OPEN SOURCE GRID (OSG) SUMMER SCHOOL POSTER SUPPLEMENTAL

ERICH A. PETERSON

1. Distributing the Problem

As the computational complexity of mining probabilistic frequent itemsets (PFIs) in uncertain data (along with the classical problem of mining frequent itemsets) is an exponential one, with the number of attributes in the database, we seek to distribute this complexity among nodes in a Condor cluster. One way in which that can be accomplished, is by letting one job find all PFIs which start with a given attribute. For example, if a given database contains attributes A, B, C, and D, then four jobs would be created—each one taking one of the aforementioned attributes and finding only those PFIs that start with their given attribute. Thus, job one given attribute A might find the itemsets A, AB, and AD to be probabilistically frequent; while job two given attribute B might find B and BD.

When the problem is split in the aforementioned way, each job is independent of the rest and need not communicate with other jobs.

2. Resources and Data Movement

The problem of mining PFIs is not married to or suited to any particular environment, being HPC or HTC; thus either one can be used, and is more dependent on the user's resources and expertise. However, because my research is more interested in the actual creation of algorithms and the theory behind them, and because that research is more interested in benchmarking the execution of the algorithm, an environment that is free of other users is preferable. Whereas, an end-user of the algorithm would be fine executing it within a shared environment.

Because of my limited knowledge of HPC, and the education received at the OSG Summer School, we will in all likelihood be using Condor. Further, because that my institution does not currently have a Condor cluster, I believe I have two main options: 1) Construct a new Condor cluster at my institution from scratch; 2) Request to use some of OSGs resources.

The size of our testing databases are in megabyte sizes. Therefore, it should be feasible to either move a copy of the full database to each job's location, or if using the Standard

Date: 08/07/2011.

Condor Universe to use remote I/O, giving all processes access to the database on the submit node. Access to the original database, represented as an FP-Tree is the only resource needed by each job.

As for workflow concerns, each job will output and return to the submit node a list of PFIs it found. The submit node, if needed, could append these output files into one larger file. Of course, this simple workflow could be accomplished with services provided by DAGMan.

3. Conclusion and Future Work

By distributing the problem as previously mentioned, researchers and data mining practitioners will hopefully be able to mine databases for PFIs which have many more attributes than they have been able to on a simple machine. For future work, I hope to follow through on this plan and benchmark just how scalable the algorithm can be and publish those results.