Heinrich Heine Universität Düsseldorf

Heinrich-Heine-Universität Düsseldorf 40204 Düsseldorf
Dekanat der Mathematisch-Naturwissenschaftlichen Fakultät

An alle
hauptamtlichen Professoren/innen
und Privatdozenten/innen
des Faches Informatik der
Mathematisch-Naturwissenschaftlichen Fakultät

Mathematisch-Naturwissenschaftliche Fakultät

Dekanat

Promotionsangelegenheiten
Universitätsstraße 1
40225 Düsseldorf
Telefon: +49 (0)211 81 15092
E-Mail: promotionmnf@hhu.de

13.01.2023

Promotionsverfahren von **Herrn M.Sc. Tobias Backes**
Auslage der Dissertation und Gutachten sowie Termin der mündlichen Prüfung
Anlage: Einseitige Zusammenfassung der Dissertation

Sehr geehrte Damen und Herren,

in dem oben genannten Promotionsverfahren wird die Annahme der Dissertation

Partial Orders and Progressive Blocking: A Matching-based Framework for Large-scale Entity Resolution in Bibliographic Data

von den Berichterstattenden Prof. Dr. S. Dietze, Prof. Dr. S. Conrad und Prof. Dr. L. Dietz beantragt. Sie kann zusammen mit den Gutachten in der Zeit

vom 23.01.2023 bis 09.02.2023

eingesehen werden. Bitte wenden Sie sich zur Einsicht an das Promotionsbüro (promotionmnf@hhu.de).

Einsprüche gegen diese Dissertation können nur zwei Tage nach der vorgenannten Frist geltend gemacht werden. Erfolgt kein Einspruch, so gilt die Dissertation als angenommen (§ 7 Ziffer (5) PO).

Sofern die Dissertation angenommen wird, findet die mündliche Prüfung am

14.02.2023 um 14:30 Uhr

im **Geb. 25.22, Hörsaal 5 H** statt. Als Prüferinnen bzw. Prüfer sind vorgesehen:
Prof. Dr. A. Dilthey und Prof. Dr. J. Rothe.

Die Öffentlichkeit ist bei der Befragung zugelassen.

Mit freundlichen Grüßen
im Auftrag

Silke Krispin

Abstract

Entity resolution (*ER*) is the task of grouping entity mentions by the real-world object they refer to. It is central to ordering and aggregating knowledge in the growing amount of available structured, semi-structured and unstructured information. At its core, ER determines whether the similarity between two entity representations is sufficient to indicate equivalence. From a technical point of view, the most essential aspect is not how to compute this similarity, but how to discover the most similar pairs efficiently. As it is infeasible to compare all mention pairs, dedicated techniques must be used to structure or partition the search space. This is known as *similarity search*. A simple approach is to establish a prior grouping based on selected key features or combinations thereof. This is known as (hash-based) *blocking* and is limited in modelling intransitive matching relationships. Alternative heuristics like alphabetical order can be used to suggest mention pairs in the order of their approximated coreference likelihood. This is known as *progressive* resolution. Progressive methods have focused on alphabetically sorting string-based entity representations, which optimistically assumes that coreference likelihood can be approximated as a total order. In this thesis, we suggest to *partially order* entity representations instead, as the subset partial order is better suited to model the matching relationships between sets of features. A notion of neighborhood as has been previously defined for total orders (e.g. alphabetical sorting) or continuous spaces (e.g. space partitioning) can also be defined on partial orders and exploited for progressive resolution. In this thesis, we explore opportunities of partially ordering entity mentions to develop a generalized set-based framework that can be adapted to ER tasks such as progressive author disambiguation, hierarchical affiliation resolution and large-scale duplicate detection. In a series of works, we have explored the topics of clustering, blocking and progressive resolution in the context of author disambiguation. This was followed by experiments with hierarchical resolution of affiliation strings and billion-scale blocking for detecting duplicate publication records. In the process, a modular entity resolution framework was refined that consists of the steps (1) representation, (2) specification, (3) generalization, (4) separation, (5) collocation and (6) conflation. Entity mentions are (1) represented as sets of attribute-value pairs, which are in some cases (2) isolated by specification if they are not informative enough. Further, (3) hypothetical representations are added by removing features that are not required for blocking equivalence. The result corresponds to sufficient overlaps for blocking equivalence. Then, (4) the representations are separated into super-blocks consisting of representations that are somehow connected in the subset partial order, which is built explicitly in (5) collocation. Finally, (6) edge-weights based on observation counts can be used in the partial order's directed acyclic graph to progressively contract edges, thereby merging nodes (blocks) to increase the size of clustering tasks. Each development step in the series of publications related to this thesis has been evaluated on gold datasets for different application scenarios in the domain of bibliographic data. Thereby, we have proven the practicability of our approach, shown where it outperforms existing baselines and where current limitations call for further research. The description of our works is complemented by a brief discussion of each individual publication and embedded in the body of existing literature by an integrative introduction and preliminaries chapter as well as a dedicated related work chapter. In the final chapter, we conclude how we have been able to design a novel ER approach that is unique in how it combines a number of beneficial properties in a modular and easily adaptable framework. From a number of inherent limitations, we derive tasks for future work before summarizing our contributions and how they have addressed gaps in the existing ER literature.