

Dimitri Busch
Fraunhofer Information Center for Planning and Building IRB
Nobelstrasse 12, 70569 Stuttgart, Germany
✉ dimitri.busch@irb.fraunhofer.de

Archives of Data Science, Series A
(Online First)
KIT Scientific Publishing
Vol. 4, No. 1, 2018

DOI: 10.5445/KSP/1000085951/15
ISSN 2363-9881

1 Introduction

ICONDA® Bibliographic (International Construction Database) is a bibliographic database which contains English-language documents in the area of planning and building construction. The Fraunhofer IRB (Fraunhofer Information Center for Planning and Building (IRB), Stuttgart, Germany) produces this database and provides it online on `http://www.irb.fraunhofer.de/iconda` (accessed: 26.05.2019). The documents are indexed with descriptors from controlled vocabularies (FINDEX thesauri, an authority list). Table 1 shows a sample document from ICONDA. Until recently, indexing was done manually. The manual assignment of the descriptors was time-consuming and expensive.

Table 1: ICONDA: Sample Document.

Original Title	Resiliency for the City and the Sea. Examples of coastal urbanism: New York and Puerto Rico
Author	Bolstad, Jennifer; Meyer, Walter
Abstract	Coastal Urbanism has to face changing climate.Two examples, from Puerto Rico and New York, show approaches that use living systems and natural landforms
Keywords	town development; planting; sample presentation; project description; flood control; tornado; bank improvement; park; cost data
Publication year	2014
Language	English
Publication type	Journal article
Source	Topos (2014), no.87, p. 64–69

This paper deals with an automated (semi-automatic) indexing system that was developed to solve the problem mentioned above. Here are some key terms that are used in this paper:

- **Document:** Metadata entry;
- **Descriptor:** Preferred term, category, keyword;
- **Indexing**: Assignment of descriptors to documents, categorization;
- **Classifier:** Computer program for indexing.

2 The State of Science and Technology

Indexing with a controlled dictionary can be considered as text categorization (De Campos et al, 2009). There are two main approaches to text categorization-knowledge engineering and machine learning (Sebastiani, 2002). These two approaches are considered more closely below.

2.1 Knowledge Engineering

The knowledge engineering approach to text categorization consists in manual building an expert system capable to make text categorization decisions. The expert system consists of a knowledge base created by a knowledge engineer, i.e. a human expert in knowledge representation. To represent knowledge in such knowledge bases, rules are usually used. The rules can be represented like this:

Rule 1. *IF <premise is true> THEN <assign descriptor>,*

where the premise is a Boolean expression that consists of terms that can be connected with logic operations like AND, OR and NOT. If a document meets the premise of the rule, the descriptor can be assigned to the document.

CONSTRUE, a classifier with manually prepared rules, can be mentioned as an example of the knowledge engineering approach. This classifier was reported to achieve the recall of 94 percent and the precision of 84 percent, i.e.

a performance comparable to a performance of a human indexer. (Hayes and Weinstein, 1990)

The main drawback of the knowledge engineering approach is a high expenditure of time and thus high cost for rule development. For example, it took approximately 2.5 person-years to develop rules for the CONSTRUE system that uses 674 categories (Hayes and Weinstein, 1990).

2.2 Machine Learning

Classifiers that support machine learning generate their knowledge bases automatically from data already indexed (training sets). There are a lot of approaches to learning classifiers, e.g. probabilistic, econometric, profile-based, neural, example based, decision trees and decision rules (Sebastiani, 2002). This paper deals mainly with linear profile-based and inductive decision rule classifiers. These two techniques are considered more closely below.

