Movie Recommender System using Map-Reduce

Team Research Investigation Presentation 1
Team Malibu

Presented By:
Sarvadeep Singh Bindra
Pratiksha Mishra
Overview

- Concept of Map Reduce
- Concept of Collaborative Filtering
- Sequential Algorithm
- Parallel Algorithm
Map Reduce

- Most commonly used programming model for large dataset, problems that needs to be solved on distributed systems, parallel computing.
- In MapReduce, the data is broken down to smaller data set, which is processed separately and the results of these smaller of dataset are sum-reduced and mapped back to provide final outcome of processed data.
- Data set used: movielens.com
  - 100,000 ratings from 1000 users on 1700 movies
Collaborative Filtering

- Technology used in recommendation systems
- Working:
  - Each user in a CF system rates the items that they have experienced.
  - CF system matches the user with people who share similar interests or tastes with him/her.
  - Ratings from those likely-minded people are used to generate recommendations for the given user
Sequential Algorithm

Step 1: Calculate the similarity matrix $S(m\times m)$ using rating matrix $R$. Pearson Correlation is adopted as the similarity measure. Each element $(u_1, u_2)$ in $S$ denotes the Pearson Correlation between user $u_1$ and user $u_2$.

Step 2: Find $USER_K$ nearest neighbors for user $u$. The larger the correlation, the more similar the two users are.

Step 3: Predict the rating values for items unrated yet by user $u$. The predictive rating values are obtained by the weighted sum of the items from $u$’s $USER_K$ nearest neighbors.

Step 4: Recommend top $ITEM_K$ items to user $u$ by the predictive values.

Large Dataset on Single Core -> Time Consuming
How we intend to parallelize?

- Split the dataset equally across cores
- Perform sequential algorithm on every core; the output will give top k/n on every core
- Summing them up to get top k rating values
  - Suppose we divide dataset on 5 cores, we need to find top 5 users.
  - Each core will give top 5 users as output using MAP REDUCE and summing them will give top 25 users
  - Again, using we will find top 5 users from 25 users by using MAP REDUCE
- Recommending the active user about interest
Reference

Any questions?