Week 3 Homework

Ben Arancibia

September 12, 2015

2.5.1

What is the communication cost of each of the following algorithms, as a function of the size of the relations, matrices, or vectors to which they are applied?

(a) The matrix-vector multiplication algorithm

Communication cost: $O(r * c)$

The matrix-vector multiplication algorithm $M * v$ creates a key value pair for each entry in the matirx.

The communication cost is $O(r * c)$ where r and c are the number of rows and columns of the matrix.

(b) The union algorithm

Communication cost: $O(r + s)$

In the union of R and S, the mapper funtion passes key value pairs for each entry in R and S. The communication cost is the total number of entries in R plus the total number of entries in S or $O(r + s)$.

(c) The aggregation algorithm

Communication cost: Number of tuples (a, b, c)

The communication cost of $R(A, B, C)$ is the number of tuples (a, b, c) in the relation R.

2.6.1

Describe the graphs that model the following problems.

(a) The multiplication of an n $\times$ n matrix by a vector of length n.

One reducer per output cell and each reducer computes $SUMj(A[i, j] * B[j, k])$

(b) The natural join of R(A,B) and S(B,C), where A, B, and C have domains of sizes a, b, and c, respectively.

The map function outputs the same value as its input, but changes the key to always be the join attribute b.

After the MapReduce system groups together the intermediate data by the intermediate key, use the reduce function and do a nested loop join over each group.

All the values from each group have the same join attribute so need check which relation each tuple comes from, so that don't join a tuple from R with itself.

(c) The grouping and aggregation on the relation R(A,B), where A is the grouping attribute and B is aggregated by the MAX operation. Assume A and B have domains of size a and b, respectively.

The graph consists of two steps, first local aggregation then a global aggregation. These steps basically correspond to Map and Reduce operations. Local aggregation is optional and raw records can be emitted, shuffled, and aggregated on a global aggregation phase.