PittComputing&Information

Dissertation Defense

Doctor of Philosophy in Computing and Information

A NOVEL APPROACH FOR IMPROVING THE QUALITY OF DATA USING AGGREGATION MECHANISM by Shadi Al-khateeb

Date: February 8, 2021
Time: 1:00pm – 3:00pm
Place: https://pitt.co1.qualtrics.com/jfe/form/SV_3f1QpeahaAgWn7T

Committee:

- Vladimir Zadorozhny, Professor, Department of Information Sciences, School of Computing and Information
- Paul Munro, Associate Professor, Department of Information Sciences, School of Computing and Information
- Konstantinos Pelechrinis, Associate Professor, Department of Information Sciences, School of Computing and Information
- John Grant, Professor, Department of Computer Science, University of Maryland

Abstract:

Due to the inception of the big data applications, it is becoming increasingly important to manage and analyze large volumes of data. However, it is not always possible to efficiently analyze very big chunks of detailed data. Thus, data aggregation techniques emerged as an efficient solution for reducing the data size and providing summary of the key information in the original data. For example, yearly stock sales are used instead of daily sales to provide a general summary of the sales. Data aggregation aims to group raw data elements in order to facilitate the assessment of higher-level concepts. However, data aggregation can result in the loss of some important details in the original data, which means that the aggregation should be done in a creative manner in order to keep the data informative even if there is a loss in some details. In some cases, we may have only aggregated versions of the data due to the data collection constraints as well as high storage and processing requirements of the big data. In these cases, we need to find the relationship between aggregated datasets and original datasets. Data disaggregation is one solution for this issue. However, accurate disaggregation is not always possible and easy to utilize.

In this proposal, we introduce a novel approach to improve the quality of data to be more informative without disaggregating the data. We propose information preserving signature based preprocessing strategy. As well as an aggregation-based information retrieval architecture using signatures. We compensate the loss of details in the raw data by highlighting the most informative parts in the aggregated data. Our approach can be used to assess similarity and correspondence between datasets and to link aggregated historical data with most related datasets.