Scalable Author Disambiguation for Bibliographic Databases

Project abstract

Undoubtedly, the unambiguous attribution of scholarly material to their authors ranks among the most critical challenges for digital libraries. The main objective of this project is to significantly improve author disambiguation within two of the most comprehensive scholarly literature databases: the dblp computer science bibliography, maintained by Schloss Dagstuhl – Leibniz Center for Informatics, and the online service of Zentralblatt MATH (zbMATH), maintained by FIZ Karlsruhe – Leibniz Institute for Information Infrastructure. The corporation with the Heidelberg Institute for Theoretical Studies (HITS), one of the leading institutions in the field of named entity recognition, will help to bridge the gap between the latest results in natural language processing (NLP) research and their concrete application in digital libraries.

Key questions

Digital bibliography services collect and organize bibliographic meta data provided by the publishers of scholarly publications. Disambiguating the authors that hide behind the bibliographic meta data comprises multiple challenges. Firstly, meta data may be incomplete and is prone to typos and mistakes. Unfortunately, the importance of assigning authorship was under-appreciated in the past, so meta data inconsistencies are quite frequent. Additionally, a single individual may appear under different names across publications (synonyms). Name abbreviations, full name changes, and spelling or transliteration variants are not uncommon. The latter is the case of different transliteration rules for Asian or Cyrillic characters, which are known to have varied significantly across countries and epochs. To complicate matters further, the same name might refer to multiple individuals (homonyms), making name-only-based disambiguation simply insufficient. In fact, common names can be shared by tens of individuals within a research discipline. This is particularly acute in the case of Chinese authors, since the top ten surnames account for 40% of the population. In the Vietnamese case, a mere one hundred family names are estimated to be in common use.

Project goals
The foremost goal of this project is to develop scalable and flexible methods for author disambiguation and identification for digital libraries and bibliographic databases. To this end, the project partners aim to adopt state-of-the-art techniques and to develop new algorithms that can be used in the live production environment of dblp and zbMATH. This real-world application has very strict requirements for algorithmic techniques. While a number of different approaches have been proposed in the literature, essentially all fall short of being ready to use in a live production environment for one or another reason, most prominently due to the precision, scalability, and incrementality requirement of the disambiguation process. It is the objective of this project to build upon existing approaches and to develop new techniques for meeting these requirements. One particular unique opportunity of this project is to link common author entities in the dblp and zbMATH data stock, which will make new contextual information visible in the joint data set. The cooperation between both databases will provide a result superior to that potentially produced by either of the two partners alone. The cooperation with HITS as third project partner could not be more appropriate: as a reliable expert on entity resolution, the NLP group will bridge the gap between the academic NLP community devoted to the abstract problem and its concrete application to the available data from dblp and zbMATH. Cooperation with international partners, especially eScience.gov.cn and MathNet.ru, will provide valuable data in the context of non-western names, and will be a significant step towards the globalization and internationalization of both databases.

Our approach

The collaboration will involve the following tasks:

- Synthesis of the data stock of dblp and zbMATH. Linking dblp and zbMATH will allow the joint data of both databases to be analyzed as a whole, and will enable both databases to mutually correct and enrich their data in the future.
- Data enrichment of both data stocks using the data provided by eScience.gov.cn, MathNet.ru and further international data delivery partners. The Unicode encoding of the actual (e.g., Asian or Cyrillic) author names is helpful in highly ambiguous non-western cases.
- Adaptation of (theoretical) state-of-the-art methods to the real-word scenario of living, dynamic digital libraries, and evaluation of their utility in this practical context, in particular (a) adapting scalable graph-based methods to the incremental data curation model of living digital libraries, and (b) building upon state-of-the-art techniques from natural language processing and machine learning such as Markov logic networks.
- Implementation of new algorithms and software tools in the production systems of dblp and zbMATH. All implemented infrastructures will be maintained and extended well beyond the duration of this project.

All results will be published. Developed methods will be of interest for digital libraries in general, and the infrastructure-providing institutes of the Leibniz Association in particular.

Outcome: publications, datasets, code

- Mark-Christoph Müller, Adam Bannister, Florian Reitz: Off-The-Shelf Semantic Author Name Disambiguation for Bibliographic Data Bases. TPDL (Demos) 2019: 397-400. [doi.org](https://doi.org/10.1007/978-3-030-30760-8_42) [https://github.com/nlpAThits/scad-tool](https://github.com/nlpAThits/scad-tool)
- Mark-Christoph Müller: Semantic Matching of Documents from Heterogeneous Collections: A Simple and Transparent Method for Practical Applications. RELATIONS Workshop 2019: 34-40. [doi.org](https://doi.org/10.18653/v1/W19-0804) [https://github.com/nlpAThits/TopNCosSimAvg](https://github.com/nlpAThits/TopNCosSimAvg)
- Mark-Christoph Müller: On the Contribution of Word-Level Semantics to Practical Author Name Disambiguation. JCDL 2018: 367-368. [https://doi.org/10.1145/3197026.3203912](https://doi.org/10.1145/3197026.3203912)
- Oliver Hoffmann, Florian Reitz. hdblp: historical data of the dblp collection, Zenodo 1213050, Apr. 2018 https://doi.org/10.5281/zenodo.3051910
- Florian Reitz: Two Test Collections for the Author Name Disambiguation Problem based on DBLP, Zenodo 1201323, Mar. 2018 https://doi.org/10.5281/zenodo.1201324
- Mark-Christoph Müller: Semantic Author Name Disambiguation with Word Embeddings. TPDL 2017: 300-311 https://doi.org/10.1007/978-3-319-67008-9_24