DEPARTMENT OF COMPUTER SCIENCE

M.Sc. Thesis Defense
Mining Frequent Patterns from Uncertain Data with MapReduce
by
Yaroslav Hayduk

Date: March 26, 2012 at 3 pm
Place: E2-155 EITC

Abstract

Traditional frequent pattern mining from precise data discovers frequently occurring patterns from traditional transaction datasets. As for frequent pattern mining of uncertain data, it allows data analysts to mine frequent patterns from probabilistic datasets, within which each item is associated with an existential probability, representing the likelihood of the presence of the item in the transaction. When compared with precise data, the solution space for mining uncertain data is often much larger due to the probabilistic nature of uncertain datasets. Thus, algorithms for mining uncertain data usually take substantially more time to execute. Since tree-based frequent pattern mining algorithms (which I am focusing on in my M.Sc. research) build a tree-based compressed representation of the input database, it may not fit into main memory in cases where the input dataset is large.

Recent research shows that MapReduce yields significant performance gains for data mining algorithms, which can be mapped to the map and reduce execution phases of MapReduce. One of the most attractive features of MapReduce is fault-tolerance, which permits detecting and restarting failed jobs on working machines in the MapReduce network. In this M.Sc. thesis, I explore the feasibility of applying MapReduce to frequent pattern mining of uncertain data.