2.2.1 Linear Profile-Based Classifiers

In the linear profile-based classifier, each descriptor C_i is represented by a vector $\vec{C_i} = (w_{i1}, w_{i2}, \cdots, w_{it})$ belonging to $|t|$ -dimensional vector space, where w_{ik} denotes the weight of a term k associated with the descriptor C_i. This vector is also called descriptor profile. Each document D_j to be indexed is also represented by a vector $\vec{D_j} = \left(w_{j1}, w_{j2}, \cdots, w_{ij}\right)$, where w_{jk} denotes the weight of the term k in the document D_j. During the indexing of a document, similarity values are calculated between descriptors and the document. The similarity between the descriptor C_i and the document D_j can be calculated e.g. as a dot product $\sum_{k=1}^{t}(w_{ik}w_{jk})$ of the vectors $\vec{C_i}$ and $\vec{D_j}$.

To build profile-based classifiers, i.e. to calculate profiles, various methods can be used. These methods can be divided into batch induction methods and online induction methods. Batch induction methods build a classifier by analysing the training set all at once (Sebastiani, 2002). Examples of the batch methods are the Rocchio algorithm (Rocchio, 1971), discriminant analysis (Blosseville et al, 1992) and cooccurrence based methods (Ferber, 1997; Pouliquen et al, 2003).

Online (incremental) methods build a classifier soon after evaluating the first training document, and incrementally refine it as they evaluate new ones (Sebastiani, 2002). Online methods are useful if the training set is not available in its entirety right from start (ibid.). Examples of online methods are the perceptron algorithm (Manning and Schütze, 1999, pp. 597-604), Winnow(Littlestone, 1988; Dagan et al, 1997), the Widrow-Hoff algorithm (Lewis et al, 1996).

2.2.2 Inductive Decision Rule Classifiers

A rule learning system constructs one or more rules of the form (Fürnkranz et al, 2014, p. 25):

Rule 2. *IF* f_1 *AND* f_2 *AND* f_L *AND THEN assign descriptor* C_i*,*

where the conditional part of the rule is a logical conjunction of features (also called literals). In the case of text categorization, a feature f_k in the premise denotes the presence of a term k in the test document D_j, while the consequence denotes the decision to assign the descriptor C_i. The automatic construction of such rules is typically done by heuristically searching for the conjunction of features that is more predictive for this descriptor. Individual rule learners vary widely in terms of the methods, heuristics etc. Among the rule learners that have been applied to text categorization are e.g. RIPPER (Cohen, 1995) and SWAP-1 (Apte et al, 1994).

2.3 Hybrid Classifiers

The main advantage of machine learning consists in the low time consumption of generating the classifiers. Most learning classifiers are based on mathematical (statistical) models. Such models are not always understandable by humans, so it is difficult to diagnose the reason for the false positives/negatives and to fine-tune the classifiers (Villena-Román et al, 2011).
One way to solve the latter problem is to construct a hybrid classifier in which the results provided by the first classifier are improved using a second classifier.

For example, a categorization system of Villena-Román et al (2011) uses a rule-based classifier created by knowledge engineers to improve results provided by an example-based classifier (k-Nearest Neighbour). Another hybrid classifier of Hess et al (2008) uses an inductive decision rule classifier with numerical features to improve results provided by a linear profile-based Rocchio classifier.

3 Controlled Vocabularies

In this section, controlled vocabularies used for the indexing of ICONDA documents are presented.

3.1 FINDEX Bau

The Facet-Oriented Indexing System for Architecture and Construction Engineering (FINDEX Bau) was created by the Fraunhofer Information Centre for Planning and Building (Fraunhofer-IRB, 1985). The thesaurus consists of a systematic and an alphabetical part. The systematic part consists of four levels, the first of these contains 20 facets, e.g. CONSTRUCTION TYPE, EXECUTION OF CONSTRUCTION WORK, BUILDING USE. In the alphabetical part, a distinction is made between descriptors (preferred names) and non-descriptors (alternative names). Between non-descriptors and descriptors, the following equivalence relationships can be established:

- BD (benutze Deskriptor, German) - USE DESCRIPTOR, which leads from a non-descriptor to a descriptor;
- BF (benutzt für, German) - USED FOR, a reciprocal relationship, which leads from a descriptor to a non-descriptor.

