600.415 - Project Phase 1

Adam Gerber Juri Ganitkevic

November 23, 2010

(1) Adam Gerber and Juri Ganitkevic

(2) In this project we will expand upon a novel optimization algorithm for multi-table intersections. The actual domain of the data and database design are therefore of secondary nature only. We are currently planning on using a sufficiently decomposed database of Twitter posts and user information to illustrate gains in join performance achieved by our algorithm.

(3) Since we will not be using an estabished SQL database, but rather sample joins performed within our own code, our sample queries will be limited to natural joins over a large number of tables.

(4) n/a

(5) n/a

(6) We currently plan on expanding upon an existing database of tweets by separating out currently textual features such as "location" into their own tables, thereby creating a larger number of tables for us to multi-join.

(7) Our reports will be detailing the number and size of relations involved in the join as well as the methods used to speed up the muti-join.

(8) We focus on join/intersection algorithms and optimization methods for those. A few approaches we have in mind are:

- Use a hiearchical key structure to efficiently perform intersections over k ordered sets.
- Use Bloom filters to precisely and efficiently estimate overlap between key sets.

(9) Our software will be written in Java and should run on any recent machine without any additional libraries. Primarily, we will conduct (and can demonstrate) our experiments on our own desktop and laptop machines (i.e. up to dual-core i7 with 4GB RAM). We also intend to run a few large-scale experiments on the CLSP cluster machines with 128GB RAM.