Relevance Graph

Collaborative-based approach

Nhan Nguyen

# The viewing history data

The current user viewing history data will be used as input for collaborative-based algorithm. Current history data are aggregated from several platforms, including STB, web, mobile, etc. Currently, we are only interested in data from AMS STB as it is the most reliable source of user viewing history data. We extract events from AMS STB from the UVH data and a summary can be seen in the below table.

|  |  |
| --- | --- |
|  | After |
| AccountID (AMS STB) | 155,289 |
| Events | 2,351,792 |
| Movies | 15,254 |

The summary statistics for the number of movies per STB:

Min. 1st Qu. Median Mean 3rd Qu. Max.

1.00 2.00 5.00 15.14 13.00 1214.00

We are removing STB that have very high movie viewing. We start with a threshold of the 3rd quartile plus 5.0 times the IQR. So we would eliminate STB with more than 68 movies. There are 7558, ~ 4.9%, STB like that.

The summary statistics of the number of STB per movies:

Min. 1st Qu. Median Mean 3rd Qu. Max.

1. 2.0 15.0 154.2 119.0 8266.0

Summary statistics (movies per STB and STB per movies) after removing high movie viewing account:

Min. 1st Qu. Median Mean 3rd Qu. Max.

1.000 2.000 4.000 9.147 11.000 68.000

Min. 1st Qu. Median Mean 3rd Qu. Max.

1.0 2.0 12.0 105.8 67.0 7119.0

The following charts look at the percentage of movies in y-axis and the number of user on x-axis. So almost 85% of the movies has fewer than 1000 viewers (< 0.15% of total 745,670 accounts).



The next chart looks at the percent of users that view movies. It shows that 80% of total accounts view less than 25 movies.



# Collaborative-based results

We ran the collaborative-based algorithms on the data set. The parallel design of the algorithm was presented in a previous report. The implementation is in RevoR and run on Hadoop MapReduce.

Number of movie-pairs with zero cosine similarity: 222344676

Summary of non-zero cosine values:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Min. | 1st Qu. | Median | Mean | 3rd Qu. | Max. |
| 0.0004854 | 0.0060800 | 0.0106300 | 0.0212800 | 0.0193400 | 1.0000000 |

The next chart looks at the distribution of (non-zero) similarity values.



More than 90% of cosine values are below 0.02.

If we pick a threshold of similarity value 0.05, there are 682702 similarity higher than this threshold out of 10M values. A summary statistics of similarity values over all the movies are as below:

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Threshold | Min | 1st Qu. | Median | Mean | 3rd Qu. | Max |
| 0.05 | 0.00 | 12.00 | 39.00 | 44.76 | 65.00 | 352.00 |
| 0.1 | 0.00 | 4.00 | 13.00 | 17.75 | 24.00 | 202.00 |
| 0.15 | 0.00 | 1.00 | 6.00 | 10.40 | 13.00 | 178.00 |

