

NATIONAL UNIVERSITY OF SINGAPORE

School of Computing

C S S E M I N A R

Title: An Extreme-Value-Theoretic Foundation for Similarity Applications

Speaker: Dr. Michael E. Houle, Visiting Professor
National Institute of Informatics (NII), Japan

Date/Time: 18 August 2015, Tuesday, 02:00 PM to 03:30 PM

Venue: Executive Classroom, COM2-04-02

Chaired by: Dr Zimmermann, Roger, Associate Professor, School of Computing
(rogerz@comp.nus.edu.sg)

Abstract:

For many large-scale applications in data mining, machine learning, and multimedia, fundamental operations such as similarity search, retrieval, classification, clustering, and anomaly detection generally suffer from an effect known as the 'curse of dimensionality'. As the dimensionality of the data increases, distance values tend to become less discriminative due to their increasing relative concentration about the mean of their distribution. For this reason, researchers have considered the analysis of similarity applications in terms of measures of the intrinsic dimensionality (ID) of the data sets. This presentation is concerned with a generalization of a discrete measure of ID, the expansion dimension, to the case of continuous distance distributions. This notion of the ID of a distance distribution is shown to precisely coincide with a natural notion of the indiscriminability of distances, thereby establishing a theoretically-founded relationship among probability density, the cumulative density (cumulative probability divided by distance), intrinsic dimensionality, and discriminability. The proposed indiscriminability function is shown to completely determine an extreme-value-theoretic representation of the distance distribution. From this representation, a characterization in terms of continuous ID is derived for the notions of outlierness and inlierness of data.

Biodata:

Michael E. Houle obtained his PhD degree from McGill University in 1989, in the area of computational geometry. Since then, he developed research interests in algorithmics, data structures, and relational visualization, first as a research associate at Kyushu University and the University of Tokyo in Japan, and from 1992 at the University of Newcastle and the University of Sydney in Australia. From 2001 to 2004, he was a Visiting Scientist at IBM Japan's Tokyo Research Laboratory, where he first began working on approximate similarity

search and shared-neighbor clustering methods for data mining applications. Since then, his research interests have expanded to include dimensionality and scalability in the context of fundamental data mining tasks such as search, clustering, classification, and outlier detection. He has co-authored award-winning conference papers on outlier detection (Best Research Paper Award at IEEE ICDM 2010) and similarity search (Best Paper Award at SISAP 2014). Currently, he is a Visiting Professor at the National Institute of Informatics (NII), Japan.