Examples:

- heat insulation BD thermal insulation
- thermal insulation BF heat insulation,

where THERMAL INSULATION is a descriptor, and HEAT INSULATION is a non-descriptor.
The thesaurus contains about 6500 terms and is bilingual: German and English.

3.2 FINDEX Raum

The Facet-Oriented Indexing System for Regional Planning, Town Planning and Housing (FINDEX Raum) was created by the Fraunhofer Information Centre for Planning and Building and is structured similar to FINDEX Bau. The thesaurus consists of a systematic and an alphabetical part. The systematic part consists of four levels, the first of these contains 18 facets, e.g. SPACE AND SETTLEMENT, MUNICIPAL ADMINISTRATION, TECHNICAL INFRASTRUCTURE. The alphabetical part distinguishes between descriptors and non-descriptors. Between non-descriptors and descriptors, as in FINDEX Bau, equivalence relations BD (USE DESCRIPTOR) and BF (USED FOR) can be established (see also Subsection 3.1). Additionally, association relations SA (SEE ALSO; siehe auch, German) can be established between related descriptors.

Example:

- regional development research SA historic spatial research

The thesaurus contains about 2300 terms and is bilingual: German and English.

3.3 IRB Keyword List

The IRB Keyword List is an authority list created by the Fraunhofer IRB. The authority list contains about 38600 terms related to architecture, construction engineering, town planning etc (state of 07.2019). The authority list contains, among other things, all the terms of the thesauri FINDEX Bau and FINDEX Raum, but does not support relationships between the terms. The authority list is bilingual: German and English.

4 Semi-Automatic Indexing Using a Statistical Classifier

The Fraunhofer IRB has developed a computer program for semi-automatic indexing of ICONDA documents, in which automatically proposed descriptors are checked and assigned by human indexers. This program is considered more closely below.

4.1 Statistical Classifier

The program for semi-automatic indexing is a Java application that uses JEX, a linear profile-based classifier. JEX (JRC EuroVoc Indexer) is a free software developed at the Joint Research Centre (JRC) of the European Union (Steinberger et al, 2012). Although JEX was originally intended for indexing according to the EuroVoc thesaurus (EU, 2018), it can be adapted for other controlled dictionaries. The advantages of this classifier include an ability to perform high speed multi-label indexing with large controlled dictionaries. JEX supports an application programming interface (API) allowing it to be called from by other programs. Because JEX is based on a statistical (linear) model, is also referred to below as Statistical Classifier.

In statistical classifier, descriptors are represented by profiles that are generated from training documents. The profiles are generated according to a cooccurence-based algorithm (Pouliquen et al, 2003). Tables 2-3 show sample profiles for the descriptors HISTORIC BUILDING and WASTE PREVENTION. Each profile record consists of a term and its weight in relation to the descriptor. The terms were selected in the profile from training documents, to which the descriptor was assigned.

Table 2: Sample profile for the descriptor HISTORIC BUILDING.

Term	Weight
historic	0.243
monument	0.205
baroque	0.198
palace	0.185
church	0.163
...	...

Table 3: Sample profile for the descriptor WASTE PREVENTION.

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

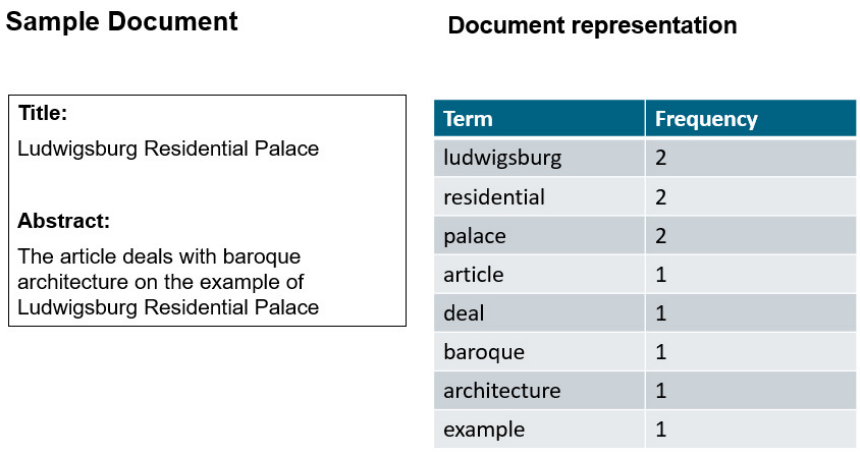

Figure 1: Statistical classifier: Document representation.

