PROGRAMME

The Final Oral Examination for the Degree of

DOCTOR OF PHILOSOPHY
Computer Science

Pui Kuen Fong
2008 University of Victoria MSc
2005 University of Victoria Hons. BSc

“Privacy Preserving Data Mining using Unrealized Data Sets—Scope Expansion and Data Compression”

May 2nd, 2013
1:00 pm
Engineering/Computer Science Bldg., Room 468

Supervisory Committee:
Dr. Jens Weber, Department of Computer Science, UVic (Co-Supervisor)
Dr. Alex Thomo, Department of Electrical and Computer Engineering, UVic (Co-Supervisor)
Dr. Aaron Gulliver, Department of Electrical and Computer Engineering, UVic (Outside Member)

External Examiner:
Dr. Yeye He, Microsoft Research

Chair of Oral Examination:
Dr. Don Vandenberg, Department of Physics & Astronomy, UVic
Abstract

In previous research, the author developed a novel PPDM method – Data Unrealization – that preserves both data privacy and utility of discrete-value training samples. That method transforms original samples into unrealized ones and guarantees 100% accurate decision tree mining results. This dissertation extends their research scope and achieves the following accomplishments: (1) it expands the application of Data Unrealization on other data mining algorithms, (2) it introduces data compression methods that reduce storage requirements for unrealized training samples and increase data mining performance and (3) it adds a second-level privacy protection that works perfectly with Data Unrealization.

From an application perspective, this dissertation proves that statistical information (i.e. counts, probability and information entropy) can be retrieved precisely from unrealized training samples, so that Data Unrealization is applicable for all counting-based, probability-based and entropy-based data mining models with 100% accuracy.

For data compression, this dissertation introduces a new number sequence – J-Sequence – as a mean to compress training samples through the J-Sampling process. J-Sampling converts the samples into a list of numbers with many replications. Applying run-length encoding on the resulting list can further compress the samples into a constant storage space regardless of the sample size. In this way, the storage requirement of the sample database becomes $O(1)$ and the time complexity of a statistical database query becomes $O(1)$.

J-Sampling is used as an encryption approach to the unrealized samples already protected by Data Unrealization; meanwhile, data mining can be performed on these samples without decryption. In order to retain privacy preservation and to handle data compression internally, a column-oriented database management system is recommended to store the encrypted samples.
Awards, Scholarships, Fellowships

2005-2006  University of Victoria Fellowships

2004      NSERC Industrial Undergraduate Student Research Award

Publications
