Item-based Collaborative Filtering for Recommendation

by Vishal Doshi

The MapReduce application is packed in 659391383_ASSIGN3.zip. Extract and import in Eclipse to view the source code. NOTE: Read code comments for better understanding.

Class and files:

IBCFHadoop: Driver class for the MapReduce Application. Get the top 100 pairs in a file "movieDetails/top100.txt"

IBCFMapper1: make user the key *[(Alice,Matrix,5) (Alice, Alien, 1)=> Alice (Matrix,5) Alice (Alien,1)]*

IBCFReducer1: create inverted index*[Alice (Matrix,5) Alice (Alien,1) => Alice (Matrix,5) (Alien, 1)]*

IBCFMapper2: emit all cooccurred ratings *[Alice (Matrix,5)(Alien,1)(Inception,4) =>Matrix,Alien (5,1) Matrix,Inception (5,4) Alien,Inception (1,4)]*

IBCFReducer2: compute similarities (Pearson's Correlation) *[Matrix,Alien (5,1) Matrix,Alien (4,3) =>Matrix,Alien (-0.47)]*

IBCFMapper3: Will club and sort the final output of Job2.

MovieRating.java: a Comparable class which sorts the list as per the movieId

Instructions to run:

Assumption: You have imported the eclipse project and exported it as a .jar file name "ibcfhadoop.jar" on your Desktop.

Step 1: Download the MovieLens 100K data from: http://grouplens.org/datasets/movielens/ *[Required files are also provided in the package]*

Step 2:Extract and copy **u.data** and **u.item** files on you Desktop

Step 3: Create 'ibcfInput' and 'movieDetails' directory in HDFS

- **./bin/hadoop fs -mkdir ibcfInput**
- **./bin/hadoop fs -mkdir movieDetails**

Step 4: Move the files from to Desktop to HDFS

- **./bin/hadoop fs -put ~/Desktop/u.data ibcfInput**
- **./bin/hadoop fs -put ~/Desktop/u.item movieDetails**

Step 5: Export the jar file to desktop and Run .jar file

- **./bin/hadoop jar ~/Desktop/ibcfhadoop.jar IBCFHadoop ibcfInput ibcfOutput1 ibcfOutput2 ibcfOutput3**

[4 parameters - 1: input directory, 2: output directory for job 1, 3: output directory for job 2, 4: output directory for job 3]

Step 6: Top 100 movie pairs in terms of similarity would be in: **movieDetails/top100.txt** in HDFS