Smart Topic Miner: Supporting Springer Nature Editors with Semantic Web Technologies

Francesco Osborne[1], Angelo Salatino[1], Aliaksandr Birukou[2], Enrico Motta[1]

[1] Knowledge Media Institute, The Open University, MK7 6AA, Milton Keynes, UK
{francesco.osborne,angelo.salatino,enrico.motta}@open.ac.uk
[2]Springer-Verlag GmbH, Tiergartenstrasse 17, 69121 Heidelberg, Germany
aliaksandr.birukou@springer.com

Abstract. Academic publishers, such as Springer Nature, annotate scholarly products with the appropriate research topics and keywords to facilitate the marketing process and to support (digital) libraries and academic search engines. This critical process is usually handled manually by experienced editors, leading to high costs and slow throughput. In this demo paper, we present Smart Topic Miner (STM), a semantic application designed to support the Springer Nature Computer Science editorial team in classifying scholarly publications. STM analyses conference proceedings and annotates them with a set of topics drawn from a large automatically generated ontology of research areas and a set of tags from Springer Nature Classification.

Keywords: Scholarly Data, Ontology Learning, Bibliographic Data, Scholarly Ontologies, Data Mining, Conference Proceedings, Metadata.

1 Introduction

An important challenge for academic publishers is to categorize their editorial products with respect to relevant research topics and keywords. This is critical for a variety of tasks that benefit both the publisher and the research community. First, the use of appropriate descriptors helps researchers in identifying relevant papers and supports academic search engines and recommender systems. In the second instance, a topic-based representation can inform marketing decisions, such as in which venues or communities to present a book. Finally, a granular description of the editorial content can be useful for producing advanced analytics about research trends, thus supporting publishing strategies.

Traditionally, editors classify proceedings manually, by associating to each proceedings book a list of categories from existing classifications (e.g., ACM, MeSH) and a set of keywords. This task is performed according to their experience in the research field, after analysing titles, abstracts, keywords and a list of additional terms from the call for papers. It is thus a costly and time-consuming process that may be biased by the editor view of the academic landscape. In addition, this manual analysis may miss the subtle emergence of some innovative topics or fail to detect the decline of traditional ones.

In this demo paper we present Smart Topic Miner (STM), a web application developed in collaboration with Springer Nature (SN) that classifies scholarly publications according to an automatically generated ontology of research areas. This

paper is complementary to the one accepted in the ISWC 2016 Applications Track and focuses on the main functionalities and the technical implementation of the system. We refer the reader to [1] for a comprehensive exposition of the set-covering algorithm, the knowledge bases and the system evaluation. The demo version of STM is available at http://rexplore.kmi.open.ac.uk/STM_demo. The reader can try it by using the 'Example Springer Nature Proceedings' option, which allows testing the application by using six default SN proceedings.

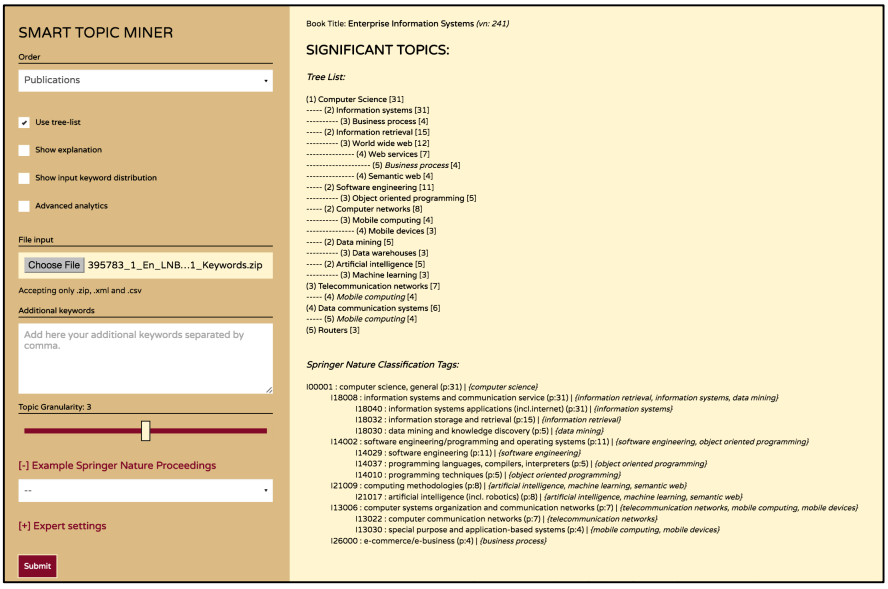

Figure 1. The STM interface.

2 Smart Topic Miner

When conference organizers send the proceedings to Springer Nature, the papers are typeset and copyedited. In the typesetting phase, XML files with relevant metadata are produced. The in-house editors analyse these metadata with the Smart Topic Miner for selecting a number of keywords and SN classification tags. The output includes: 1) a set of research areas structured according to an ontology of research areas, 2) a set of Springer Nature Classification tags, and 3) a variety of analytics for allowing editors to analyse the content of the proceedings and the quality of the classification. The web interface of STM is shown in Figure 1.

2.1 Knowledge Bases

STM categorizes publications according to two classifications: the Klink-2 Computer Science Ontology (CSO) and the SN Classification for Computer Science (SNC). CSO was created by applying the Klink-2 algorithm [2] on a dataset of about 16 million publications, mainly drawn from Computer Science. Klink-2 is an algorithm which generates an ontology of research areas by inferring semantic relationships

from scholarly metadata and external sources – e.g., DBpedia, calls for papers, web pages. It is integrated in Rexplore [3], an innovative system which uses semantic technologies for exploring and making sense of scholarly data. The current version of CSO includes about 17k topics linked by 70k semantic relations and structured in terms of 8 levels of granularity.