Figure 1 shows a sample document and its representation in the system. On the left side, fields analysed by the classifier are presented. On the right side, a representation of the document in the indexing system is shown. The document is represented by its terms and frequencies of the terms. For example, the term PALACE appears in the title and in the abstract and thus has a frequency of 2.

4.2 Semi-Automatic Indexing

In order to index a document, the indexing 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.

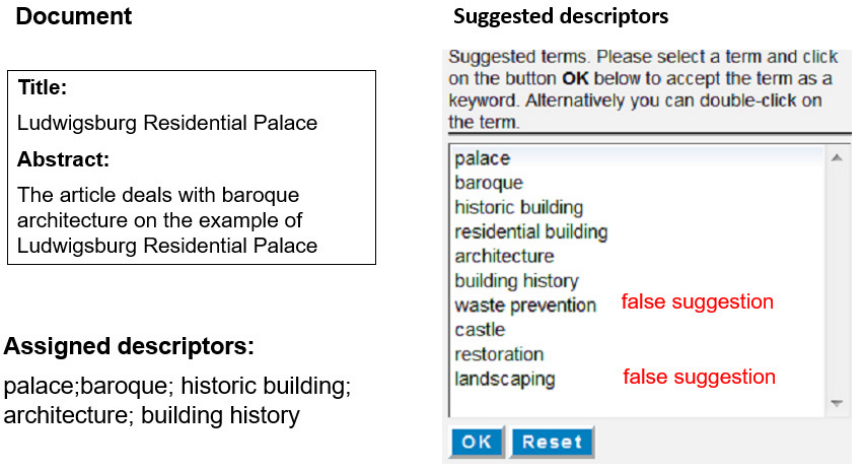

Figure 2: Example of semi-automatic indexing.

Figure 2 shows indexing of the document shown in Figure 1. In the right part of the figure, descriptors suggested by the indexing system are presented. There are both correct and incorrect suggestions. WASTE PREVENTION is e.g. incorrect. To clarify the cause of the incorrect proposed descriptor, the profile for the descriptor can be checked (Table 3).

As can be seen, the profile contains the incorrect term LUDWIGSBURG, i.e. the name of a city in Germany, that has no relation to the topic. Because the both the document and the profile contain LUDWIGSBURG, the descriptor WASTE PREVENTION was proposed for the document. Such incorrect profile entries can cause incorrect indexing, but they are unavoidable if the profiles are generated automatically. The example (Figure 2) also shows that some evident descriptors, e.g. ARCHITECTURAL STYLE, which are clearly derivable from the document content, are not suggested. One solution of the above mentioned problems is the hybrid indexing that will be presented in Section 5.

4.3 Evaluation of the Statistical Classifier

In order to test the statistical classifier, two evaluation types were applied: The evaluation through comparison with a gold standard, i.e. descriptors previously manually assigned by indexers, and the manual evaluation. In the last case, descriptors automatically suggested were then rated manually. Each proposed descriptor was assigned one of the following grades: GOOD, RATHER GOOD, IRRELEVANT, FALSE. The evaluation types mentioned above, i.e. the comparison with the gold standard and the manual rating, were used because some studies show that their results may be different (Pouliquen et al, 2003). Because of high time consumption, the manual evaluation has only been done for a small test set of 20 documents already indexed. The automatic indexing was performed with the ranking of 20, i.e. up to 20 descriptors were proposed for each test document. The following values were calculated for the following evaluation measures: Precision, Recall and F-Score. For the manual evaluation, GOOD and RATHER GOOD proposals were taken into account, and all descriptors already assigned were rated as GOOD. The results of the evaluations showed that the manual rating gives substantially better results (Table 4). One reason for this is that gold standards often contain only descriptors that have been considered best by indexing experts. Such decisions are often made subjectively (see Chen, 2008).

