Abschließender Sachstandsbericht
Leibniz-Wettbewerb

Scalable Author Disambiguation for Bibliographic Databases
Antragsnummer: K184/2014

Berichtszeitraum: 01.06.2015 – 31.12.2018

Federführendes Leibniz-Institut:
Schloss Dagstuhl – Leibniz-Zentrum für Informatik GmbH

Projektleiter/in:
Raimund Seidel, Ph.D.

1. Achieved objectives and implemented milestones

The main objective of this project was to significantly improve the author name disambiguation (AND) within the dblp computer science bibliography and the online service zbMATH. As specified in the work plan of the proposal, a data exchange and record linkage workflow has been established between dblp and zbMATH which will be maintained well beyond the duration of this project. A number of high-quality training and test datasets have been constructed and published. We specified appropriate evaluation metrics and developed algorithmic methods to support the identification of errors in the existing data stock as well as the low-error inclusion of new metadata. These methods have been integrated into the live production environment of dblp and zbMATH and are actively contributing towards improving the metadata quality in both services. All results have been published in international, peer-reviewed outlets and made available as open-access.

The adoption of state-of-the-art techniques from the current scientific literature to the real-world scenario of a practical bibliographic database, however, has proven to be practically infeasible. None of the existing implemented methods did scale well for corpora of several millions of items and most of them were too demanding in terms of resources. Additionally, the planned integration of metadata from non-Western data partners has been reevaluated and deemed to promise only little potential to improve our algorithms, so only very little attention has been given to this task. Instead, a lot more attention has been directed towards the emerging field of artificial neural networks and embedding based machine learning algorithms, which has not been part of the original proposal of 2014.

More detailed information on the activities, obstacles, results, and achievements of this project can be found in Section 2 and 3.

2. Activities and obstacles

Adaptation of state-of-the-art techniques to the problem-specific domain

The initial goal of adapting state-of-the-art algorithms to the individual data sets of the projects proved to be difficult. The major obstacles were that (1) existing literature mostly ignores the aspect of data blocking and (2) it proved to be difficult to transfer algorithmic solutions between projects with different data properties.

Blocking is a divide-and-conquer approach to reduce the execution time and resource demand of algorithms. State-of-the-art algorithms are usually evaluated with very small test data sets that either make blocking unnecessary or provide a predefined blocking. We attempted to directly implement state-of-the art approaches but failed because of the size of our data sets. Each project had already developed blocking approaches based on the properties of their data and the underlying data model. E.g., dblp relies heavily on a multi step blocking based on coauthors. This is not easily transferable to zbMATH as, traditionally in the field of Mathematics, there are fewer coauthors per publication and the underlying data model at zbMATH allows for fuzzy mappings between authors and documents.

The existing blocking solutions also proved difficult to adapt to state-of-the-art algorithms. To circumvent these problems, we concentrated on identifying small modules in the disambiguation process that can be improved with machine learning and integrated into our existing framework.

Specification of evaluation metrics

Traditionally, research in AND has used a set of task-specific evaluation metrics which were not widely applied outside that domain, like e.g. ACP, AAP, and the related K-metric. Discourse-oriented NLP (Natural Language Processing) as done at HITS, on the other hand, has been working for quite some time on tasks like co-reference and entity resolution, which are structurally very similar to AND. For these tasks, there is a considerable body of work on the definition of correct and plausible evaluation metrics, incl. the B-Cubed measure. From the outset of the project, we applied this more appropriate metric, both for quality control and for disambiguation evaluation proper.

NLP methods for author name disambiguation

One of the key assumptions concerning the potential role of NLP in AND was that semantic similarity between publication titles and/or abstracts is an important indicator of author identity. Traditionally, AND used to rely strongly on coauthor information, while publication title similarity, if addressed at all, was mainly tackled on the string or n-gram level.

In the project, we used vector-based representations of word semantics ('word embeddings') from NLP for computing discriminative similarity scores for word sequences (i.e., publication titles). At the beginning of the project, these scores were employed as one class of features in an AND machine learning classifier. In later work, the focus shifted towards simpler, more transparent machine learning, and towards rule-based systems, while at the same time the methods employed for computing semantic similarity became more refined. These methods included the training of domain-specific word embeddings as 'pluggable' semantic resources, and algorithms for computing semantic similarity of word sequences that were both more scalable and more transparent.

