Minimising the Long-tail Problem in Collaborative Filtering Based Recommender Systems Using Clustering

Author: Yash Mundhra

Supervisors: Aleksander Czechowski, Davide Mambelli, Oussama Azizi, Frans Oliehoek

24th June 2022

Introduction: 1

Recommender System: Algorithms aimed at recommending attractive items to users based on user profiles, previous purchases, and ratings.

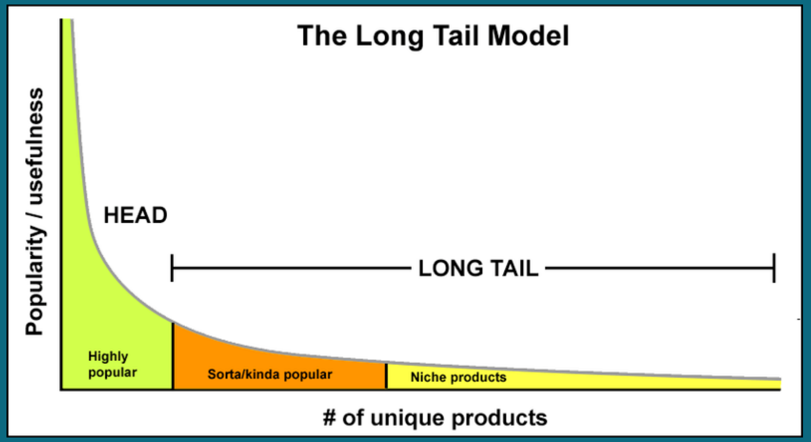


Figure 1: The long tail model visualized

The long tail problem: "recommender systems ignore unpopular or newly introduced items having only few ratings and focus only on those items having enough ratings to be of real use in the recommendation algorithms" [1].

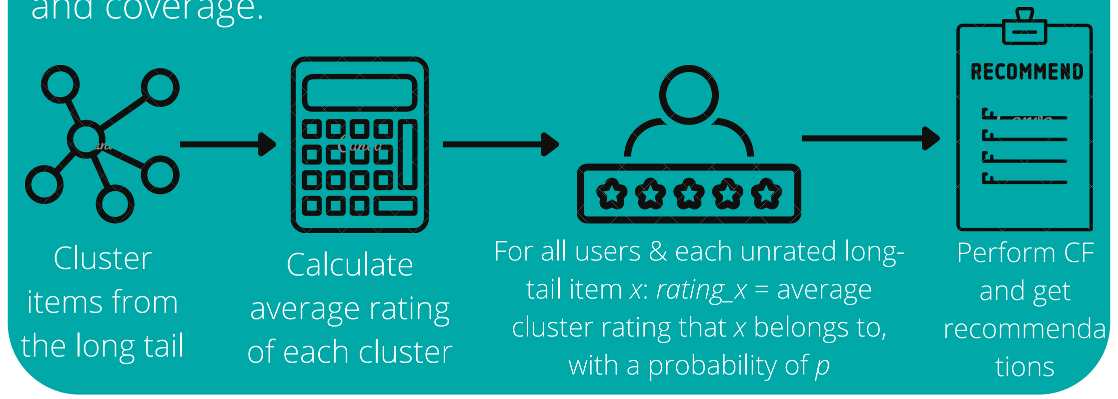
The long tail problem is caused due to **data sparsity**. Some items have too few ratings to make accurate recommendations.

Research Question:

To what extent can clustering be applied to the long-tail of collaborative filtering recommender systems such that more long-tail items are included in the set of recommendations while not affecting the accuracy of the recommender system and how do the number of clusters and the cutting point have an influence on this?

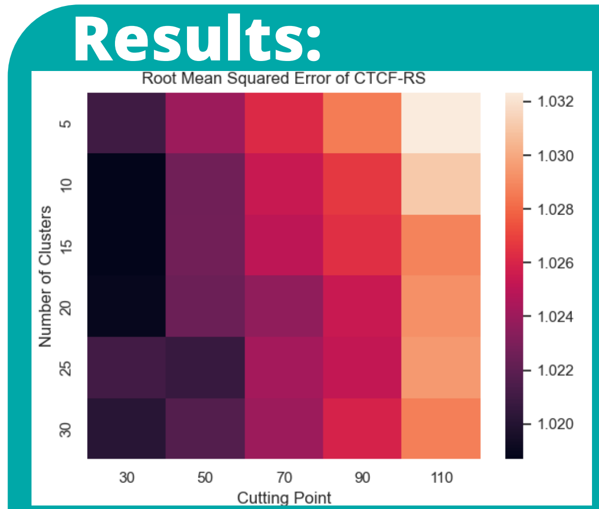
Proposed Solution: 2

Adaptation of clustered tail method (CT) proposed by Park and Tuzhilin [1], to work in the context of Collaborative Filtering based recommender system. Investigate the impact of number of clusters and cutting point on the accuracy, diversity and coverage.



