Recommending similar items in large-scale online marketplaces

Brief Summary of Research Paper(http://bit.ly/2oYesuo)

Motivation

- Large Scale dynamic marketplace requires faster/scalable results for recommendation system
- Quality of recommendation should increase user engagement and business metrics

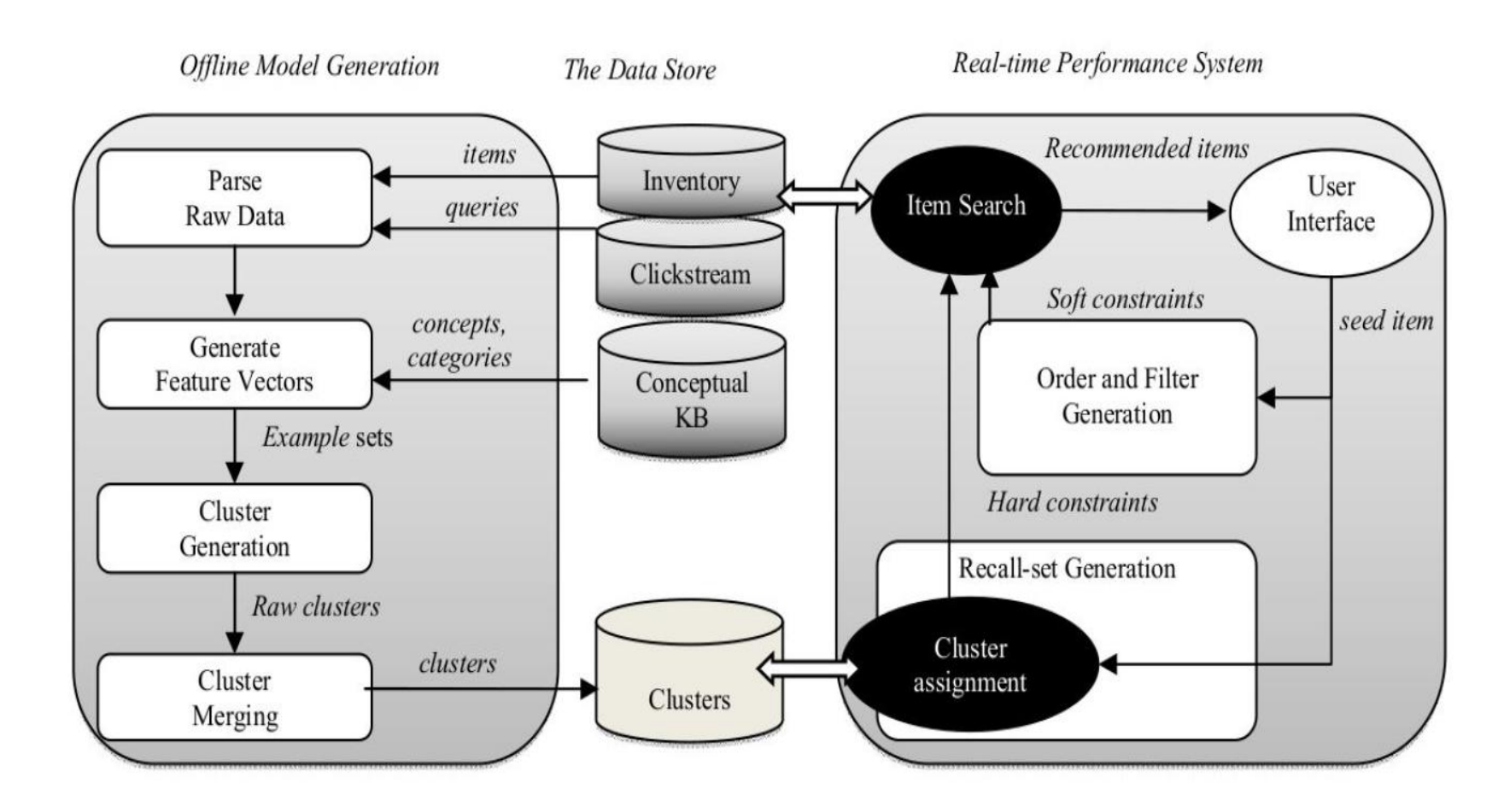
Key Ideas

- Balance – Similarity and Quality
  - Quality – User trust in seller
  - Similarity – Items bought similar to these items on user queries
- User queries to generate clusters
- Generate Clusters based on expressions to support better online infrastructure

Key Ideas - Combination

- Combine online and offline:
  - Offline – Use for cluster generation
  - Online – Item Similarity and quality combined with offline
- User queries use separate clustering process and algorithm highly parallel

Key Ideas - Architecture



Key Ideas – 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}

Key Ideas – Cluster Assignment

- 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

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

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- Clusters might get outdated if a large number of new items appear within a short period.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