Building Scalable Machine Learning Solutions for Data Cleaning

Ihab Ilyas[1]

Abstract

Machine learning tools promise to help solve data curation problems. While the principles are well understood, the engineering details in configuring and deploying ML techniques are the biggest hurdle. In this talk I discuss why leveraging data semantics and domain-specific knowledge is key in delivering the optimizations necessary for truly scalable ML curation solutions. The talk focuses on two main problems: (1) entity consolidation, which is arguably the most difficult data curation challenge because it is notoriously complex and hard to scale; and (2) using probabilistic inference to suggest data repair for identified errors and anomalies using our new system called HoloClean. Both problems have been challenging researchers and practitioners for decades due to the fundamentally combinatorial explosion in the space of solutions and the lack of ground truth. There's a large body of work on this problem by both academia and industry. Techniques have included human curation, rules-based systems, and automatic discovery of clusters using predefined thresholds on record similarity Unfortunately, none of these techniques alone has been able to provide sufficient accuracy and scalability. The talk aims at providing deeper insight into the entity consolidation and data repair problems and discusses how machine learning, human expertise, and problem semantics collectively can deliver a scalable, high-accuracy solution.

Author

Ihab Ilyas is a professor in the Cheriton School of Computer Science and the NSERC-Thomson Reuters Research Chair on data quality at the University of Waterloo. His main research focuses on the areas of big data and database systems, with special interest in data quality and integration, managing uncertain data, rank-aware query processing, and information extraction. Ihab is also a co-founder of Tamr, a startup focusing on large-scale data integration and cleaning. He is a recipient of the Ontario Early Researcher Award (2009), a Cheriton Faculty Fellowship (2013), an NSERC

[1] Cheriton School of Computer Science, University of Waterloo, Canada, ilyas@uwaterloo.ca

doi:10.18420/btw2019-02

Discovery Accelerator Award (2014), and a Google Faculty Award (2014), and he is an ACM Distinguished Scientist. Ihab is an elected member of the VLDB Endowment board of trustees, elected SIGMOD vice , and an associate editor of the ACM Transactions of Database Systems (TODS). He holds a PhD in computer science from Purdue University, West Lafayette.