Smart Topics Miner 2: Improving Proceedings Retrievability at Springer Nature

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

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

Abstract. Producing a robust and comprehensive representation of the research topics covered by a scientific publication is a crucial task that has a major impact on its retrievability and consequently on the diffusion of the relevant scientific ideas. Springer Nature, the world's largest academic book publisher, has typically entrusted this task to the most expert editors, which had to manually analyse new books and produce a list of the most relevant topics. To support Springer Nature in this task, we developed Smart Topic Miner, an application that assists the editorial team in annotating proceedings books according to a large-scale ontology of research areas. Over the past three years, we evolved this application according to the editors' feedback and developed a new engine, a new interface, and several other functionalities. In this demo paper, we present Smart Topic Miner 2, the most recent version of the tool, which is being regularly utilized by editors in Germany, China, Brazil, and Japan to annotate all book series covering conference proceedings in Computer Science, for a total of about 800 volumes per year.

Keywords: Scholarly Data, Bibliographic Metadata, Topic Classification, Topic Detection, Scholarly Ontologies, Data Mining.

1 Introduction

Producing a robust and comprehensive representation of the research topics covered by a scientific publication is a crucial task that has a major impact on its retrievability and consequently on the diffusion of the relevant scientific ideas. A high-quality representation of research papers, journals, and books will also enable several approaches for discovering and querying scientific articles, exploring research dynamics, extracting knowledge, and recommending papers. Springer Nature, the world's largest academic book publisher, has typically entrusted this task to their most expert editors, which had to manually analyse new books and produce a list of the most relevant topics. However, this process was expensive, time-consuming, and weighted exclusively on a small group of senior editors. For these reasons, in 2016 we developed Smart Topic Miner [1], an application which assisted the editorial team at Springer Nature in annotating proceedings books according to the Computer Science Ontology [2], a large-scale ontology of research areas.

In this demo paper we present the second version of Smart Topic Miner (STM) [3], which introduces several novelties including: i) a new approach for identifying research topics and producing an explanation for each suggested topic, ii) a new interface (showed in Figure 1), iii) the ability to take into account the annotation of previous editions of the conference in question, and iv) the integration with the CSO Portal [2] and the Springer Nature (SN) editorial systems. This paper is complementary to the one accepted in the ISWC 2019 In-Use Track and focuses on the main functionalities and the technical implementation of the system. We refer the reader to [3] for a comprehensive exposition of the main components, the evolution of the system in the recent years, the background data, and the system evaluation. The demo version of STM is available at http://stm-demo.kmi.open.ac.uk.

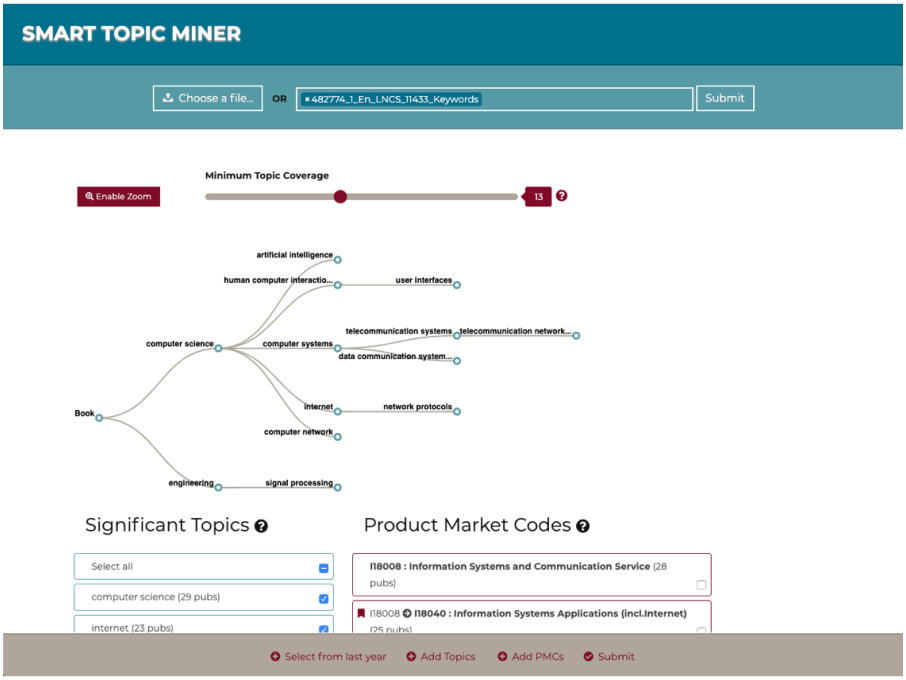

Figure 1. Smart Topic Miner 2.0 interface.

2 Smart Topic Miner 2

Smart Topic Miner 2 is a web application that assists editors in classifying books and more generally any collection of research papers. Specifically, it takes as input XML files describing the metadata of one or more books and returns: (i) a taxonomy of the relevant topics drawn from the Computer Science Ontology (CSO) [2], which is the largest taxonomy of research topics in the field; (ii) a set of relevant Product Market Codes (PMCs), Springer Nature internal classification system; (iii) an explanation for each topic, in terms of the text excerpts that triggered the topic identification; and (iv) a list of chapters from the book annotated with topics from CSO.

