Practical Experience with Social Network Data for Analysis

Péter Molnár
Department of Computer and Information Science
Clark Atlanta University
223 James P. Brawley Dr. SW
Atlanta, Georgia 30314
pmolnar@cau.edu

ABSTRACT

The objective of this project is to understand the dynamics of social impact on users in social networks like Twitter. We define the impact that one user has on another based on the frequency and type of communications between them. We expect that only a very small portion of all Twitter users actually have those interaction. The majority of tweets will never be retweeted, and most users on Twitter are hardly ever mentioned by someone else. Therefore, we need a huge amount of data to extract a significant number of users for our research.

In our initial phase, we collected close to 170 million tweets over a period of less than three month, leading up to the 2012 presidential elections. On certain days, we received about 4 million tweets per day. To specify the connection to the Twitter Public Stream API, we choose keywords pertaining current political topics, as well as terms that relate to student life.

The technical issue that we encouterd was neither bandwidth, nor storage, but the performance of our SQL database. The SQL schema did not work as expected. While queries on single tables, like ranking hash tags from a list of about 40 million toke about twenty minutes, queries on multiple joint tables with aggregation needed to be broken up into multiple queries creating additional tables and indexes.

The poster presents alternative approaches for gathering and processing the data. We compare the implementation and performance of various NoSQL databases, including MongoDB and Neo4j, and other forms of data storage to make use of parallel distributed processing. Implementations of popular algorithms, including community identification, page ranking, and collaborative filtering, serve as benchmark.

Keywords

social network analysis, NoSQL databases, graph analysis