Towards Semantic Quality Control of Automatic Subject Indexing

Martin Toepfer[1(✉)] and Christin Seifert[2]

[1] ZBW – Leibniz Information Centre for Economics,
Düsternbrooker Weg 120, 24105 Kiel, Germany
m.toepfer@zbw.eu
[2] University of Passau, Innstraße 33, 94032 Passau, Germany
christin.seifert@uni-passau.de

Abstract. Automatic subject indexing is a key technology for digital libraries, however, factors like concept drift hinder its success in practice. Releasing high-quality results into productive retrieval systems may still be possible when thorough quality control is applied, which may support algorithmic improvements and allow to create high precision filters. Since errors and their relevance can depend on characteristics of concepts and their relations, evaluations should take semantic aspects into account. For this reason, we present the prototype of a web-based reviewing tool which especially aims at fostering semantic analysis and visualization, that is, considering relations, properties and semantic categories of concepts, algorithms and reviews. The tool uses techniques of the Semantic Web. Its application is demonstrated by example.

Keywords: Quality control · Automatic subject indexing · Semantics

1 Introduction

Accurate indexing of documents with subject headings (descriptors, concepts) of controlled vocabularies enables high-quality semantic access to digital libraries. Automation of this task has been addressed by many researchers, for instance, in the field of machine learning and multi-label classification. In practice, different factors hinder the success of automatic methods, thus libraries apply them either only as assistants [1,3], or as autonomous agents restricted to special types of documents [7]. In particular if predictions are passed to productive retrieval systems without human intervention, continuous testing and control becomes crucial to ensure high-quality results over time. In this paper, we present a web-based application for reviewing automatically predicted subject headings. In order to recognize semantic patterns in errors, integration of background knowledge from thesauri is desirable. We build upon technology from the Semantic Web for data modelling to foster analysis and visualization of relations, properties and semantic categories of concepts, indexing approaches and ratings.

J. Kamps et al. (Eds.): TPDL 2017, LNCS 10450, pp. 616–619, 2017.
DOI: 10.1007/978-3-319-67008-9_56

2 Background

Put briefly, subject indexing aims to determine the most relevant subjects of documents comprehensively, precisely and concisely. Controlled vocabularies are used to reduce ambiguity and enable further semantic applications. In this work, we use the STW thesaurus 9.02[1] [2], which addresses economics and related subject areas. It has more than 6,000 concepts with links between broader (BT), narrower (NT), and semantically related (RT) concepts. Descriptors are additionally linked to semantic categories. Regarding automatic subject indexing, we assume that there is a main system under review, for instance, a fusion system [6] that combines lexical approaches, which use keyword matching, and associative approaches, which learn synonymous expressions from examples.

Common approaches for evaluation of automatic methods leverage corpora with documents that have already been indexed professionally and compare them with subjects predicted by algorithms. Several classification metrics, like precision, recall, and F1, or ranking metrics can be computed and different averaging techniques may be used, for instance, aggregation by concept or by instance. Beyond these evaluation approaches, subject-specific analysis and fine-grained ratings are used as well, for instance, at the German National Library [7].

Since cleansing and evaluation tasks are crucial but often costly parts of projects, general purpose tools like OpenRefine[2] and various specialized user interfaces for annotation and evaluation tasks have been developed in different domains, for instance, ontology alignment and object-vocabulary automatic linking, which also have to deal with fuzzy matching problems.

3 Reviewing Subject Headings

The main view for reviewing subject headings is depicted in Fig. 1. On the top, meta-data (title, author keywords, abstract) is shown 1 for determining relevant subjects manually. A table 2 summarizes the concepts (rows) that have been proposed by different indexing approaches (columns). Each concept can be rated individually 3. Missing concepts can be added 4. When finished reviewing the concepts, the reviewer enters a final decision for the document on a 3-point scale 5, which especially determines if the automatically generated descriptors must be rejected because the proposed subjects would be misleading. A graph visualization[3] depicts relations between proposed concepts (direct RT relations and paths of BT) and their semantic categories 6. Some decisions may be subtle, like disambiguation between Germany, Germans and German (language).

The tool especially targets the precision of automatic indexing, thus the most relevant role of concept-level ratings is to prevent misleading concepts (−) in the output. The other levels (0, +, ++) denote increasing preciseness and relevance. Detailed information on the rating guidelines can be accessed by a dialog 7.

[1] www.zbw.eu/en/stw-info/ (accessed: 10.04.2017).
[2] www.openrefine.org/ (accessed: 15.06.2017).
[3] The graph visualization is below the table in the user interface, but depicted next to it due to space constraints.

