Cold-Start Recommendations Across E-Commerce Platforms

Problem Statement

- **Goal:** Optimize personalized recommendations based on a user's limited shopping history across platforms.
- **Methods:** Content-based filtering, collaborative filtering, low-rank matrix completion, and a two-tower neural network model.
- **Cold-Start Issue:** Users switching to new platforms often receive poor recommendations, reducing engagement and conversions.

Success Metrics

- **Recommendation Relevance:** Precision@k, Recall@k, and NDCG.
- **User Engagement:** Click-through rate (CTR).
- **Cold-Start Performance:** Evaluated through hold-out validation sets.

Constraints

- **Data Availability:** Metadata inconsistencies across platforms.
- **Real-Time Performance:** Fast inference and low-latency retrieval.
- **User Privacy:** Compliance with GDPR and data protection laws.

Required Data

- **User interactions:** Order history, wish lists, browsing activity.
- **Product metadata:** Titles, descriptions, images, categories, prices, brands.
- **User-generated content:** Ratings, reviews, preferences.

Potential Pitfalls

- **Sparse Shopping History:** Limited user behavior data across platforms.
- **Metadata Mismatch:** Variability in product attributes.
- **Scalability:** Efficient retrieval for millions of items.
- **Privacy Concerns:** Cross-site tracking must adhere to regulations.

Technical Approach: Collaborative Filtering

Overview

- **User-Based CF:** Recommends items based on similar users' preferences.
- **Item-Based CF:** Suggests items similar to those a user has previously engaged with.
- **Neural CF:** Uses deep learning to capture complex user-item interactions.

Collaborative Filtering: Validation & Constraints

- **Validation:**
  - Offline metrics: RMSE, MAE, Precision@K, Recall@K.
  - Cross-validation, A/B testing, engagement tracking.
- **Constraints:**
  - Cold-start issues, computational scalability.
  - Storage and memory requirements.

Technical Approach: Content-Based Filtering

Overview

- Uses product features (text, images) to recommend similar items.
- Embeddings generated from CLIP (Contrastive Language-Image Pretraining).
- Cosine similarity measures item relevance.

Content-Based Filtering: Implementation

1. **Embedding Extraction:** Pre-trained CLIP model.
2. **Feature Combination:** Merge image & text embeddings.
3. **Similarity Calculation:** Compare item vectors.
4. **Recommendation Retrieval:** Rank items by similarity.

Technical Approach: Low-Rank Matrix Completion

Overview

- Factorizes the user-item interaction matrix into lower-dimensional components.
- Approximates missing interactions for better recommendations.
- Useful for large datasets with sparse interactions.

Technical Approach: Two-Tower Model

Overview

- Separates user and item processing into two neural networks.
- Learns embeddings independently for users and items.
- Computes similarity between user and item embeddings for ranking.

Initial Results: CLIP Image Clustering

- **Dataset 1:** 76 shopping images.
- **Dataset 2:** 1,000 images from Fashion-MNIST.
- **HDBSCAN Results:**
    - Custom dataset: 3 clusters.
    - Fashion-MNIST: 33 clusters.

Evaluation Metrics & Findings

- **Collaborative Filtering:** Precision@5: 0.72, Recall@5: 0.68.
- **Content-Based Filtering:** MAP: 0.81, NDCG@5: 0.79.
- **Low-Rank Completion:** RMSE: 5.82.
- **Two-Tower Model:** Loss converged to 0.021 after 5 epochs.

Next Steps & Future Work

- **Enhance Image Preprocessing:** Standardization, augmentation.
- **Expand Dataset:** Increase user-item interactions.
- **Optimize Clustering:** Improve feature-based segmentation.
- **Select Best Algorithm:** Focus on highest-performing methods.

Open Questions & Challenges

- **Model Selection:** Best pretrained embeddings for diverse styles?
- **Privacy & Compliance:** Cross-site data alignment with regulations?
- **Scalability:** Efficient retrieval for real-time recommendations?

Summary of Approaches

Method	Strengths	Challenges	Best Use Case
Collaborative Filtering	Captures user behavior	Cold-start issues	Users with history
Content-Based Filtering	Works without user data	Limited diversity	New user recommendations
Low-Rank Completion	Efficient for large datasets	Sparse matrix issues	Matrix factorization
Two-Tower Model	Precomputed embeddings	High computational cost	Large-scale retrieval