3. Results and achievements

Record linkage between the dblp and zbMATH data stock

During the course of the project, a process has been established to identify and link common author entities in both dblp and zbMATH. Here, we chose a conservative semi-automated approach that favors the precision of the matching and aims to avoid false-positive matches. In detail, profiles have only been matched if they are based on quality-checked (“manual”) document assignments. Moreover, we analyzed the fragmentation of the coauthor communities on any given author profile in dblp and zbMATH and only linked if all coauthor communities were uniquely connected by documents common in both databases. In case of ambiguities, the document assignment have been manually checked and numerous misassignments could be detected and corrected in both databases.

Using this workflow, by the end of 2018, there have been 2,435 curated profiles in dblp that point to their counterpart in zbMATH and 4,588 zbMATH profiles with a verified link to dblp. It should be noted that these selections of curated profiles are representing particularly interesting authors in both databases. E.g, while a random dblp author profile lists 5.9 publications on average, the linked zbMATH profiles list 71.7 publications on average.

Special attention has been given to the emerging role of open, persistent identifier (PID) schemes in the process of linking both databases. In particular, during the course of the project, WikiData (wikidata.org) has emerged as a central, community driven authority linking hub for PIDs. Both dblp and zbMATH contribute continuously to WikiData by providing quality-checked linkage (17,405 dblp PIDs and 8,336 zbMATH author IDs as of end of 2018) towards their WikiData items. In doing so, further PIDs like ORCID, ISNI, and others can be retrieved from WikiData and used as a valuable linkage indicator.

Publication of high-quality training and test data-set

In order to create a training and test dataset for author disambiguation we took advantage on the fact that the dblp and zbMATH datasets of documents overlap significantly. We first restricted both sides to manually curated author profiles, and took then the actual intersection by mapping records from the two individual files on the basis of DOI and author position. This produced an intersection file of 2,886 authorship records in 2,779 publications, for which we computed the

B^3 metric between the manual author disambiguations at zbMATH and dblp. The obtained F1-score of 0.995 indicates a near-perfect agreement between these independent author disambiguation decisions, which we took as a sufficient evidence for the actual correctness of the disambiguations.

We concluded from this that the quality level of the manual author disambiguation in the zbMATH database is sufficient for the data to be used as the basis of a new AND data set, without the need for additional manual annotation. This gave rise to the SCAD-zbMATH dataset [1], consisting of 30000 authorship records (this relatively "small" number is mainly due to licensing restrictions), all manually disambiguated, and with only fully disambiguated documents (for each document, all the authorship records are disambiguated). This new AND dataset has also some additional properties that make AND difficult in practice, and that make it novel compared to those already existing: "high author name variability", "high name homography", "high proportion of single-author publications", "low number of publications per author".

In a slightly different approach, we created another test collection that is based on known defects in dblp [2]. From the curation history of the database, we automatically identified defects and their correction. Based on these defects, we build a test collection that can be used to study properties of defects and to test how well algorithms handle defects. This dataset, as well as SCAD-zbMATH, is published in an open data repository.

New methods for identification of errors in the existing data stock

Identifying (and fixing) homonymous and synonymous author profiles is one of the major tasks of curating personalized bibliographies like the ones provided by dblp and zbMATH. We developed an efficient and scalable machine learning approach to identify homonymous bibliographies using artificial neural networks [3]. We trained our model using the gold-standard curation data set derived from the past years of active, manual curation at the dblp computer science bibliography. Furthermore, we developed a new vectorization method for bibliographies into 40 data dimension which had been identified from studying the actual, manual bibliography curation practice at dblp. As a diagnostic test, the labeling computed by our classifier shows a Matthews correlation of 0.54 (i.e., a strongly positive relation), and an accuracy of 74% in correctly reporting homonymous bibliographies. To the best of our knowledge, this is the first published result addressing the homonym bibliography detection problem.

