Afternoon Tutorial

Entity Resolution for Big Data

Lise Getoor
University of Maryland, College Park

Ashwin Machanavajjhala
Duke University

Abstract

Entity resolution (ER), the problem of extracting, matching and resolving entity mentions in structured and unstructured data, is a long-standing challenge in database management, information retrieval, machine learning, natural language processing and statistics. Accurate and fast entity resolution has huge practical implications in a wide variety of commercial, scientific and security domains. Despite the long history of work on entity resolution, there is still a surprising diversity of approaches, and lack of guiding theory. Meanwhile, in the age of big data, the need for high quality entity resolution is growing, as we are inundated with more and more data, all of which needs to be integrated, aligned and matched, before further utility can be extracted. In this tutorial, we bring together perspectives on entity resolution from a variety of fields, including databases, information retrieval, natural language processing and machine learning, to provide, in one setting, a survey of a large body of work. We discuss both the practical aspects and theoretical underpinnings of ER. We describe existing solutions, current challenges and open research problems. In addition to giving attendees a thorough understanding of existing ER models, algorithms and evaluation methods, the tutorial will cover important research topics such as scalable ER, active and lightly supervised ER, and query-driven ER.

KDD'13, August 11–14, 2013, Chicago, Illinois, USA.
ACM 978-1-4503-2174-7/13/08.

Instructors

Lise Getoor is a professor in the Computer Science Department at the University of Maryland, College Park. Her primary research interests are in machine learning and reasoning with uncertainty, applied to structured and semi-structured data. She also works on data integration, social network analysis and visual analytics. She has six best paper awards, an NSF Career Award, has served as associate editor for the Machine Learning Journal, JAIR, and TKDD, is elected member of the International Machine Learning Society board and AAAI Executive council, was PC co-chair of ICML 2011, and has served on a variety of program committees including AAAI, ICML, IJCAI, ISWC, KDD, SIGMOD, UAI, VLDB, WSDM and WWW. She received her Ph.D. from Stanford University, her M.S. from UC Berkeley, and her B.S. from UC Santa Barbara.

Ashwin Machanavajjhala is an Assistant Professor in the Department of Computer Science, Duke University. Previously, he was a Senior Research Scientist in the Knowledge Management group at Yahoo! Research. His primary research interests lie in data privacy, systems for massive data analytics, and statistical methods for information extraction and entity resolution. He is a recipient of the NSFCAREER award in 2013 and the ACM SIGMOD Jim Gray Dissertation Award Honorable Mention in 2008. He received his Ph.D. from Cornell University and a B.Tech in Computer Science and Engineering from the Indian Institute of Technology, Madras.