Table 4: Evaluation of the statistical classifier: 20 documents, Rank=20.

Evaluation measure	Evaluation through direct assessment by evaluator	Evaluation through comparison with a gold standard
Precision	0.46	0.15
Recall	0.55	0.27
F-score	0.50	0.19

5 Prototype Hybrid Classifiers

The example of semi-automatic indexing (Figure 2, see also Subsection 4.2) shows that the statistical classifier can suggest so-called false positive (incorrect) descriptors. The example of also shows that some evident descriptors which are clearly derivable from the document content, are not suggested. In order to solve these problems, the Fraunhofer IRB has developed prototype hybrid classifiers that use categorization rules to complement and to correct the results of the statistical classifier. This prototype classifiers are considered more closely below.

5.1 Categorization Rules

The hybrid classifiers support two rule types: Sufficient rules and necessary rules. All rule types consist of two parts. The left part of the rule contains one or more terms that can be connected with Boolean operators. The right part of a rule consists of a descriptor. Sufficient rules are used to assign descriptors to a document. If the expression in the left part of the rule is applicable to the document the descriptor can be assigned to the document. For example, Rule 3 says that if the document contains the word BAROQUE then the descriptor ARCHITECTURAL STYLE can be assigned to the document. Necessary rules are used to prove descriptors proposed by the classifier. For example, Rule 4 says that the descriptor WASTE PREVENTION can only be assigned if the document contains the word WASTE or the word REFUSE.

Rule 3. *baroque → architectural style*

Rule 4. *waste or refuse ← waste prevention*

Rules can be created in different ways. They can be created manually by human experts. They can be generated automatically from documents already indexed.

5.2 Use of manually created Rules

To create and try rules, a prototype has been developed by extending the program for semi-automatic indexing (see Subsection 4.2) with a simple rule editor and a rule-based classifier. Figure 3 shows an application of rules created manually (Rule 3 and Rule 4) to results of the statistical indexing from Figure 2. The first rule caused the addition of the descriptor to the suggestion list. The second rule caused deletion of the incorrect descriptor WASTE PREVENTION from the suggestion list.

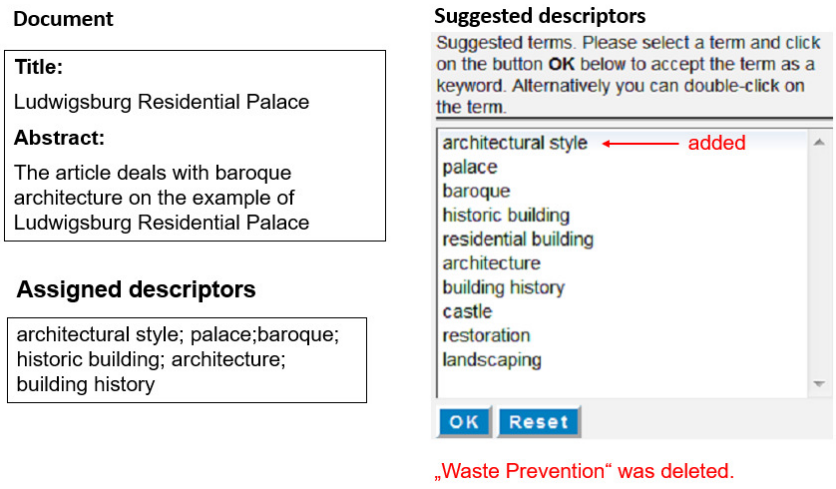

Figure 3: Application of manually created rules to results of the statistical indexing.

5.3 Induction of Necessary Rules

To induce and evaluate rules, another prototype has been developed. The prototype is a Java application that is running in batch mode. The application uses JEX (see Subsection 4.1) for statistical indexing and the JRip classifier for rule induction. JRip implements the RIPPER algorithm (Cohen, 1995) and is a component of WEKA, an open source data mining system (Witten et al, 2011).