Another way to obtain hints about homonymous author profiles is using the Mathematics Subject Classification (MSC) scheme. Using MSC, one can find areas of study that are seldom seen together. We constructed a metric using the frequency where certain MSCs are combined with other MSCs in papers, citations and our manually cleaned author data. With this metric, we were able to search for outliers in our data. This meant we could score an author based on the distance of the MSCs in the papers they wrote. This has proven to be very successful at finding incorrectly assigned authors who otherwise would have been overlooked.

New methods for low-error inclusion of new data

Supporting the inclusion of new publication data into an existing, production-level bibliographic database can be considered the most important of the application-oriented tasks in AND. In contrast to pure research or academic approaches, where entire data sets of publications and/or author profiles are clustered at once, a production system has to deal with one single new publication at a time. Also, in the former approaches, system recall and precision are weighted equally, while a production environment is required to be more conservative, putting more emphasis on precision to prevent low-quality data from entering the database.

In the project, these special aspects of application-oriented AND were addressed in several ways. We identified the comparison of two publications with a shared identical (or highly similar) author name as an atomic AND procedure. From this procedure, higher-level methods can easily be derived. All methods implemented for this comparison (both machine learning and rule-based) were designed to provide a similarity score for each classification [4,5]. Using this score as a confidence threshold in a fully automatic system, data quality can be controlled. In addition, a recently developed method [6,7] can also provide a compact qualitative representation of the word-level similarities between the publication titles that were compared which, in combination with the similarity score, can be used for human sanity checking.

Integration of methods into the dblp and zbMATH live production environment

The results of the project have been integrated into the workflow of dblp and zbMATH primarily in three ways. (1) The availability of additional test sets and our improved understanding of task specific evaluation metrics allowed us to increase the performance of existing production components. E.g., the means for systematic evaluation helped us to incorporate new features in existing AI algorithms or to tune parameters specifically for our data set. The test sets were also used for project specific training of workflow components. Several machine learning techniques and strategies have been implemented in a semi-automatic workflow to match author entities with those of other relevant databases (for example, at zbMATH, more than 38,840 links to MGP entities, 30,000 to ORCID and 3,900 to WikiData). (2) The project created a number of software artifacts that are now integrated in the workflow of our projects. This includes methods to measure semantic distance between publications, adjusted specifically to the properties of our data sets. It also includes components that are trained to identify profiles with specific defects. E.g., we can now automatically identify homonyms in dblp using a system trained on the test collections created in this project. (3) The data exchange between dblp and zbMATH helped to identify defects in both collections. More importantly, the exchange provided a test ground for semi automatic reconciliation with other curated databases such as MGP, ORCID and WikiData. These reconciliations have already identified thousands of defective profiles and helped repairing them, thus zbMATH and dblp will continue to exchange data.

4. Equal opportunity

All partner institutes promote gender equality in research and in the workplace. Job openings have been advertised in a gender neutral fashion. Disabled applicants whose qualification and aptitude were equal to that of other applicants had been given preferential consideration in vacancies.

5. Quality assurance

The project produced a large-scale gold standard AND data set on the basis of the zbMATH data, although, for licensing reasons, only a smaller subset could be made publicly available (SCAD-zbMATH, see Section 2 above). Data quality for this data set was controlled by measuring the disambiguation agreement on a common subset of dblp and zbMATH data. We found a near-perfect agreement (above 0.99 B-Cubed F-measure), which allowed us to assume overall correctness of both data sets.

As for the developed disambiguation methods, we employed commonly accepted evaluation measures (B-Cubed, Matthews Correlation Coefficient), and performed cross-validation and significance testing where applicable. In order to facilitate comparison of our methods to existing ones, we also employed an off-the-shelf AND data set (KISTI), using both our preferred evaluation metric B-Cubed and the K-score metric which is more widely used in the AND community.

Results were published in conferences with Open Access proceedings where possible, or with an accompanying, open access author copy in one case. One article in a Closed Access journal was published as Open Access for a fee, and as a preview version on arXiv.org.

Where possible, implemented methods and code were also made available at the relevant places, e.g. on GitHub.com.

