COMP47580

Recommender Systems & Collective Intelligence

Recommender Systems Report

<Troitskii Arsenii>

<20204701>

Declaration of Authorship

I declare that all material in this assessment is my own work except where there is clear acknowledgement and appropriate reference to the work of others.

Signed ......Troitskii Arsenii................... Date ....4/24/2025................................

1. Analysis of Ranking Approaches

1.1 Popularity Ranker

We start our analysis with Popularity Ranker, which shows the highest relevance (3.7129) across all rankers. It represents that recommendations of popular domains align well with users' preferences. That trend was expected, as popular domains (such as films in our case) are appreciated by many users.
Popularity Ranker reached excellent metrics (1.0) in coverage and item space coverage, which means that the ranker can make recommendations for all users and for different preferences, using the whole list of domains.

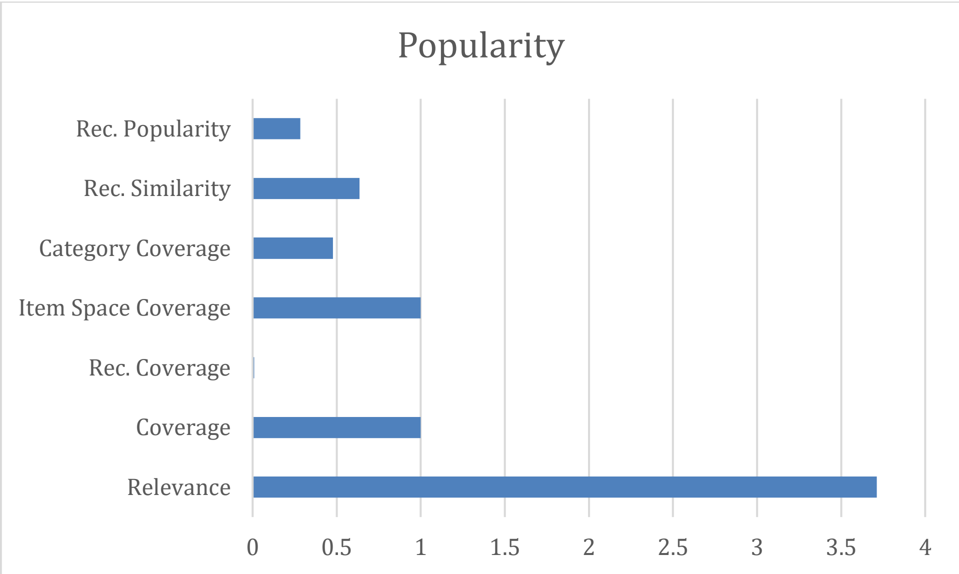

Figure 1: Evaluation metrics for Popularity Ranker

On the other hand, Popularity Ranker shows poor results on recommendation coverage (0.009), which represents that it tends to recommend the same several popular domains to every user. This can result in creating a group of popular domains where users receive only limited content.
In addition, the highest recommendation popularity (0.2838) also highlights that this approach is mostly focused on the most popular items (films in our case).
Category coverage is moderate, which means popular items could be included in several categories, but not in all of them. Recommendation similarity for Popularity Ranker is relatively lower than for other rankers, which is unusual for a popularity-based approach.

1.2 Genre Jaccard Ranker

For Genre Jaccard Ranker, as represented on Figure, we observe that Jaccard Ranker shows the lowest relevance (3.2819) across all rankers, which means that genre-based approach by itself is not the best algorithm for users. It has excellent coverage (1.0), but has moderate item space coverage (0.6068).
Based on recommendation coverage, which is equal to 0.7824, it is good enough but not perfect, which indicates that it recommends a reasonably diverse set of films across all users.
However, the category coverage is the lowest one across all rankers (0.2148), which is surprising for a genre-based approach, as it can lead to focus on only specific genres.
Recommendation similarity is relatively high (0.7507), indicating similar suggestions for users with similar genre preferences.
Also, in comparison to Popularity Ranker, Jaccard Ranker has lower recommendation popularity (0.0471), which means it recommends less mainstream items.

