Pregel: A System for Large-Scale Graph Processing

Miles Welsh
11/24/2013

Malewicz, Grzegorz, Matthew H. Austern, Aart J. Bik, James C. Dehnert, Ilan Horn, Naty Leiser, and Grzegorz Czajkowski. "Pregel: A System for Large-Scale Graph Processing." 135-45. Web. 24 Nov. 2013.

Main Idea

- Pregel was designed by Google to process large graphs and output them
- The output is determined by a user-defined algorithm
- This algorithm is run and is checked each step by Pregel. When the algorithm is complete, it terminates and the output is received.

How the idea is Implemented

- The idea is implemented using messages that go between vertices during a process called supersteps.
- These vertices can read the previous message(S-1) and send a message(S+1) to another vertex.
- Once the desired outcome has been achieved the vertex goes into a halted state. Communication ceases when all vertices have been halted
- Graphs can be mutated through partial ordering. Removing edges are done before removing vertices, while adding vertices are done before adding edges. The rest of the mutation is handled by the user.
- Fault tolerance is achieved through “Checkpointing”. Ever superstep is saved to a local storage system. Errors are detected by pings sent from the master to the workers. If the ping is not sent then the workers shut down and the previous superstep is reloaded.

My Analysis and Thoughts

- Pregel is extremely efficient and allows for huge graphs to be processed easily.
- I wondered how the were going to handle mutation, but putting vertices above edges is genius because without a vertex there can be no edges.
- My one question is with the fault tolerance, if there is a ping sent at every superstep, then does every worker receive the ping at the same time? Do workers always finish their jobs at the same time? The paper doesn't talk about this.

Advantages and Disadvantages

- **Advantages**
- Allows for scalable graphs
- Easy to modify vertices
- As shown in Figure 7 and 8, this model is very efficient.
- **Disadvantages**
- API does not look like it can be changed
- The process uses RAM which limit the amount of data that can be used
-

Potential Uses

- Linkedin(Job networking)
- Facebook(Social networking)
- Disease Vectors(Keeping track of potential contagious individuals)