The prototype induces necessary rules for descriptors that have not been successfully trained. The rule induction is done in the following way: First, the statistical classifier is applied to the training set. Descriptors that were incorrectly suggested are determined and a new separate training set is created for each descriptor. This new training set contains true positive and false positive examples (documents) for the descriptors and is used for rule induction.

The introduced approach is similar to the hybrid classifier of Hess et al (2008), but conditional parts of rules contain terms rather than numeric features. For example, the following rule can be induced:

Rule 5. *ultrasonic* ← *ultrasound*

5.4 Evaluation of the Hybrid Classifier that Supports Rule Induction

The hybrid classifier that supports rule induction (see Subsection 5.3) was evaluated with the following data:

- 21 descriptors from FINDEX Bau;
- Primary training set: 1365 documents;
- Test set: 458 documents.

Tables 5-6 show evaluation results for the ranking of 3 and 5. The evaluation values show that the use of the hybrid method leads to an increase in the precision and the F-score compared to the statistical method.

Table 5: Evaluation of the hybrid classifier with Rank=3.

Evaluation measure	Statistical indexing	Hybrid indexing
Precision	0.30	0.46
Recall	0.67	0.50
F-score	0.42	0.48

Table 6: Evaluation of the hybrid classifier with Rank=5.

Evaluation measure	Statistical indexing	Hybrid indexing
Precision	0.21	0.35
Recall	0.79	0.47
F-score	0.33	0.40

6 Conclusion

The Fraunhofer IRB has integrated a statistical classifier in a computer program for semi-automatic indexing of ICONDA documents, in which automatically proposed descriptors are checked and assigned by the human indexers. The program was evaluated through direct assessment by indexers and put into application in 2017. Currently it is used mainly for the indexing of CIB [1] conference papers.

The Fraunhofer IRB has also developed prototype hybrid classifiers that use a rule-based classifier to complement and to correct the results of the statistical classifier. The rules can be created manually or induced automatically. An evaluation of the hybrid classifier that support rule induction shows that the use of the hybrid method leads to an increase in performance (precision and the F-score) compared to the statistical method. The hybrid approach is particularly useful when a descriptor can not be successfully trained by the statistical classifier. Similar approaches can probably be used for the indexing of German-language documents.

References

Apte C, Damerau F, Weiss S (1994) Automated Learning of Decision Rules for Text Categorization. ACM Transactions on Information Systems 12(3):233–251, New York. DOI: 10.1145/183422.183423.

Blosseville M, Hébrail G, Monteil M, Pénot N (1992) Automatic Document Classification: Language Processing, statistcal Analysis, and Expert System Techniques used together. In: Proceedings of the 15th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Copenhagen, June 21 - 24, 1992, Fox E, Belkin N, Ingwersen P, Pejtersen A (eds), ACM Press, New York, pp. 51–58. DOI: 10.1145/133160.133175.

Chen X (2008) Indexing Consistency between Online Catalogues. PhD thesis, Berlin, Humboldt University of Berlin. DOI: 10.18452/15777.

Cohen W (1995) Fast Effective Rule Induction. In: Proceedings of the Twelfth International Conference on International Conference on Machine Learning (ICML'95), Prieditis A, Russel S (eds), Morgan Kaufmann Publishers, Tahoe City, pp. 115–123. DOI: 10.1016/B978-1-55860-377-6.50023-2.

Dagan I, Karov Y, Roth D (1997) Mistake-Driven Learning in Text Categorization. In: Proceedings of the Second Conference on Empirical Methods in Natural Language Processing, Providence, August 1-2, 1997, Cardie C, Weischedel R (eds), ACL, Somerset, NJ, pp. 55–63.

De Campos L, Fernandez-Luna J, Huete J, Romero A (2009) Handbook of Research of Text and Web Mining Technologies, Vol. 1, IGI Global, Hershey (NY), New Jersey, chap. Thesaurus-Based Automatic Indexingpp. 331–345. ISBN: 978-1-599049-90-8, DOI: 10.4018/978-1-59904-990-8.

