15-400 Report 2

Lichen Zhang

February 12, 2020

1 Major Changes

So far, no big changes. I have to mention that one of my friends who is working on societal computing wants to work on some graphs built from sinaweibo in China, I am quite interested in his plan, and currently we are evaluating whether it's possible for us to build such graph: it has almost 1 billion users, so we are not sure to what extent we can do that. If it looks plausible, I might switch gear to apply those algorithms to these real-life graphs.

2 Accomplishments

I have almost finished the vanilla version of CRD algorithm, due to the complication of `networkx` package and I was trying to meet the exact complexity bound described in paper, I haven't completely finished the goal for this milestone.

3 Meet Milestones

Previously I was planning to finish the implementation of first version and completed test for it, however, currently I've only finished 80% of the implementation.

4 Surprises

It is in fact quite hard to implement a theoretical algorithm, especially they use some exotic data structures to meet their cost bound. In CRD algorithm, they propose the use of a priority list data structure that has been dedicated to this specific task, so it takes me quite a while to finish and debug it. Also, Andy has finished a version that simplifies the algorithm by using a standard priority queue, and it seems the algorithm gives some unsatisfactory results on our constructed testing graph.

5 Look Ahead

The good thing is since I have put most of the complexities in using libraries and self-implemented data structure, it will be much easier to integrate into the GraphSAGE architecture, since they setup all their tests based on `networkx`. Hopefully, before next meeting I can integrate it into GraphSAGE.

6 Revisions

Currently no big changes. If the CRD algorithm we are working on really does not perform very well, I will quickly switch gear into other local algorithms, such as Nibble or Page-Rank Nibble, which is considered to be easier to implement. Also, collaborate on clustering sinaweibo graph remains a choice.

7 Resources

If I stick on current project, then there is no resource problem. If I start to work on the sinaweibo one, we might encounter several problems: first of all, in order to construct some graph, we have to use the API provided by sina, which, due to special conditions in China, might not be easy to use, and we might not be able to get a desired result. Second, suppose we can really configure that API and have some fancy graph built, then it remains a big problem that where do we store such graph, since weibo has a prohibitively large group of users.