Relevance Graph

Collaborative-based approach

Nhan Nguyen

# The viewing history data

The current AMS dataset contains accounts that have multiple tokenid. Therefore, viewing history of such accounts are being partitioned to several parts corresponding to their tokenids. Even though such issue has been addressed, it has not been resolved totally. We need to clean our dataset by removing all accounts that have multiple tokenid.

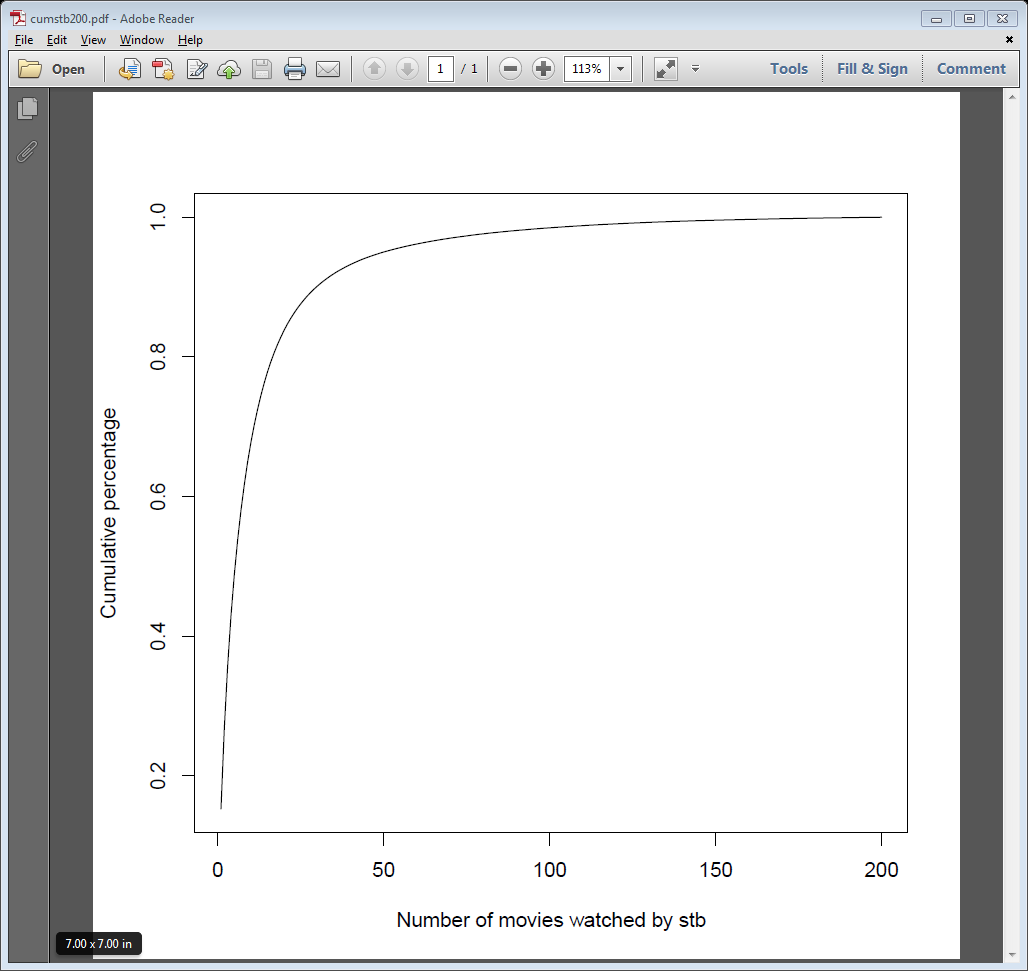
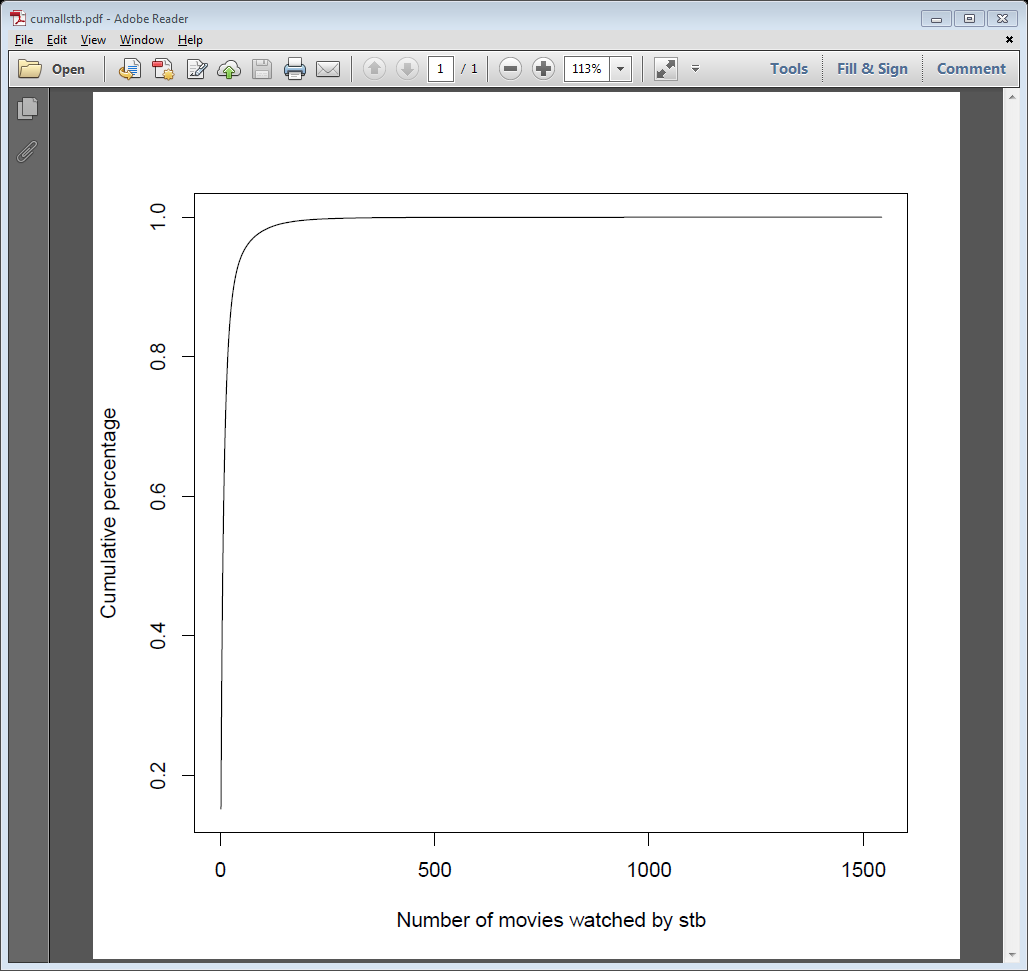
|  |  |  |
| --- | --- | --- |
|  | Before | After |
| AccountID (STB) | 873,052 | 745,670 |
| Events | 13,884,713 | 10,530,792 |
| Movies | 20,448 | 20,448 |

