Amazon Access Samples Analysis

Shortcomings of Recommender Systems and Alternative Approach

xxx, 28 of August 2023

Problem Description

Predict access of a specific user based on all-user/all-access table

- Sparse, highly unbalanced dataset (only 0,05% of values are one) with ~9000 features

Access

User

```
[[0 0 0 ... 0 0 0]
 [0 0 0 ... 0 0 0]
 [0 0 0 ... 0 0 0]
 ...
 [0 0 0 ... 0 0 0]
 [0 0 0 ... 0 0 0]
 [0 0 0 ... 0 0 0]]
```

- Preprocessing via cleaning (removal of outliers), transformation (differentiation no-access and no-data) and correlation analysis to reduce problems in model building

Recommender System

Using Jaccard and Cosine Similarity

- Similarity Metrics between single users yield a similarity matrix as basis for the prediction of the access status.

Access

User

```
[[0 0 0 ... 0 0 0]
 [0 0 0 ... 0 0 0]
 [0 0 0 ... 0 0 0]
 ...
 [0 0 0 ... 0 0 0]
 [0 0 0 ... 0 0 0]
 [0 0 0 ... 0 0 0]]
```

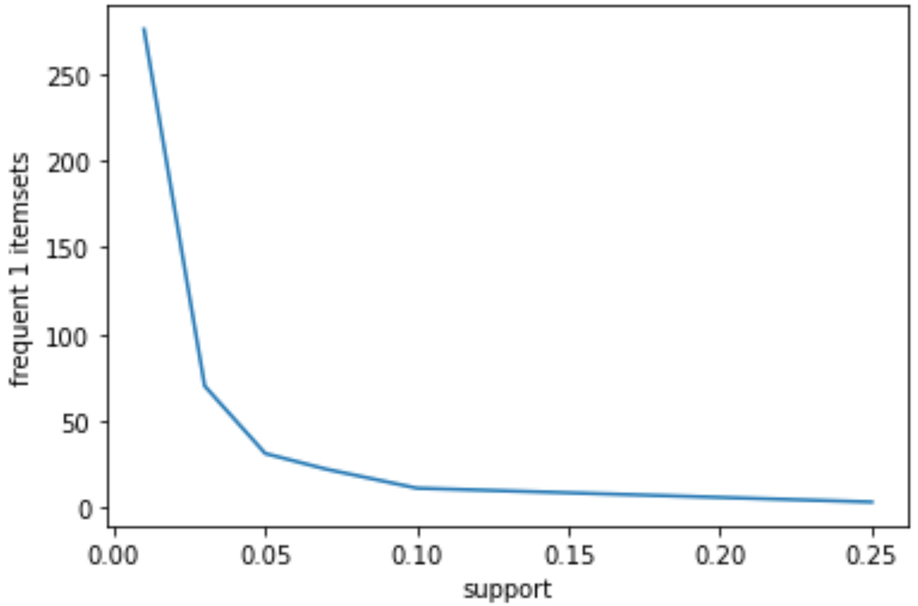
User

```
[[1.         0.94124492 0.93165308 ... 0.93909368 0.93200183 0.93225557]
 [0.94124492 1.         0.9425761  ... 0.94264793 0.93960254 0.94161049]
 [0.93165308 0.9425761  1.         ... 0.9371506  0.93367911 0.93393078]
 ...
 [0.93909368 0.94264793 0.9371506  ... 1.         0.94287965 0.93912115]
 [0.93200183 0.93960254 0.93367911 ... 0.94287965 1.         0.93750286]
 [0.93225557 0.94161049 0.93393078 ... 0.93912115 0.93750286 1.        ]]
(5166, 5166)
```

Alternative Approach

Association Rule Mining via Apriori Algorithm

- Frequent itemsets are calculated in order starting with frequent 1 sets
- Support value decides over minimum number of user in a combination of permission groups to be considered frequent



Results

Recommender System not suitable for highly unbalanced data, Association Rule Mining yields valuable results

- Recommender System predicts always zero for both Jaccard and Cosine Similarity, because of the sparsity and therefore imbalance in the input data

	Predicted no-access	**Predicted access**
Actually no-access	780915	0
Actually access	6900	0

- Association Rule Mining identifies the following maximum frequent itemsets with a support of 3%, so a user with access 427 is highly likely to have accesses 830, 1102, 1159 and 1268 as well, etc.

427	427	427	805	830	855	855	855
830	855	904	814	904	899	904	913
1102	1102	1102	1102	1102	1102	1102	1102
1159	1159	1159	1159	1159	1159	1159	1159
1268	1268	1268	1268	1268	1268	1268	1268