University of Zurich[UZH]

Department of Informatics

University of Zürich
Department of Informatics
Binzmühlestr. 14
CH-8050 Zürich
Phone. +41 44 635 43 11
Fax +41 44 635 68 09
www.ifi.uzh.ch/dbtg

UZH, Dept. of Informatics, Binzmühlestr. 14, CH-8050 Zürich

Prof. Dr. Michael Böhlen
Professor
Phone +41 44 635 43 33
Fax +41 44 635 68 09
boehlen@ifi.uzh.ch

Zürich, August 17, 2022

MSc Project: Implementing and evaluating different algorithms for entity matching

Entities, such as customers or products, may be recorded more than once in an information system during different transactions or there may be different records stored in separate source systems that are to be integrated. Entity matching is the process of identifying the different records describing the same entity, even if these records are not identical. Records may deviate from each other due to data entry errors, such as typos, or because the overall structure of the records differs. Over the course of the last decades, many different algorithms and approaches have been developed, but there is still not silver bullet for solving this problem.

The goal of the MSc project is to implement several different algorithms, evaluate them using benchmarks employing real-world datasets, and identify their strengths and weaknesses. In particular, the work is structured into the following tasks:

- **T1: Reading and understanding the literature**
  This encompasses reading the papers by Brunner and Stockinger[1] and Helmer et al.[2] and the book chapter on data duplication[3] (and any other relevant literature, such as technical reports and supplementary material) in order to gain a deeper understanding of the problem and the inner workings of the algorithms.
- **T2: Implementing the algorithms**
  The second task is about actually implementing the algorithms after studying them in the previous task. The algorithms described in [1] and [2] (cross-parsing) have to be implemented, some of the other algorithms (e.g. from [3]) are chosen according to preferences and after some further discussion. During the implementation, the algorithms also have to be tested for correctness.
- **T3: Setting up the benchmark**

A benchmark has to be set up for the evaluation in the following task. This includes preprocessing the data to prepare it for the evaluation and to decide on various parameters of the benchmarks. Possible datasets are the Magellan dataset mentioned in [1] and the EU financial sanction list (available here: `https://data.europa.eu/data/datasets/consolidated-list-of-persons-groups-and-entities-subject-to-eu-financial-sanctions?locale=en`).

- **T4: Evaluating the algorithms**
  After the algorithms have been implemented and tested for correctness and the benchmark has been set up, the next step is to evaluate the performance of the algorithms. This includes such parameters as the run-time of entity matching itself, the training time of machine-learning approaches, and the memory footprint of the algorithms.
- **T5: Summarizing the findings in a report**
  At the end of the project, a report describing the implemented approaches, the benchmark, the evaluation and the findings needs to be written up.

References

[1] U. Brunner and K. Stockinger. Entity matching with transformer architectures - A step forward in data integration. In *Proc. of the 23rd Int. Conf. on Extending Database Technology (EDBT'20)*, pages 463–473, Copenhagen, Denmark, 2020.

[2] S. Helmer, N. Augsten, and M. H. Böhlen. Measuring structural similarity of semistructured data based on information-theoretic approaches. *VLDB J.*, 21(5):677–702, 2012.

[3] I. F. Ilyas and X. Chu. *Data Cleaning*. Association for Computing Machinery, New York, NY, USA, 2019.

Supervisor: Sven Helmer

Department of Informatics, University of Zurich

Prof. Dr. Michael Böhlen