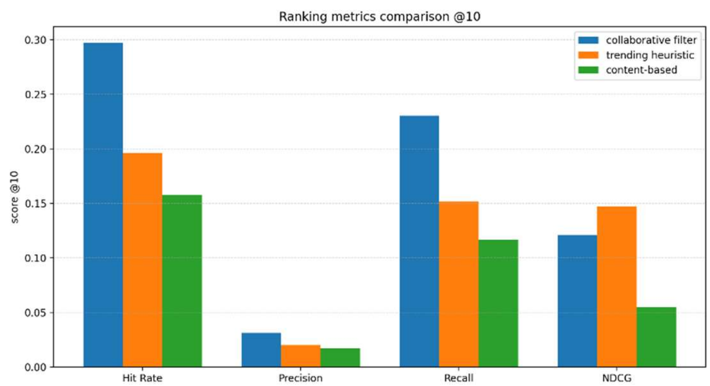
Building New Recommendation Systems

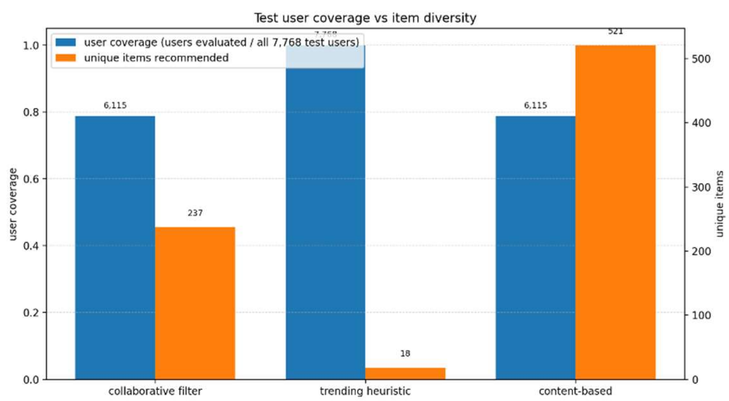
Three, mutually exclusive, models for recommending content to adventurers in our fantasy data set were built. Each individual model uses a singular approach, namely collaborative filtering, content-based, and a trending-content heuristic. The models all operated on the same temporal test-train split, which was split at approximately 85% time coverage, producing a conventional 80-20 training to testing ratio. I used the same definition of a “relevance” across all models for when an item was considered to have been properly watched. A content item is counted as watched if the user’s views of that item had been for over 60% of that content’s length. This was true for both training relevance and evaluation using the test set, which helps make the evaluation target consistent.



The results were slightly underwhelming, but out of the three models collaborative filtering performed the strongest with metric scores of HitRate@10 = 0.297, Precision@10 = 0.0311, Recall@10 = 0.230, NDCG@10 = 0.1208. This stronger performance may be because of the model learning from co-consumption patterns that inherently exist in the user behavior. Rather than try to match items to items or use a bland global heuristic, the model relies on a mix of content meaningfully watched and other content similar users have watched. This provides a balance of matching your taste preference while also introducing a bit of novelty too, as similar users may watch things outside the target users’ profile. The users who this serves best are those with richer qualified histories in training, because the model has more evidence about their preferences and more chances to find meaningful neighbors. Users with no qualified history (cold start users) are inherently impossible to serve with a pure collaborative filter approach and this model avoids them completely. Overall collaborative filtering is a powerful basis for a recommendation system, but this specific implementation may benefit from further refinement, hyperparameter tweaking, and matrix factorization.

Content-based was the weakest predictor of future qualified engagement. It had scores of HitRate@10 = 0.158, Precision@10 = 0.0169, Recall@10 = 0.1167, and NDCG@10 =

0.0548. This is likely because the metadata similarity used (title, genre, studio, language, duration) is a weaker signal on its own for what users will actually watch. It does not capture popularity, seasonal trends, or any other implicit signal. User behavior has strong exposure effects embedded in it that metadata alone cannot capture. However, it did have offer far greater breadth of content than either collaborative filtering (237) or the trending heuristic (18). This outcome is obvious when considering how the systems work. Content-based is always looking for item to item metadata similarity, resulting in a web of neighboring items that has the highest catalog coverage. On the other hand, both collaborative filtering and trending's recommendations identify the phenomenon that a large proportion of views are over a tiny proportion of the total content catalog (essentially popular items). These approaches discard the rest of the catalog, while content-based recommends even potentially unseen, low-quality content. The result is higher coverage but at the cost of quality.



The trending model received HitRate@10 = 0.196, Precision@10 = 0.0200, Recall@10 = 0.1516, and NDCG@10 = 0.1469. It is considerably worse than CF, which is expected because it is not personalized, but surprisingly its NDCG was higher, likely because when trending "hits" it tends to hit very high in the ranking and so NDCG heavily rewards these earlier hits. Its major weakness is of course that it offers repetitive recommendations and it had the lowest item coverage. Regarding cold start behavior, trending is naturally strong for cold users because it does not require any history, but it completely abandons cold start items as it will never even introduce them to the populace. Both CF and content-based require user history to personalize, and therefore they struggle with cold users, but for cold items content-based would serve best as it doesn't rely on interaction history, while trending and CF would fail. A hybrid model would utilize the best of all three worlds, with cold users getting trending items, while warm users get personalized recommendations via collaborative filtering. Content-based can be sprinkled into both these strategies in order to diversify and further personalize recommendations, as well as promote cold start items.