2.1 Knowledge bases

Smart Topic Miner 2 relies on three main knowledge sources: i) CSO, ii) PMCs, and iii) the metadata of Springer Nature publications.

CSO [2] is a large-scale ontology of research areas that currently includes 14K topics linked by 162K semantic relationships. It was automatically generated by running the Klink-2 algorithm [4] on a large dataset of research publications. The CSO data model[1] is an extension of SKOS[2]. It includes three main semantic relations: *superTopicOf*, which indicates that a topic is a super-area of another one (e.g., Semantic Web is a super-area of Linked Data), and *relatedEquivalent*, which indicates that two topics can be treated as equivalent for the purpose of exploring research data (e.g., Ontology Matching and Ontology Mapping).

The PMCs is a three-level mono-hierarchical classification system used by Springer Nature to categorize proceedings, books, and journals, and is used in the metadata describing the contents for the Springer Nature library[3] as well as third-party libraries and bookshops. The Computer Science section includes 103 categories characterizing both research fields and domains. We integrated CSO and PMCs by means of 332 relationships, so that every PMC is now associated to a set of related CSO concepts.

STM 2 also exploits a database of metadata which contains titles, abstracts, keywords and other information describing about 50K books published by SN in the field of Computer Science, including 10K proceedings. In this dataset each proceedings book is associated with an ID identifying its conference series, as well as with the topics and PMCs chosen by editors. This information is used to identify the previous edition of a conference and retrieve all relevant data.

2.2 STM Back-End

STM 2 parses the input files (generally XMLs) to retrieve the metadata associated to each chapter, which include title, abstract, list of keywords, and others. These metadata are then fed into the STM Engine, which identifies the relevant topics by running the CSO Classifier [5] on the title, abstract and keywords of each chapter.

The CSO Classifier is a tool that we developed for automatically classifying research papers in terms of relevant concepts drawn from CSO. The interested reader can refer [5] for additional details, and the open-source codebase is available on a GitHub repository: https://github.com/angelosalatino/cso-classifier.

The STM Engine also generates an explanation for each topic in terms of a distribution of text excerpts from which each topic was inferred. This process is important both for building trust in the system and for detecting possible mistakes. Next, STM uses the mapping between the PMCs and CSO to infer all relevant PMC identifiers, counting also the number of chapters associated to each PMC. Finally, STM identifies the conference series associated to the input proceedings (e.g., ISWC) and it retrieves the topics and PMCs associated to the same conference in the previous year. This information will be displayed alongside the current list of topics and PMCs.

To support the presentation of the results, STM 2 generates also a taxonomy of topics for each input book, using the *subTopicOf* relationships in CSO.

---

[1] CSO Data Model - https://cso.kmi.open.ac.uk/schema/cso

[2] SKOS Simple Knowledge Organization System - http://www.w3.org/2004/02/skos

[3] Springer Link - https://link.springer.com

2.3 User Interface

The user interface of STM 2 was redrawn and simplified following the editors' feedback which highlighted how the previous solution was sometimes too busy and unclear [1]. Figure 1 show the new interface which now consists of a top menu for loading the metadata of one or more books, a main panel for inspecting and selecting the CSO topics and PMCs, and a bottom menu for further options and submitting the classification to the production system at Springer Nature.

The editors can interact with this interface to explore the output, check why specific topics were inferred by the system, compare them with the annotations produced in previous editions, and include or exclude specific topics or PMCs according to their expertise. They can also inspect each chapter and see the relevant list of topics and text excerpts. The resulting topics and PMCs will be used for classifying proceedings in digital and physical libraries.

A statistical analysis of the user downloads in SpringerLink [3], the SN online library, showed that annotating books with STM had a massive impact on their discoverability. This resulted in about 9 million additional downloads over the last 3 years.

3 Conclusions

In this demo paper we presented Smart Topic Miner 2, a web application using semantic technologies, designed to assist Springer Nature editors in classifying conference proceedings. We are currently working on extending it in order to cover other fields of Science, such as Engineering. We are also investigating methods for learning from the feedback of editors in order to increase the STM accuracy. Finally, we plan to develop a version of STM to be used by authors for classifying their own research papers when producing the camera ready.

References

1. Osborne, F., Salatino, A., Birukou, A., Motta, E.: Automatic Classification of Springer Nature Proceedings with Smart Topic Miner. Semant. Web -- ISWC 2016. 9982 LNCS, 383–399 (2016).
2. Salatino, A.A., Thanapalasingam, T., Mannocci, A., Osborne, F., Motta, E.: The Computer Science Ontology : A Large-Scale Taxonomy of Research Areas. In: The Semantic Web -- ISWC 2018. Springer (2018).
3. Salatino, A.A., Osborne, F., Birukou, A., Motta, E.: Improving Editorial Workflow and Metadata Quality at Springer Nature. In: The Semantic Web – ISWC 2019. Springer Verlag (2019).
4. Osborne, F., Motta, E.: Klink-2: Integrating Multiple Web Sources to Generate Semantic Topic Networks. In: The Semantic Web - ISWC 2015. pp. 408–424 (2015).
5. Salatino, A.A., Osborne, F., Thanapalasingam, T., Motta, E.: The CSO Classifier: Ontology-Driven Detection of Research Topics in Scholarly Articles. In: TPDL 2019: 23rd International Conference on Theory and Practice of Digital Libraries. Springer.