The Pattern of Search and Click v1

dgliu

2018-07-16

1 Feature Extraction

1.1 Session Level Features

- Queries per Session
- Level1 Categories per Session
- Leaf Categories per Session
- Clicks per Session

1.2 Query Level Features

- Query Length (mean)
- Initial Query Length
- Consecutive Queries Similarity (mean, Jaccard index)
- Head&Tail Queries Similarity (Jaccard index)
- Consecutive Queries Similarity (mean, Embedding Vector)
- Head&Tail Queries Similarity (Embedding Vector)

1.3 Click Level Features

- Search Depth per Query
- Position per Query
- Clicks per Query
- Level1 Categories per Query
- Leaf Categories per Query

1.4 Potential Features

- Spaces per Query
- Shopper per Query
- Shopper per Session
- Entropy per Query
- Entropy per Session
- Add Cart per Query
- Add Cart per Session
- Add Collect per Query
- Add Collect per Session
- ...

2 Results

The features used in the section include $S_1, S_2, S_3, S_4, Q_1, Q_2, Q_3, Q_4, C_1, C_3, C_4$ and C_5. We choose about 50 million samples by judging whether uniform distributed random numbers less than 0.05. Using the k-means algorithm on the PAI platform, we get the preliminary clustering results. The parameter settings follow the default settings, except centerCount=6 and accuracy=0.001.

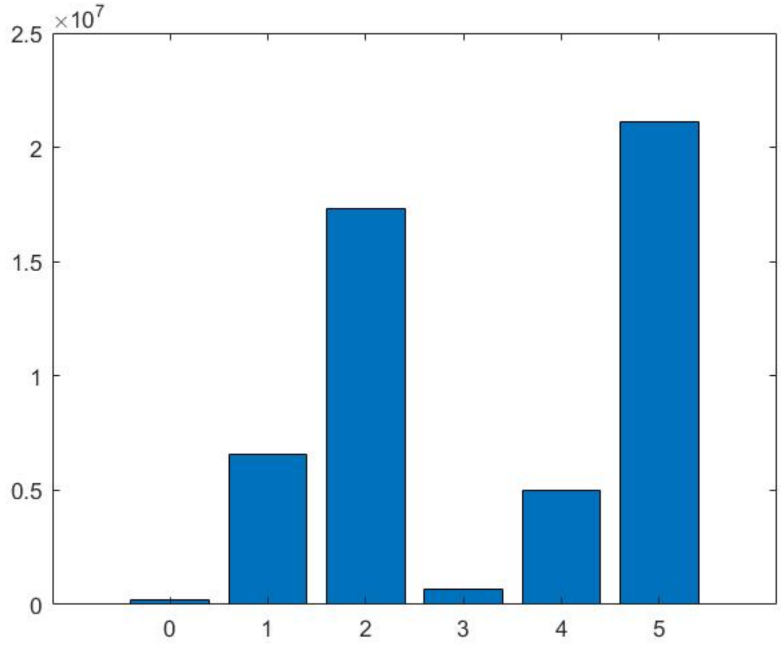

Figure 1: Histogram of each cluster size.

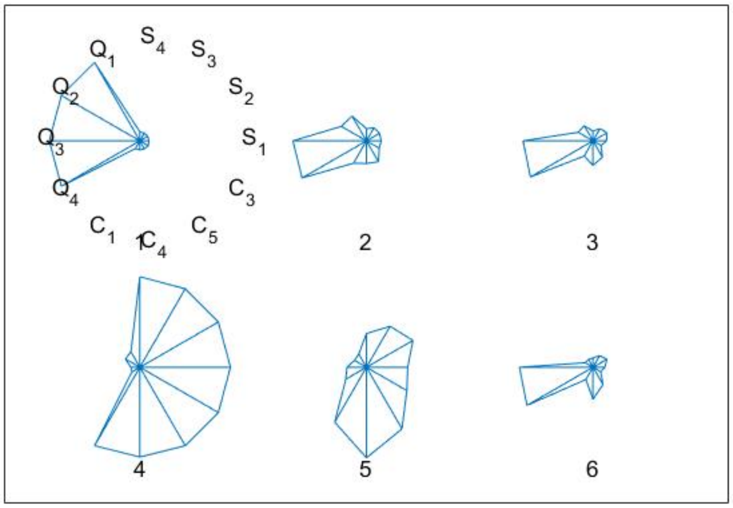

Figure 2: Cluster center visualization.