The Springer Nature Classification for Computer Science is an internal company classification, which is used to categorize proceedings, books and journals. It contains 76 categories in a three level taxonomy and was mapped to CSO by means of 349 relationships, so that every SN category is associated to a set of related topics.

2.2 Architecture

Figure 2 shows the STM architecture. When the user submits a collection of XML files, the parser extracts the relevant metadata, which are then sent as a JSON file to the background API via a POST query.

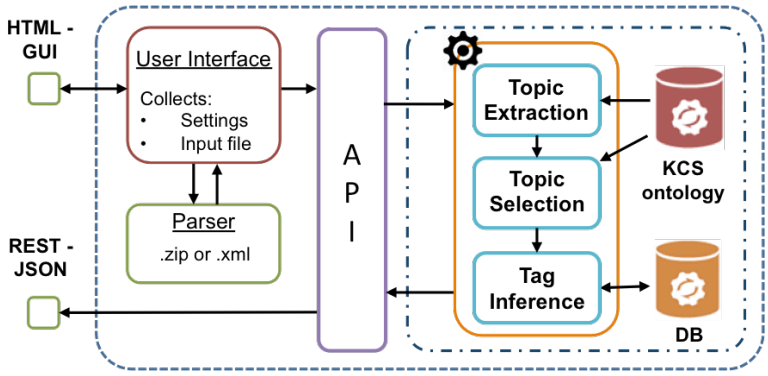

Figure 2. The STM architecture.

The backend tags each paper with a list of frequent terms extracted from its abstract, title, and keywords. Then it associates to each topic in the CSO ontology all papers tagged with its label or the label of a sub-area. For example, a paper containing the keyword "support vector machines" will be associated with the Support Vector Machines research area and with all its super-topics, such as Machine Learning and Artificial Intelligence. The set of topics is then pruned by applying a greedy set-covering algorithm. The size of the resulting set and the granularity of the topics depend on a number of parameters controlled by the users, as discussed in the next section. STM then uses the mapping between CSO and SNC to produce a relevant set of SNC tags. Finally, the outcome of the process is cached to improve performances of future queries and returned to the front-end.

2.3 Main Functionalities

The STM web interface was iteratively improved by taking in consideration the feedback of experienced SN editors and thus includes many functionalities to enhance their ability to customize the output and to assess its quality and coverage. In fact, editors did not want a completely automatic process, but a flexible tool that could be

used to investigate the proceedings and to produce different kinds of annotations according to their needs. In the following we will discuss the main functionalities and their rationales (further details are discussed in [1]).

The most used setting is the *granularity* value, which goes from 1 to 5 (default is 3) and allows users to intuitively choose how comprehensive the classification should be. Every level of granularity is associated with a number of settings of the set-covering algorithm. Figure 3 shows as an example a proceedings book processed with different granularities. A second important functionality is the *show explanation* one, which displays near each topic (e.g., Semantic Web) the list of keywords that were used to infer it (e.g., "OWL", "linked data", "ontology matching") and how many papers they cover. In fact, editors often want to investigate new or unexpected topics to decide if they have to be included in the final version of the annotations. Editors need also to check how representative a certain set of topics and tags actually is. For this reason, STM offers an *advanced analytics* functionality that provides additional information, such as the percentage coverage of the outcome and a list of uncovered and covered papers associated with their keywords and topics.

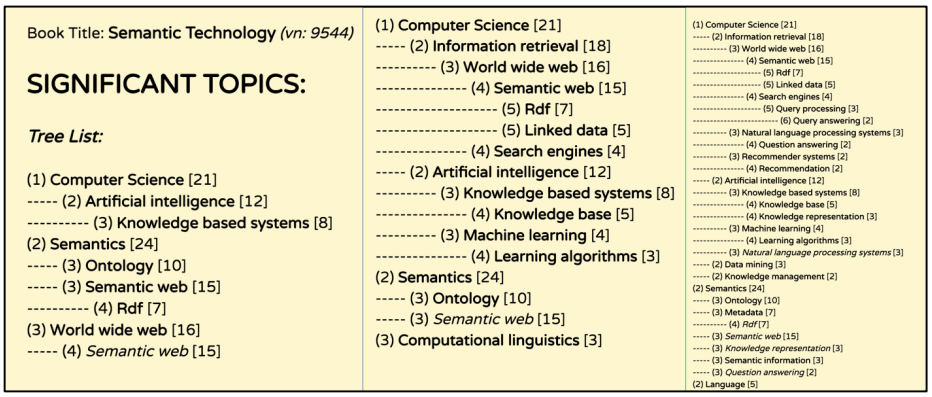

Figure 3. The same proceedings book processed with granularity 2, 3 and 4.

3 Conclusions

In this demo paper we summarized the main characteristics of Smart Topic Miner, a Semantic Web application designed to assist Springer Nature editors in classifying conference proceedings. We are now working on integrating STM into the Springer Nature workflow and we also plan to release a public version of the application to help researchers in choosing the set of topics which best describe their work.

References

1. Osborne, F., Salatino, A., Birukou, A., Motta, E.: Automatic Classification of Springer Nature Proceedings with Smart Topic Miner. In ISWC 2016 Application Track. (2016)
2. Osborne, F., Motta, E.: Klink-2: integrating multiple web sources to generate semantic topic networks. The Semantic Web-ISWC 2015, pp. 408-424. Springer. (2015)
3. Osborne, F., Motta, E. and Mulholland, P.: Exploring scholarly data with Rexplore. In International Semantic Web Conference (pp. 460-477). Springer. (2013)