The summary statistics for the number of movies per STB:

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

1.00 2.00 6.00 14.12 14.00 1543.00

We are removing STB that have very high movie viewing. We start with a threshold of the 3rd quartile plus 1.5 times the IQR. So we would eliminate STB with more than 32 movies. There are 70481, ~ 9.4%, STB like that. As this is a significant amount of STB, we cannot consider all of them as outliers and remove them from the dataset. We are now looking further to the distribution of the STB by number of viewing movies to help distinguish STB with very high movie viewing. The next charts looks at the cumulative distribution function of the STB by movie viewings, with the right one zoom into x-axis value of 0-200.



The summary statistics of the number of STB per movies:

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

1 4 24 515 254 55400

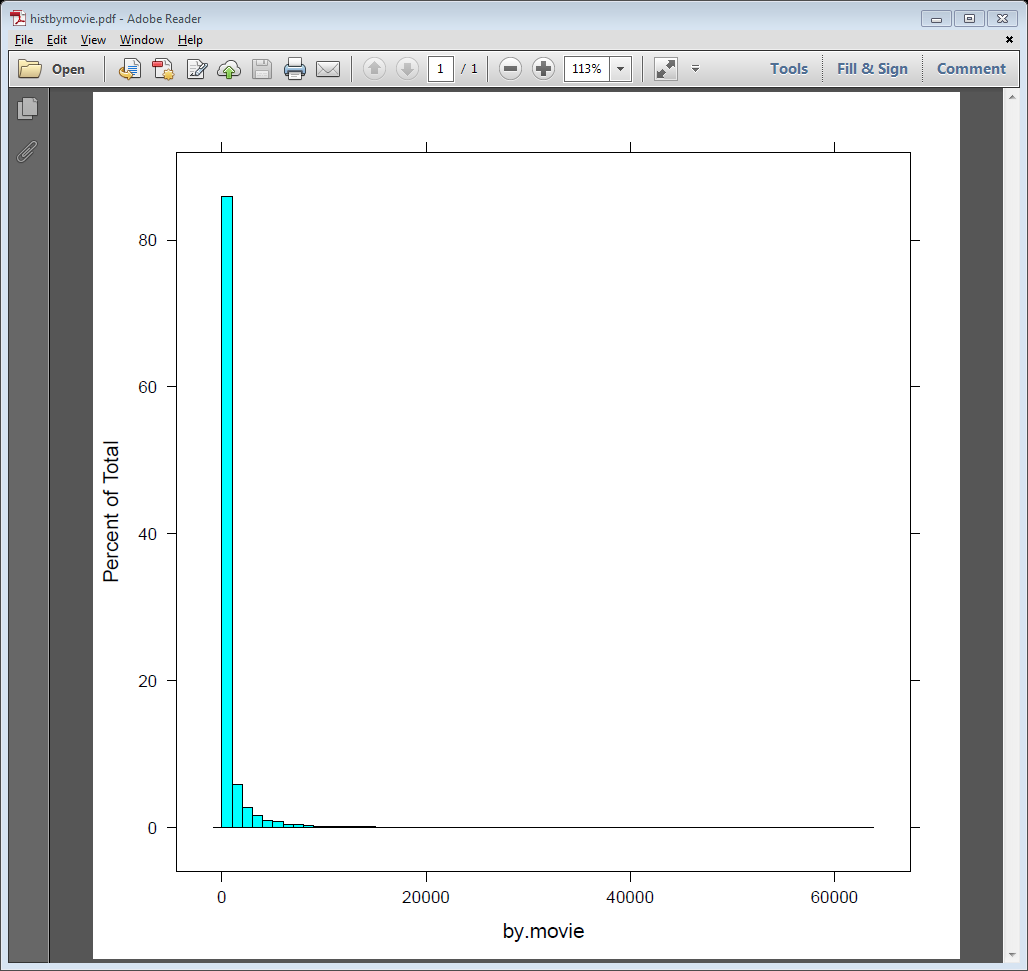
Summary statistics after removing high movie viewing account:

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

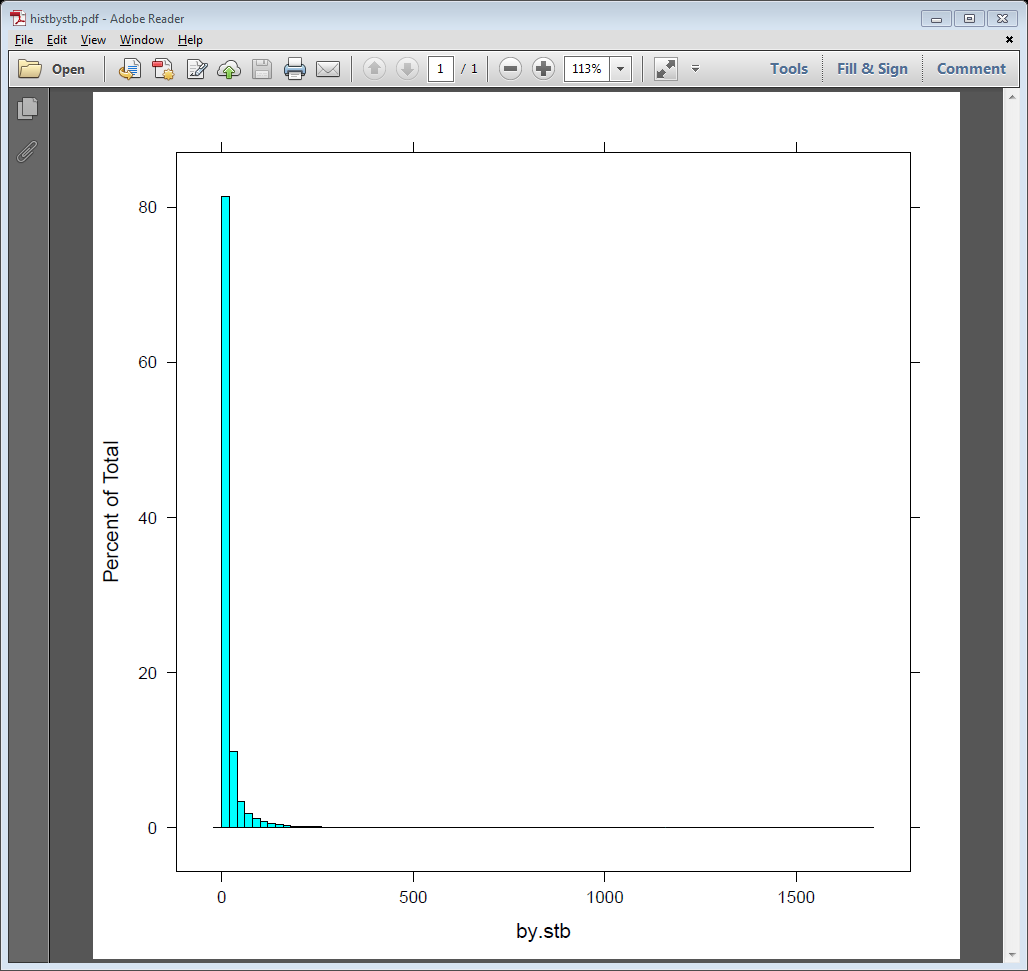
1.000 2.000 5.000 7.558 11.000 32.000

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

1.0 3.0 15.0 307.2 94.0 46620.0

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.



