SciDB for Clustering Analysis of MovieLens Rating Data

Wusuo Liu and Chia Ying Lee

# Overview and Summary of Project

Many scientific applications, such as in genomics and finance, involve performing complex linear algebraic operations, such as matrix multiplication and singular value decomposition, on massive amounts of data, often on the order of terabytes. These mathematical operations scale poorly, typically for many linear algebraic operations, and therefore pose a significant challenge.

SciDB is a database with efficient data management and distributed architecture, designed specifically for fast mathematical and statistical operations on a massive scale. This project demonstrates the capabilities of SciDB for mathematical and statistical operations for big data, with applications to clustering analysis of movie ratings data.

Our big data application is to use [MovieLens](https://movielens.org/) ratings data to cluster users based on which similar movies they liked. We perform two clustering methods, in order to demonstrate two key linear algebra operations of SciDB:

1. Cluster using correlation: demonstrates sparse matrix multiply
2. Cluster using Principle Component Analysis: demonstrates SVD

We also use the scidb R package, which provides an interface between the database and R, allowing import of data into R for visualization.

In this report, we will first describe the basic data structures and query languages of SciDB, and its installation. Then we shown how to download and load the data into SciDB, and illustrate some basic statistical operations (e.g. avg, max). Finally, we demonstrate the linear algebra operations in SciDB in the context of clustering analysis.

### Youtube URLs

* 2 min video: <https://youtu.be/1o5jLqXAXZs>
* 15 min video: <https://youtu.be/qlwy4fOXLvk>