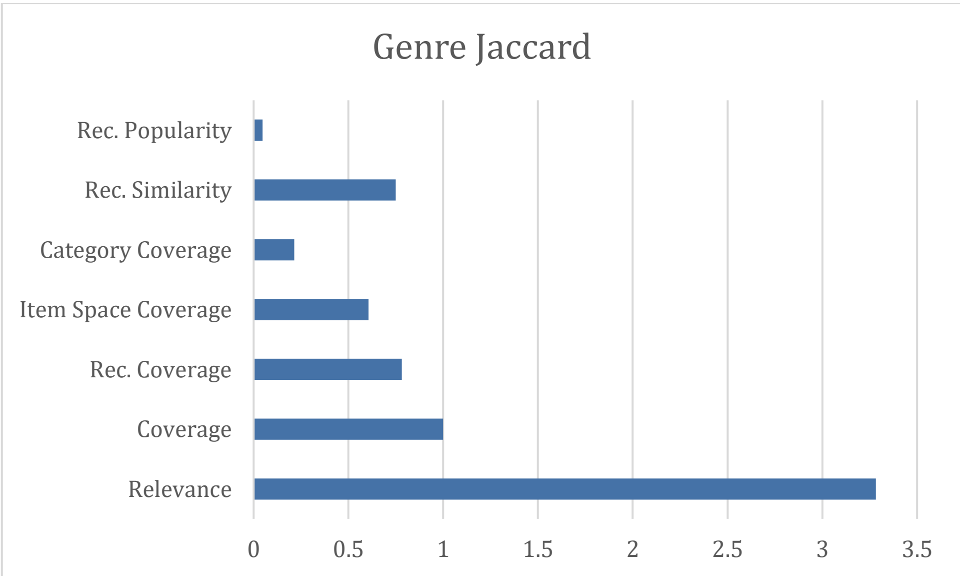

Figure 2. Metrics for Genre Jaccard

1.3 Genome Cosine Ranker

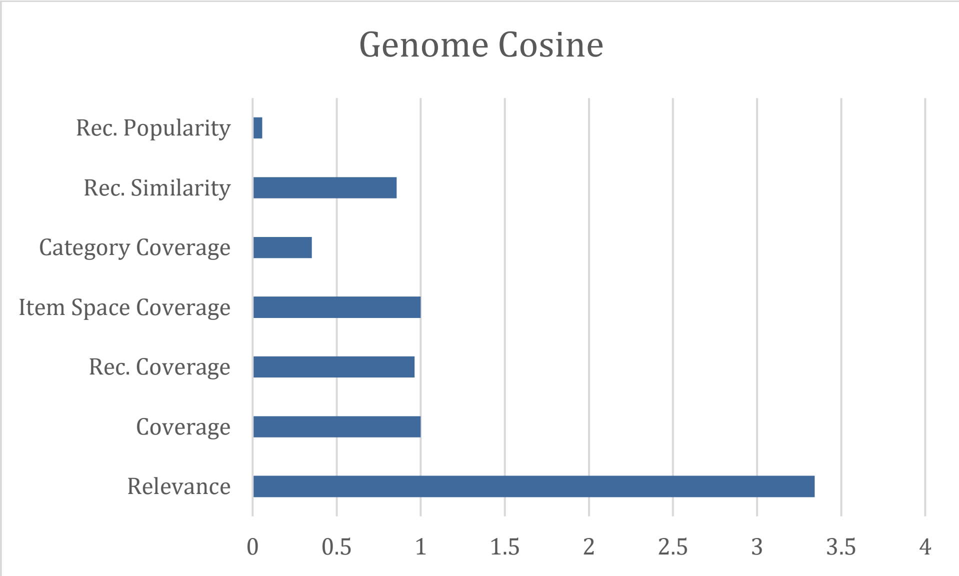

Figure 3: Genome Cosine Ranker

For Genome Cosine Ranker, which metrics represented above, it can be observed, that that content-based similarity using genome tags correlates well with user preferences, as relevance shows good metrics (3.4333). It was also achieved maximum value for coverage and item space coverage (1.0 both), indicating comprehensive users and items coverage. Its Recommendation Coverage (0.9639) is excellent, which means it recommends a diverse set of items across users. The Recommendation Similarity (0.8557) is the highest across all rankers, suggesting that genome-based similarity tends to consistent recommendation patterns across users. However, its Recommendation Popularity (0.0575) is low, which is great, as it suggest less mainstream items but more new and unique once for specific user preferences.

1.4 Rating Pearson Ranker

Talking about Rating Person Ranker, as represented below shows relatively high Relevance (3.2972), which means ranker suggesting correlates well with users preferences. However it has completely different cases for coverage and space coverage metrics (1.0 and 0.3463) respectively, that could tend of using only a third of available item catalog. On the other hand, Pearson Ranker reached the highest recommendation coverage across all rankers (0.9934), indicating it recommend the most possible diverse set of items across users. The category coverage (0.4768) is good enough, represented that rankers recommendation includes many categories. The recommendation similarity is relatively high (0.7542), while recommendation popularity is pretty low (0.049), indicating it also recommends less mainstream items, similar to Genome Cosine Ranker

