Karlsruher Institut für Technologie

Institut für Angewandte Informatik und Formale Beschreibungsverfahren

Graduiertenkolloquium Angewandte Informatik

Data Mining and Information Extraction Methods for Large-Scale High-Quality Representations of Scientific Publications

M. Sc. Tarek Saier (AIFB)

Machine-readable representations of scientific publication corpora are the foundation of vital digital systems in academia. They capture information about publication contents as well as their interconnections, thereby enabling, informing, and accelerating processes such as academic search, decision making, and large-scale bibliographic analyses. As such, they are key to mitigating the challenges caused by the ever-increasing rate of publication and progress in academia. Especially disciplines of current focus, such as the rapidly evolving area of machine learning, can therefore benefit from improved digital records of the scientific progress. However, current research based on scientific publication corpora, such as the training and evaluation of machine learning models as well as bibliographic analyses, are based on limited data. Specifically, the volume, completeness, correctness, and granularity of publication representations used are limited. For example, only abstracts are considered instead of full-texts, or corpora with highly incomplete citation networks are used. As a consequence, the results and conclusions drawn from evaluations and analyses are of limited validity.

To address these issues, we present a set of data representation and information extraction approaches that enable the creation of machine-readable publication corpora more extensive, more complete, less noisy, and of finer granularity than previously available. In particular, we present the following. As the foundation of our research, we introduce unarXive, a large-scale data set comprising the full-text and citation network of over 1 million publications. Utilizing unarXive, we further present approaches yielding advances in three areas. First, we demonstrate improvements of the completeness of citation networks though the use of blocking, as well as improvements to the granularity of document representations. Second, we show advances in the language coverage of document interconnections through a large-scale analysis of cross-lingual citations. Third, we present information extraction approaches for fine-granular representations of research artifact citations. Overall, our approaches address key shortcomings of currently used machine-readable representations of scientific publication corpora. With our approaches we demonstrate their benefits through evaluations and practical large-scale applications. Significant parts of our work have already seen adoption within the research community, which confirms their utility.

Termin: Freitag, 27.10.2023, 14:00 Uhr

Ort: Kaiserstr. 89, 76133 Karlsruhe
Kollegiengebäude am Kronenplatz (Geb. 05.20), 1. OG, Raum 1C-04
(Hinweise für Besucher: http://www.aifb.kit.edu/web/Kontakt)

Veranstalter: Institut AIFB, Forschungsgruppe Web Science

Zu diesem Vortrag lädt das Institut für Angewandte Informatik und Formale Beschreibungsverfahren alle Interessierten herzlich ein.

M. Färber (Org.), S. Lazarova-Molnar, A. Oberweis, H. Sack,
A. Sunyaev, Y. Sure-Vetter, A. Vinel, M. Volkamer, J. M. Zöllner

Besucheranschrift:
KIT-Campus Süd
Institut AIFB – Geb. 05.20
Kaiserstr. 89
76133 Karlsruhe

Postanschrift:
KIT-Campus Süd
Institut AIFB – Geb. 05.20
76128 Karlsruhe