HPI Colloquium
07.02.2019, 4 p.m.

Hasso-Plattner-Institut, Hauptgebäude, H-E.51/52
Campus Griebnitzsee, 14482 Potsdam

"Data Profiling at Scale"

Dr. Thorsten Papenbrock
Hasso-Plattner-Institut, Potsdam

Abstract

According to a CrowdFlower study [1] and common experience, data scientists spend about 80 percent of their time not on data science but data preparation. In industry, the same is true for IT-professionals, who aim to integrate, connect, and consolidate business data from third party sources. A major part of that data preparation effort is spend on understanding the structure of the data and finding correspondences to existing datasets or schemata. This process, i.e., the search for structural patterns and dependencies is called data profiling and it involves various metadata discovery tasks of exponential complexity; some of them are even amongst the hardest tasks in computer science. Most automatic data profiling algorithms do, for this reason, not scale well with the volume of the data.
In this talk, I will provide an overview of our research in the field of data profiling and discuss the challenges ahead. We developed several algorithms that improved the efficiency of automatic data profiling by many orders of magnitude and published them with a practical tool called Metanome. This tool is used by various research groups and companies all over the world and we aim to drive its development further. The three main objectives for our future research are metadata interpretation (filtering, ranking, and selection), metadata application (data linkage, cleaning, integration, and query optimization), and metadata search parallelization/distribution (for scalability, robustness, and efficiency).

Short CV

Thorsten Papenbrock is a researcher and lecturer at the Hasso Plattner Institute. He received his M.Sc. in IT-Systems Engineering in 2014 and his Ph.D. in Computer Science in 2017. The goal of his research is to create efficient tools that make data accessible. For this purpose, he develops novel algorithms and approaches to profile, cleanse, integrate, link, and structure data. His interests also involve techniques for distributed and parallel computing, database systems, and data analytics. More details about his research and teaching activities can be found online [2].

Host: Prof. Dr. Felix Naumann

[1] https://visit.figure-eight.com/rs/416-ZBE-142/images/CrowdFlower_DataScienceReport_2016.pdf
[2] https://hpi.de/naumann/people/thorsten-papenbrock.html