Seminar Series on Trustworthy Data Science and AI

SFU SIMON FRASER UNIVERSITY

Divesh Srivastava

Head of Database Research, AT&T
ACM Fellow; President of VLDB Endowment

Exploring and Analyzing Change: The Janus Project

Friday, February 10, 2023
10:30am – 12:00pm Pacific time
TASC1 9204, SFU Burnaby / Zoom

Register for online access: https://data.cs.sfu.ca/gbtu

Abstract

Data change, all the time. The Janus project seeks to address the Variability dimension of Big Data by modeling, exploring, and analyzing such change, providing valuable insights into the evolving real world and the ways in which data about it are collected and used.

We start by identifying technical challenges that need to be addressed to realize the Janus vision. Towards this end, we have extracted and worked with the histories of various structured datasets, including DBLP, IMDB, open government data, and Wikipedia, for which a detailed history of every edit is available. Our DBChEx (Database Change Explorer) prototype enables interactive exploration of data and schema changes, and we show how DBChEx can help users gain valuable insights by exploring two real-world datasets, IMDB and Wikipedia infoboxes.

Based on an analysis of the history of 3.5M tables on the English Wikipedia for a total of 53.8M table versions, we then illustrate the rich history of structured Wikipedia data: we show that tables are created in certain locations, they change their shape, they move, they grow, they shrink, their data change, they vanish, and they re-appear; indeed, each table has a life of its own. Finally, to help automatically interpret the useful knowledge harbored in the history of Wikipedia tables, we present recent results on two technical problems: (i) identifying Natural Keys, a particularly important piece of metadata, which serves as a primary key in tables over time and consists of attributes inherent to an entity, and (ii) matching tables, infoboxes and lists within a Wikipedia page across page revisions. We solve these problems at scale and make the resulting curated datasets available to the community to facilitate future research.

This is joint work with Tobias Bleifuß, Leon Bornemann, Dmitri Kalashnikov, and Felix Naumann.

Biography

Divesh Srivastava is the Head of Database Research at AT&T. He is a Fellow of the ACM, the President of the VLDB Endowment, co-chair of the ACM Publications Board, and on the Board of Directors of the Computing Research Association. He has served as PC co-chair of many international conferences including SIGMOD 2021, VLDB 2020 (Industrial), SIGMOD 2020 (Industrial), and ICDE 2019. He has presented keynote talks at several international conferences, and his research interests and publications span a variety of topics in data management. He received his Ph.D. from the University of Wisconsin, Madison, USA, and his Bachelor of Technology from the Indian Institute of Technology, Bombay, India.

All are welcome.
Enquires: Tianzheng Wang (tzwang@sfu.ca)

SFU Data Science Research Group
https://data.cs.sfu.ca