Graph Analytics

CloudSuite 2.0 Benchmark Suite

We use the GraphLab machine learning and data mining software for the graph analytics benchmark. We implemented TunkRank on GraphLab, which provides the influence of a Twitter user based on the number of that user's followers. Although GraphLab can perform distributed graph processing, in this document, we provide instructions for a single-machine setup. Instructions for cluster deployment can be found at the GraphLab website.

Prerequisite Software Packages

1. GraphLab 2.1
2. TunkRank implementation (provided in the the CloudSuite benchmark package)
3. gcc.x86_64, zlib.x86_64, openmpi.x86_64, openmpi-devel.x86_64

Download the Graph Benchmark

Building TunkRank on GrapLab

1. Untar the GraphLab package and configure.

   ```
   tar zxvf graph.tar.gz
   cd graph-release
   ./configure
   ```

2. Build TunkRank.

   ```
   cd release/toolkits/graph_analytics
   make tunkrank
   ```

Running TunkRank

We provide three options for the graph to be processed with TunkRank.

1. The first option is to use the synthetic graph generated by GraphLab with an out-degree power law of N vertices with a particular alpha parameter (alpha=2). As the generated graph size depends on the input argument N, this option is convenient when there is a need to scale the dataset.

   For example, to run TunkRank on a graph with 10M vertices utilizing 2 cpus:

   ```
   ./tunkrank --powerlaw=10000000 --ncpus=2 --engine=asynchronous
   ```

2. The second option is to use the Twitter dataset with 41M vertices (Twitter users), which was extracted in 2009. The dataset size is 25GB and GraphLab requires around 45GB heap memory to process this graph on a single machine.

   As the original dataset format is different from the GraphLab input format, the appropriate dataset can be generated running:

tar zxvf twitter_rv.tar.gz
cat twitter_rv.net | awk '{print $2, $1}' > twitter_data_graplab.in

To run TunkRank using this graph as the input:

./tunkrank --graph=/path/to/twitter_data_graplab.in --format=tsv --ncpus=2 --engine=asynchronous

3. The third option is to use a smaller Twitter dataset with 11M vertices (Twitter users). The dataset size is 1.3GB and GraphLab requires around 4GB heap memory to process this graph on a single machine.

   As the original dataset format is different from the GraphLab input format, the appropriate dataset can be generated running:

   unzip Twitter-dataset.zip
   cd Twitter-dataset/data
   cat edges.csv | awk -F"," '{print $1, $2}' > twitter_small_data_graplab.in

   To run TunkRank using this graph as the input:

   ./tunkrank --graph=/path/to/twitter_small_data_graplab.in --format=tsv --ncpus=2 --engine=asynchronous