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
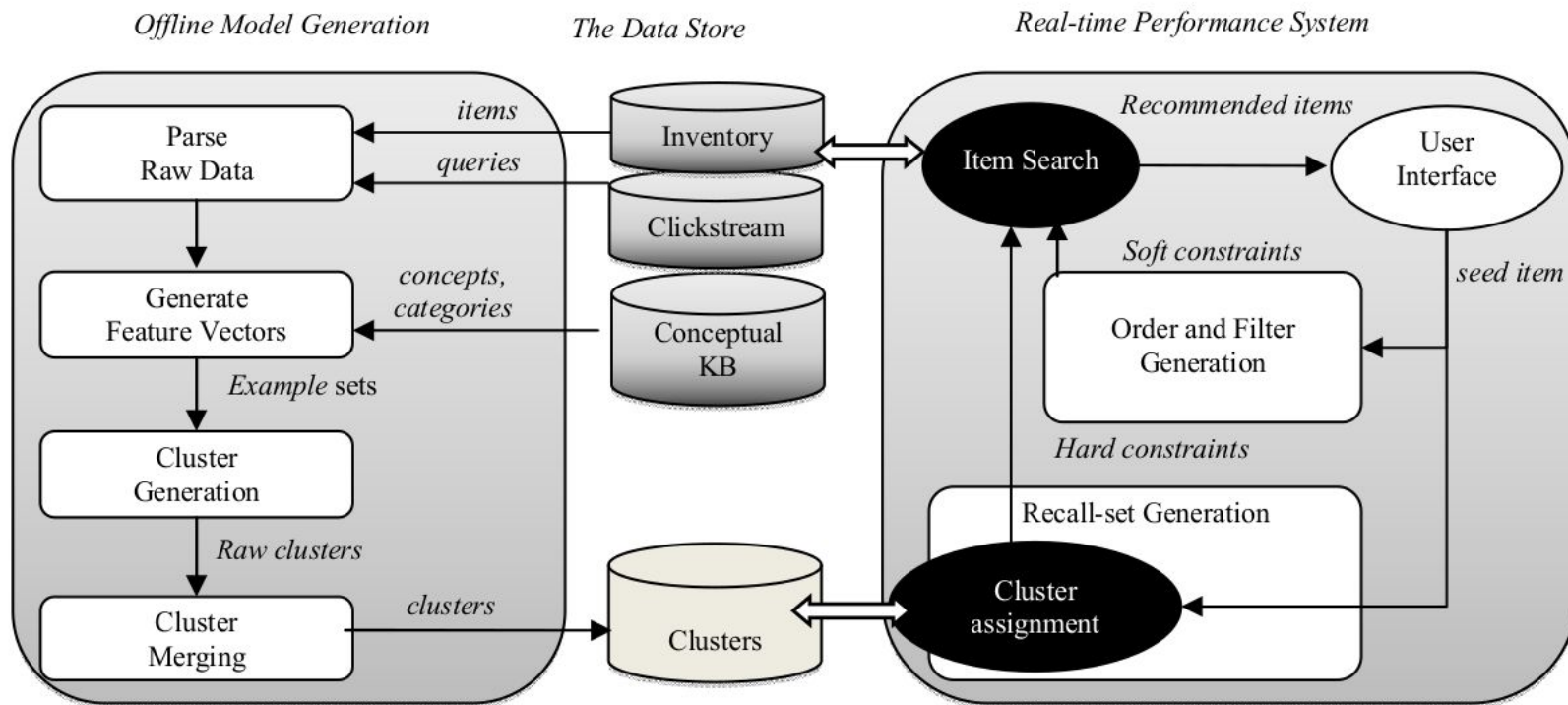
- Scalability

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

Offline Model Generation

The Data Store

Real-time Performance System

Parse Raw Data

items

queries

Inventory

Clickstream

Recommended items

Item Search

User Interface

Generate Feature Vectors

concepts, categories

Conceptual KB

Soft constraints

seed item

Order and Filter Generation

Example sets

Cluster Generation

Hard constraints

Recall-set Generation

Raw clusters

Cluster Merging

clusters

Clusters

Cluster assignment

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- Clusters might get outdated if a large number of new items appear within a short period.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!