Mining Structures from Massive Text Data: Will It Help Software Engineering?

Jiawei Han

Abel Bliss Professor
Department of Computer Science
University of Illinois at Urbana-Champaign, USA
hanj@cs.uiuc.edu

Abstract—The real-world big data are largely unstructured, interconnected text data. One of the grand challenges is to turn such massive unstructured text data into structured, actionable knowledge. We propose a text mining approach that requires only distant or minimal supervision but relies on massive text data. We show quality phrases can be mined from such massive text data, types can be extracted from massive text data with distant supervision, and entities/attributes/values can be discovered by meta-path directed pattern discovery. We show text-rich and structure-rich networks can be constructed from massive unstructured data. Finally, we speculate whether such a paradigm could be useful for turning massive software repositories into multidimensional structures to help searching and mining software repositories.

BIOGRAPHY
Jiawei Han is Abel Bliss Professor in the Department of Computer Science, University of Illinois at Urbana-Champaign. He has been researching into data mining, information network analysis, database systems, and data warehousing, with over 800 journal and conference publications. He has chaired or served on many program committees of international conferences in most data mining and database conferences. He also served as the founding Editor-In-Chief of ACM Transactions on Knowledge Discovery from Data and the Director of Information Network Academic Research Center supported by U.S. Army Research Lab (2009-2016), and is the co-Director of KnowEnG, an NIH funded Center of Excellence in Big Data Computing since 2014. He is Fellow of ACM, Fellow of IEEE, and received 2004 ACM SIGKDD Innovations Award, 2005 IEEE Computer Society Technical Achievement Award, and 2009 M. Wallace McDowell Award from IEEE Computer Society. His co-authored book "Data Mining: Concepts and Techniques" has been adopted as a textbook popularly worldwide.