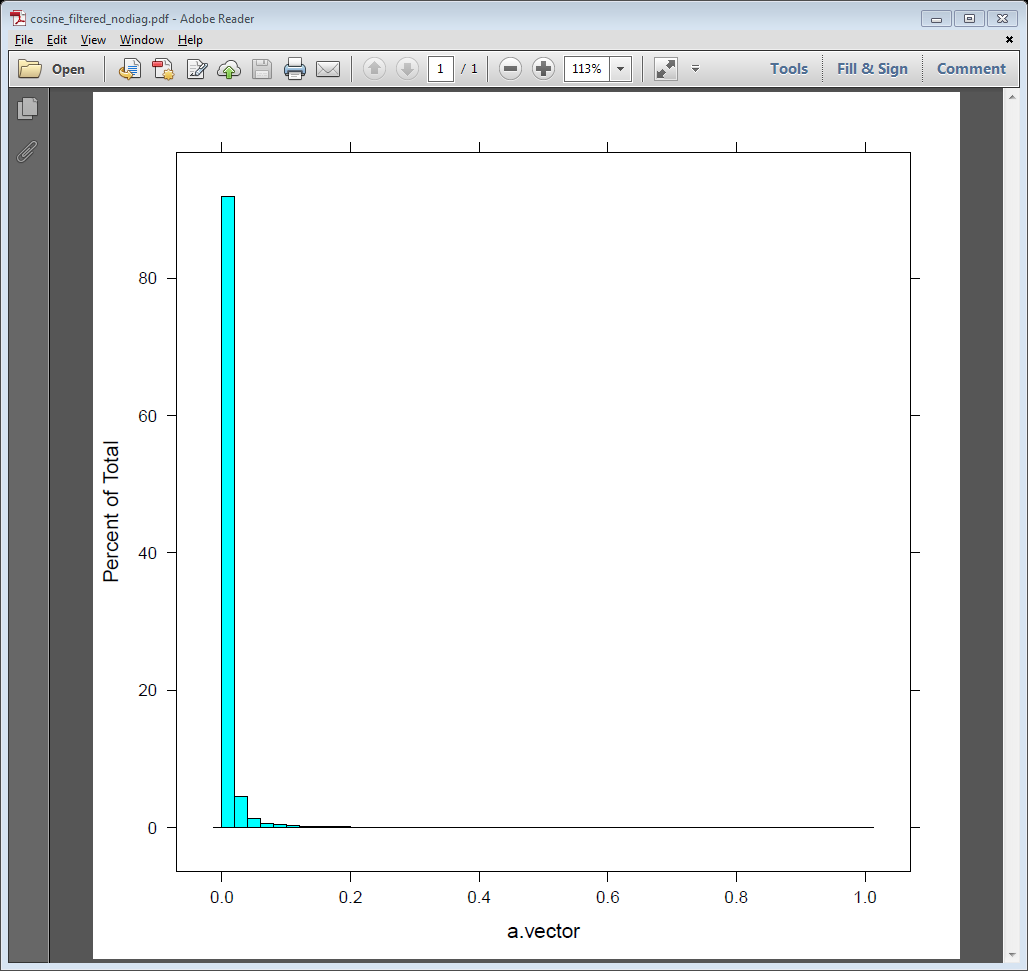
# Collaborative-based results

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.

The summary statistics on the cosine similarity value:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Min. | 1st Qu. | Median | Mean | 3rd Qu. | Max. |
| 0.0002259 | 0.0022370 | 0.0040320 | 0.0105600 | 0.0079330 | 1.0000000 |

The next chart looks at the distribution of similarity values.



More than 90% of cosine values are below 0.02.

If we pick a threshold of similarity value 0.05, there are 342082 similarity higher than this threshold out of 11M values. A summary statistics of similarity values over all the movies are as below:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Min | 1st Qu. | Median | Mean | 3rd Qu. | Max |
| 1.00 | 3.000 | 13.00 | 16.73 | 24.00 | 180.00 |

