Recommending Similar Items in Large-scale Online Marketplaces

Paper assessment for Data Science role application at OLX Buenos Aires

What is the motivation?

- Recommending similar items to increase user engagement
- Dealing with short-lived items
- Scalability

What are the key ideas?

- Trade-off between similarity and quality
  - After placing a bid: more specific results (similarity)
  - Coming from a search result: broader results (more weight to quality)

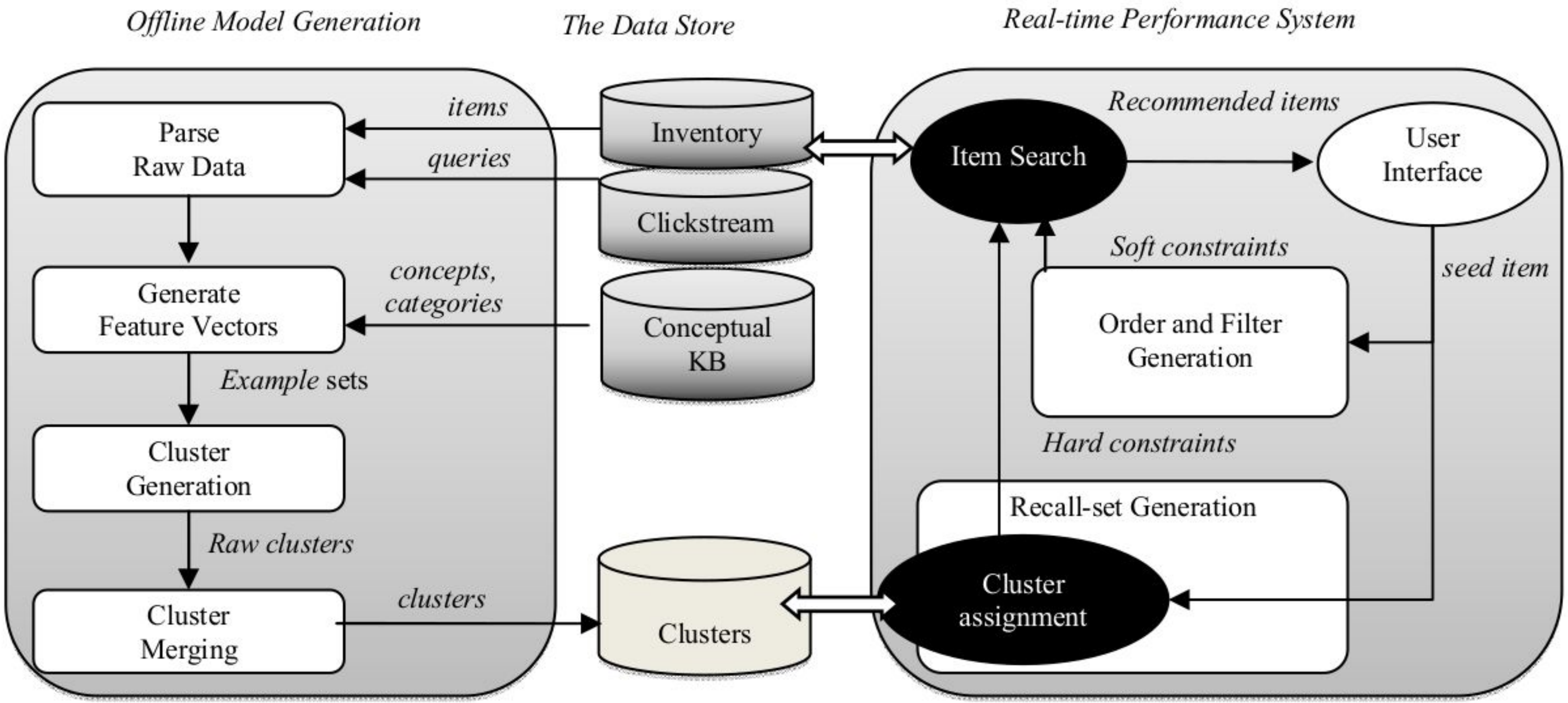
- Short-lived items, but long-term clusters

- Use user queries to learn how they conceptually group items

What are the key ideas? (II)

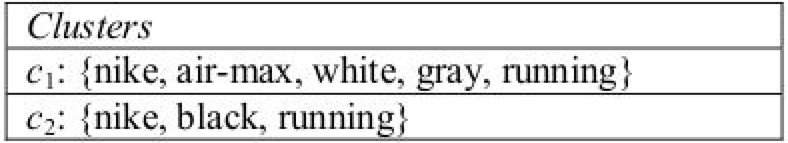
- Offline (heavy) process to generate long-term cluster definitions
- Online (fast) process to refine similarity and include item quality features
- A separate clustering process can run for each user query and therefore the algorithm is highly parallel

Architecture overview



Cluster generation

- Parent clusters: sets of items for most frequent user queries
- Feature vectors: token features weighted by *mutual information* with the item's category
- Algorithm: Bisecting K-Means
- Merging step: remove (near) duplicates and mark parent-child relations
- Cluster expressions: bags of phrases



Clusters
c_1: {nike, air-max, white, gray, running}
c_2: {nike, black, running}

Cluster assignment

$$score(i,c) = C(i,c) \sum_{f \in i} idf(f)^2 \cdot B(f) \cdot N(f,c)$$

- C(i,c): Cluster definitions indexed in Lucene, based on matching phrases
- idf(f): importance in corpus of clusters
- B(f): boosting factor based on user behavioral data
- N(f,c): index time boosting factor

Differences with collaborative filtering and Naive information retrieval

- In marketplaces with short-lived items, pre-computing recommendations using traditional item-to-item collaborative filtering is not feasible.
- It is not solely based on information about the individual items, it also uses user queries to create clusters
- Traditional IR systems are focused on item similarity. This one enables a balance between quality and similarity

Possible shortcomings

- Clustering based on occurrence of terms may not capture some semantic similarities.
- Clusters might get outdated if a large number of new items appear within a short period.

Possible extensions

- Use topic modeling to replace fixed-term clusters with term distributions
- Use an incremental clustering approach to keep clusters updated without the need of expensive model re-training.

Thanks!