Cluster items from the long tail → Calculate average rating of each cluster → For all users & each unrated long-tail item x: *rating_x* = average cluster rating that x belongs to, with a probability of *p* → Perform CF and get recommendations

Results: 3

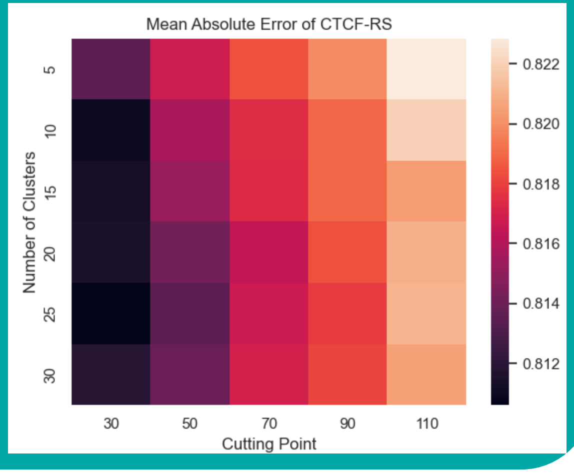


RMSE:

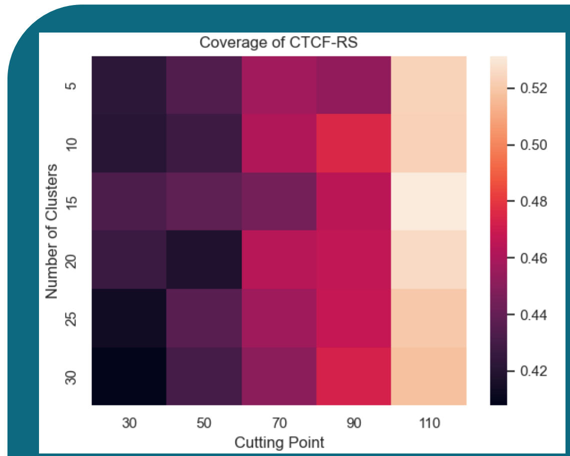
- Increase in cut point leads to increase in RMSE.
- Higher number of clusters lead to better accuracy.

MAE:

- Similar trend as RMSE.
- Baseline RS performed better for cutting point greater than 50.
- MAE slightly different than RMSE due to square of error.

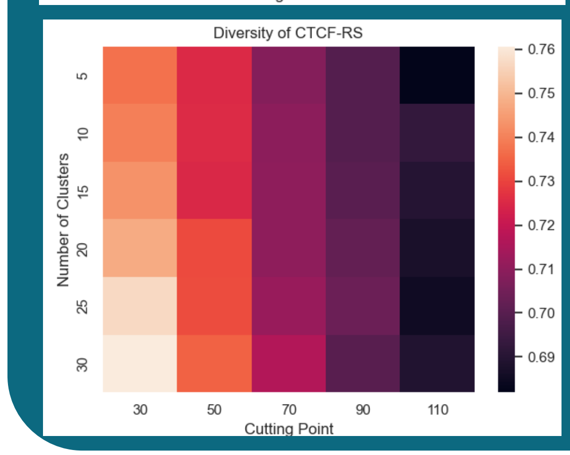


4



Coverage:

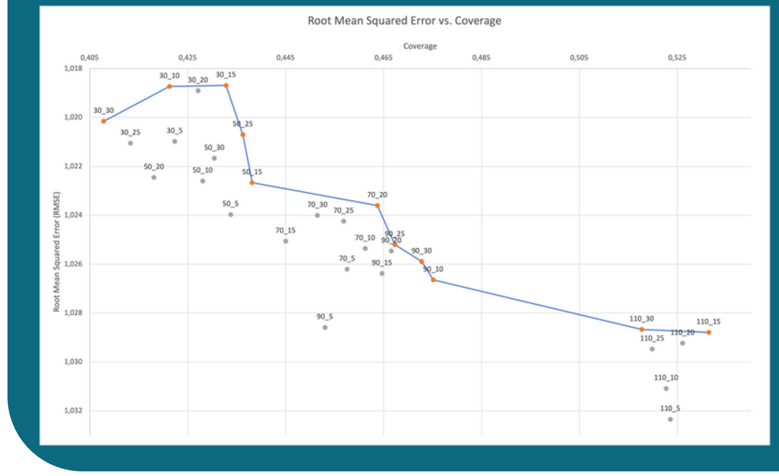
- Significant improvement in coverage (approx 13%) in comparison to baseline RS.
- Positive correlation between cutting point and coverage.
- Highest at cutting point 110 and lowest at cutting point 30. Jump from 90 to 110.



Diversity:

- Small improvement in diversity for cutting points 30 and 50.
- Baseline RS has relatively high diversity and outperform CTCF-RS for cutting points 70 and higher.

Discussion & Conclusion: 5



- Tradeoff between accuracy/diversity and coverage must be made.
- Increase in coverage = decrease in accuracy.
- Pareto frontier determines optimum point between two parameters.
- Several future extensions; run-time, generalisability, parameter tuning.

References: 6

[1]. Yoon-Joo Park and Alexander Tuzhilin. The long tail of recommender systems and how to leverage it. In Pro-ceedings of the 2008 ACM Conference on Recommender Systems, RecSys '08, page 11–18, New York, NY, USA, 2008. Association for Computing Machinery