School of Informatics and Computing
Colloquium Series

Ping Li

Dept. of Statistical Science, Faculty of Computing
and Information Science, Cornell University

Thursday, January 24, 2013

3:00-4:00pm

Dogwood Room, IMU

Topic: BigData: Probabilistic Methods for Efficient Search and Statistical Learning in Extremely High-Dimensional Data

Abstract: This talk will present a series of work on probabilistic hashing methods which typically transform a challenging (or infeasible) massive data computational problem into a probability and statistical estimation problem. For example, fitting a logistic regression (or SVM) model on a dataset with billion observations and billion (or billion square) variables would be difficult. Searching for similar documents (or images) in a repository of billion web pages (or images) is another challenging example. In certain important applications in the search industry, a web page is often represented as a binary (0/1) vector in billion square (2 to power 64) dimensions. For those data, both data reduction (i.e., reducing number of nonzero entries) and dimensionality reduction are crucial for achieving efficient search and statistical learning.

This talk will present two closely related probabilistic methods: (1) b-bit minwise hashing and (2) one permutation hashing, which simultaneously perform effective data reduction and dimensionality reduction on massive, high-dimensional, binary data. For example, training an SVM for classification on a text dataset of size 24GB took only 3 seconds after reducing the dataset to merely 70MB using our probabilistic methods. Experiments on close to 1TB data will also be presented. Several challenging probability problems still remain open.

Key references: [1] P. Li, A. Owen, C-H Zhang, On Permutation Hashing, NIPS 2012; [2] P. Li, C. Konig, Theory and Applications of b-Bit Minwise Hashing, Research Highlights in Communications of the ACM 2011.

Biography: Ping Li is an Assistant Professor in the Department of Statistical Science at Cornell University. His research interests include BigData, randomized algorithms, boosting and trees, information retrieval, etc. Ping Li won a prize in the Yahoo! 2010 Learning to Rank Grand Challenge. He is also a recipient of the ONR (Office of Naval Research) Young Investigator Award in 2009.

SCHOOL OF INFORMATICS
AND COMPUTING
INDIANA UNIVERSITY
Bloomington