

## Doctoral Thesis Defense

Speaker:	Tongyuan Wang
Supervisor:	Dr. B. Desai
Examining Committee:	Drs. D. Goswami, N. Shiri, R. Soleymani, P. Valtchev
Title:	Measures and Adjustments of Pattern Frequency Distributions
Date:	Wednesday, April 21, 2010
Time:	10:30 a.m.
Place:	EV 1.162

### ABSTRACT

Frequent pattern mining over large database is fundamental to many data mining applications, while pattern frequency distribution plays a central role in pattern mining. Various approaches have been proposed for pattern mining with respectable computational performance. However, there is a lack of well established theory to pursue mining correctness. Under a systematic investigation, this thesis has identified a set of fundamental problems embodied in conventional mining approaches. The radical ones include the extensively used but ill formed frequentness measure “support”, and the full enumeration based pattern generation, which produces excessive number of patterns in an application. Most conventional mining approaches are based on these two imperfect measures, and hence the correctness of the related mining results are unavoidably affected and symbolized with overfitting and underfitting, probability anomaly, bias for generated against original observations, etc.. Even worse, these results are delivered to users without any refinement. Overcoming these drawbacks is challenging, since these problems are rather philosophical than computational and hence their resolution demands a well established theory to reform the mining foundations and to pursue graceful knowledge degeneration.

Based on the problems identified, this thesis first proposes a reformulation of the frequentness measure, which effectively resolves the probability anomaly and other related issues. To deal with the profound full enumeration mode, we first explore a set of properties governing raw pattern frequency distributions, such that a number of important mining parameters can be predetermined. Based on these explorations, an approach to adjust the raw pattern frequency distributions is established and its theoretical merits are justified. This refinement theory shows that unconditional pattern reduction is achievable before domain constraints are imposed. The thesis then presents a maximum likelihood pattern sampling model and strategies to realize the adjustment.

Findings presented in this thesis are based on known set theory, combinatorics, and probability theory, and they are theoretically fundamental and applicable to every item based or key words based pattern mining and the improvement of mining effectiveness. We expect that these findings would pave a way to replace the full enumeration pattern generation with selective generation mode, which would then radically change the state of the art of pattern mining.