EU (2018) EuroVoc, the EU's multilingual thesaurus. URL: `https://data.europa.eu/euodp/en/data/dataset/eurovoc`.

Ferber R (1997) Automated indexing with thesaurus descriptors: A co-occurrence based approach to multilingual retrieval. In: Research and Advanced Technology for Digital Libraries: First European Conference, ECDL'97 Pisa, Italy, September 1–3, 1997 Proceedings, Peters C, Thanos C (eds), Springer, Berlin. ISBN: 978-3-540635-54-3, DOI: 10.1007/BFb0026731.

Fraunhofer-IRB (ed) (1985) FINDEX. Facet-Oriented Indexing System for Architecture and Construction Engineering. IRB Verlag, Stuttgart. ISBN: 978-3-816705-89-5.

Fürnkranz J, Gamberger D, Lavrac N (2014) Foundations of Rule Learning. Springer, Heidelberg. ISBN: 978-3-540751-97-7, ISSN: 1611-2482, DOI: 10.1007/978-3-540-75197-7.

Hayes P, Weinstein S (1990) CONSTRUE/TIS: A System for Content-Based Indexing of a Database of News Stories. In: Innovative Applications of Artificial Intelligence 2. The Second Conference on Innovative Applications of Artificial Intelligence, Rappaport A, Smith R (eds), AAAI Press, Menlo Park, pp. 49–64.

Hess A, Dopichaj P, Maass C (2008) Multi-Value Classification of Very Short Texts. In: 31st Annual German Conference on Artificial Intelligence (KI 2008), Dengel A, Berns K, Breuel T, Bomarius F, Roth-Berghofer T (eds), Springer, Berlin, pp. 70–77. DOI: 10.1007/978-3-540-85845-4_9.

Lewis D, Schapire R, Callan J, Papka R (1996) Training Algorithms for Linear Text Classifiers. In: Proceedings of the 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Zurich, August 18-22, 1996, Frei H, Harman D, Schäuble P, Wilkinson R (eds), ACM, New York, NY, pp. 298–306. DOI: 10.1145/243199.243277.

Littlestone N (1988) Learning Quickly When Irrelevant Attributes Abound: A New Linear-Threshold Algorithm. Machine Learning 2(4):285–318. DOI: 10.1023/A:1022869011914.

Manning C, Schütze H (1999) Foundations of Statistical Natural Language Processing. The MIT Press, Cambridge, MA.

Pouliquen B, Steinberger R, Ignat C (2003) Automatic Annotation of Multilingual Text Collections with a Conceptual Thesaurus. In: Proceedings of the Workshop in Ontologies and Information Extraction (EUROLAN2003), Bucharest, pp. 9–28.

Rocchio J (1971) The SMART Retrieval System - Experiments in Automatic Document Processing, Prentice Hall, Upper Saddle River, NJ, chap. Relevance Feedback in Information Retrievalpp. 313–323.

Sebastiani F (2002) Machine Learning in Automated Text Categorization. ACM Computing Surveys 34(1):1–47, New York. DOI: 10.1145/505282.505283.

Steinberger R, Ebrahim M, Turchi M (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), European Language Resources Association (ELRA), Istanbul, pp. 798–805.

Villena-Román J, Collada-Pérez S, Serrano S, González-Cristóbal J (2011) Hybrid Approach Combining Machine Learning and a Rule-Based Expert System for Text Categorization. In: Proceedings of the 24th International Florida Artificial Intelligence Research Society, FLAIRS- 24. Palm Beach, Florida. May 18-20, 2011, Murray RC, McCarthy P (eds), The AAAI Press, Menlo Park(CA), Palm Beach, pp. 323–328.

Witten I, Frank E, Hall M (2011) Data Mining: Practical Machine Learning Tools and Techniques, 3rd edn. Morgan Kaufmann, Amsterdam. DOI: 10.1016/C2009-0-19715-5.