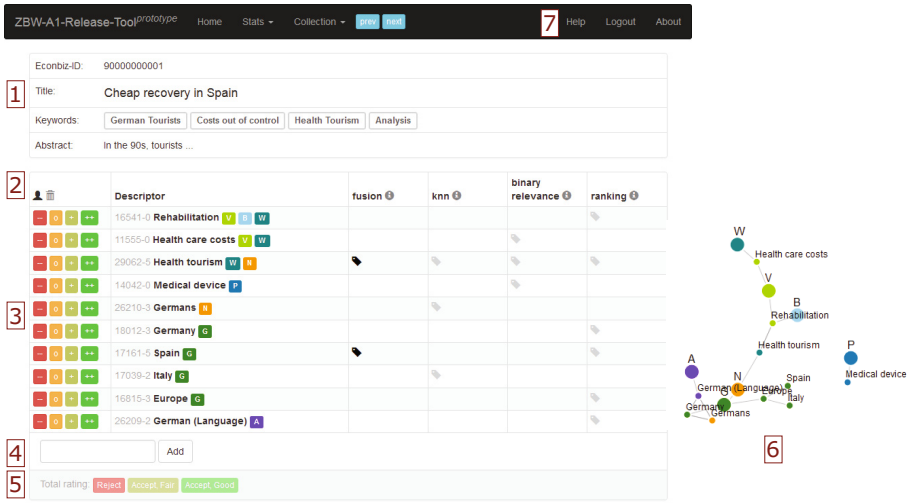

Fig. 1. User interface for reviewing subject headings.

4 Semantic Quality Control

In contrast to plain applications of precision, recall and F1, the tool proposed here aims at semantic quality control, that is, taking semantic categories and relations between concepts as specified in thesauri into account. In the context of measuring inter-indexer consistency, disregarding semantic relationships between concepts has been criticised, for instance, by Medelyan and Witten [5] who proposed a measure that incorporates RT and BT/NT relations. Some thesauri provide further structure beyond these relations, for instance, the STW (cf. Sect. 2). At the top level, it has seven semantic categories (row names in Fig. 2). In order to leverage this background knowledge, we build upon Semantic Web technology for data modeling which can be used by digital libraries to expose Linked Open Data [4]. We utilize well-known schemas: dublin core[4], SKOS[5], rev[6], and MUTO[7]. With this representation, queries on reviews can be formulated in SPARQL, accessing semantic properties and relations.

To illustrate a simple case of semantic quality control, Fig. 2 shows an analysis of artificial concept ratings by rating value, indexing algorithm (Agent) and semantic category. It can be seen that severe errors are imbalanced, e.g. fusion does not make any errors for geographic names. Such insights can help to improve the overall indexing system by weighting assignments for each category dependent on the algorithm. Also rules for filtering can be developed. For instance, geographic names proposed by knn may be blocked.

[4] dublincore.org/documents/dcmi-terms/ (accessed: 10.04.2017).
[5] www.w3.org/2004/02/skos/ (accessed: 10.04.2017).
[6] vocab.org/review/ (accessed: 10.04.2017).
[7] muto.socialtagging.org/core/v1.html (accessed: 10.04.2017).

	rating value	-				0				+				++			
	Agent	fusion	ranking	knn	binary relev.	fusion	ranking	knn	binary relev.	fusion	ranking	knn	binary relev.	fusion	ranking	knn	binary relev.
Semantic Category																	
V	Economics		1										1				1
B	Business economics		1										1				1
W	Economic sectors		1										1	2	1	2	1
P	Commodities		1		1												
N	Related subject areas													3	2	3	2
G	Geographic names		1	3	1				1					3	2		1
A	General descriptors		1						1								
	Total		6	3	2				2				3	8	5	5	6

Fig. 2. Contingency table of ratings aggregated by method and semantic category.

5 Discussion and Future Work

The software is under active development. In particular, we plan to support confidence information and quality estimation. Experiments have to be conducted to evaluate and improve the system. Some aspects of the implementation are currently tuned to the STW, especially regarding semantic properties that are beyond the scope of the SKOS specification, and thus differ among thesauri.

Acknowledgements. We would like to thank the indexing professionals who set the reviewing guidelines for valuable feedback, and thank the anonymous reviewer, who pointed us to related work in other domains, for constructive advice.

References

1. Berrios, D.C., Cucina, R.J., Fagan, L.M.: Methods for semi-automated indexing for high precision information retrieval. JAMIA **9**, 637–652 (2002)
2. Gastmeyer, M., Wannags, M., Neubert, J.: Relaunch des Standard-Thesaurus Wirtschaft - Dynamik in der Wissensrepräsentation. Inf. Wiss. Praxis **67**(4), 217–240 (2016)
3. Hinrichs, I., Milmeister, G., Schäuble, P., Steenweg, H.: Computerunterstützte Sacherschließung mit dem Digitalen Assistenten (DA-2). o-bib. Das offene Bibliotheksjournal **3**(4), 156–185 (2016)
4. Latif, A., Borst, T., Tochtermann, K.: Exposing data from an open access repository for economics as linked data. D-Lib Mag. **20**(9/10), 7 (2014)
5. Medelyan, O., Witten, I.H.: Measuring inter-indexer consistency using a thesaurus. In: Proceedings of Joint Conference on Digital Libraries, pp. 274–275. ACM (2006)
6. Toepfer, M., Seifert, C.: Descriptor-invariant fusion architectures for automatic subject indexing. In: Proceedings of Joint Conference on Digital Libraries (2017, accepted)
7. Uhlmann, S.: Automatische Beschlagwortung von deutschsprachigen Netzpublikationen mit dem Vokabular der Gemeinsamen Normdatei (GND). Dialog mit Bibliotheken **25**(2), 26–36 (2013)