6. Additional resources

HITS contributed further EUR 115,403.88 (18 PM of a post-doc researcher) in personnel costs and EUR 95.37 in material costs to the project.

7. Structures and cooperation

Continued and strengthened cooperation of dblp and zbMATH

Schloss Dagstuhl LZI and FIZ Karlsruhe agreed to a cooperation agreement for the purpose of improving their respective databases dblp and zbMATH in 2012. The creation of this project was a direct result of this earlier cooperation agreement. Since the project started, a number of workflows have established in order to improve the quality of the author identification in both databases. In particular, this includes the continuous exchange of metadata, the growing inter-linkage of bibliographies, and the common efforts for strengthening central open authority hubs like ORCID and WikiData by contributing their data.

Furthermore, all three project partners intend to continue their cooperation through the joint publication of further research results and the exchange of their data in order to enable further research. The developed APIs and packages will also serve as the basis for future joint infrastructure and research services.

Outreach to further external data delivery partners in non-western regions

The initially planned integration of data from the all-Chinese portal eScience.org.cn (former: eScience.gov.cn) has not been brought to fruition. An initial inspection of sample data from the portal did not show the anticipated overlap with our databases. Likewise, data from the Chinese Academy of Sciences would only have been available for additional fees, and without an open license for further distribution. Also, as the focus of the project shifted towards artificial neural network based machine learning algorithms, the integration of non-Western metadata was no longer a priority.

However, the developed tools and techniques served to deepen the existing cooperation of zbMATH and MathNet.ru with respect to mutual interlinking, quality control, and correction feedback. Likewise, a new exchange layer for cyrillic data could be introduced, allowing for a more effective merging of different transliterations. As a result, this component resulted in the proper merging of 14,476 hitherto separated spellings from transliterations to 4,765 joint identities.

8. Outlook

Concerning automatic methods for AND, important results and lessons learned of the project were that (1) NLP-based methods for short-text matching, like those addressing the semantics of publication titles, abstracts, or keywords, can significantly improve the performance of AND, (2) efficiency and robustness are of key importance for production-level (in contrast to academic) systems, ruling out many state-of-the-art NLP approaches as too demanding, (3) at the same time, for reasons of data quality, real-life AND (as done at dblp and zbMATH) needs to be conservative, shifting the required level of performance towards high precision (rather than recall). This, combined with the fact that production-level AND will always have a human expert in the loop, indicates the following route for further AND research: Develop methods for short-text matching which combine state-of-the-art semantic resources (like word embeddings) with efficient, light-weight algorithms. At the same time, the matching algorithms should not behave as black boxes which just output a matching score, but they should provide additional information which explains the matching decision and makes it transparent to a human expert.

References

[1] Mark-Christoph Müller, Florian Reitz, Nicolas Roy: Data sets for author name disambiguation: An empirical analysis and a new resource. Scientometrics 111(3): 1467-1500 (2017). doi:10.1007/s11192-017-2363-5

[2] Florian Reitz: Harnessing Historical Corrections to Build Test Collections for Named Entity Disambiguation. TPDL 2018: 47-58. doi:10.1007/978-3-030-00066-0_4

[3] Marcel R. Ackermann, Florian Reitz: Homonym Detection in Curated Bibliographies: Learning from dblp's Experience. TPDL 2018: 59-65. doi:10.1007/978-3-030-00066-0_5

[4] Mark-Christoph Müller: Semantic Author Name Disambiguation with Word Embeddings. TPDL 2017: 300-311. doi:10.1007/978-3-319-67008-9_24

[5] Mark-Christoph Müller: On the Contribution of Word-Level Semantics to Practical Author Name Disambiguation. JCDL 2018: 367-368. doi:10.1145/3197026.3203912

[6] Mark-Christoph Müller: Semantic Matching of Documents from Heterogeneous Collections: A Simple and Transparent Method for Practical Applications. RELATIONS Workshop 2019: 34-40. To appear.

[7] Mark-Christoph Müller, Adam Bannister, Florian Reitz: Off-The-Shelf Semantic Author Name Disambiguation for Bibliographic Data Bases. TPDL (Demos) 2019. To appear.