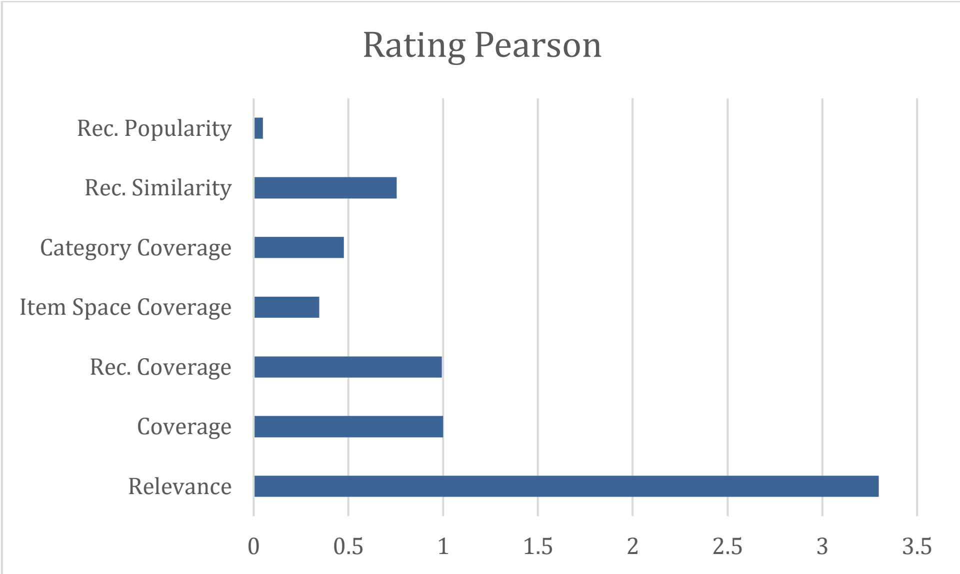

Figure 4: Rating Pearson

1.5 Inc. Confidence Ranker

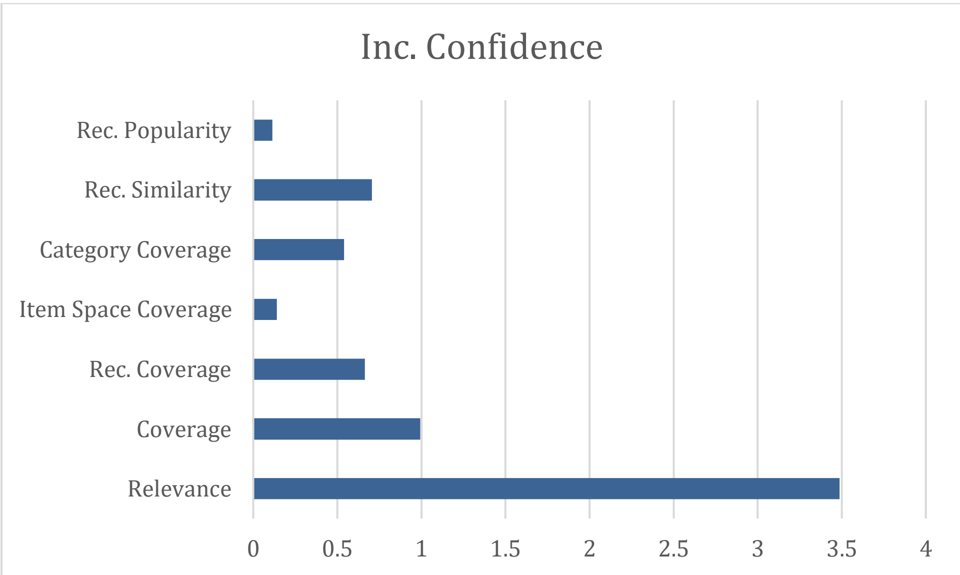

Figure 5: Inc. Confidence Ranker

For In. Confidence Ranker, as it can be observed form figure above, it has the second highest relevance (3.4868), after the Popularity Ranker, suggesting that confidence-weighted approach is also align well with user preferences. Coverage is slightly below the maximum value, but still excellent (0.9926). However it has the lowest item space coverage, indicating it uses only the small amount of all available items. Recommendation coverage and recommendation similarity are moderate (0.6634 and 0.7057 respectively). The category coverage is the highest across all rankers, indicating it recommends items from the widest range of categories. Its recommendation popularity is the second highest (0.1132) after the Popularity ranker, suggesting it recommends more popular items than other rankers, but not popular as popularity-based approach.

1.6 Comparative Analysis and Trends

Figure below represents the correlation heatmap between the evaluation metrics discussed above. Green cells represents strong positive correlation, red ones strong negative correlation. The brighter colors are, the stronger correlation their mean.

	Relevance	*Coverage*	*Rec. Coverage*	*Item Space Coverage*	*Category Coverage*	*Rec. Similarity*	*Rec. Popularity*
Relevance	1						
Coverage	-0.193291483	1					
Rec. Coverage	-0.942805499	0.026585583	1				
Item Space Coverage	0.307822684	0.693846466	-0.367074205	1			
Category Coverage	0.552541115	-0.548854289	-0.304483108	-0.368759071	1		
Rec. Similarity	-0.774068786	0.243618628	0.846943518	0.138781925	-0.463104836	1	
Rec. Popularity	0.98039047	-0.017079455	-0.974380863	0.399618005	0.448562142	-0.796659705	1

Figure 6: Correlation heatmap

It also necessary to have the overall comparison of all evaluation metrics, so it would more clearly visible the trade-off process between all ranking approaches.

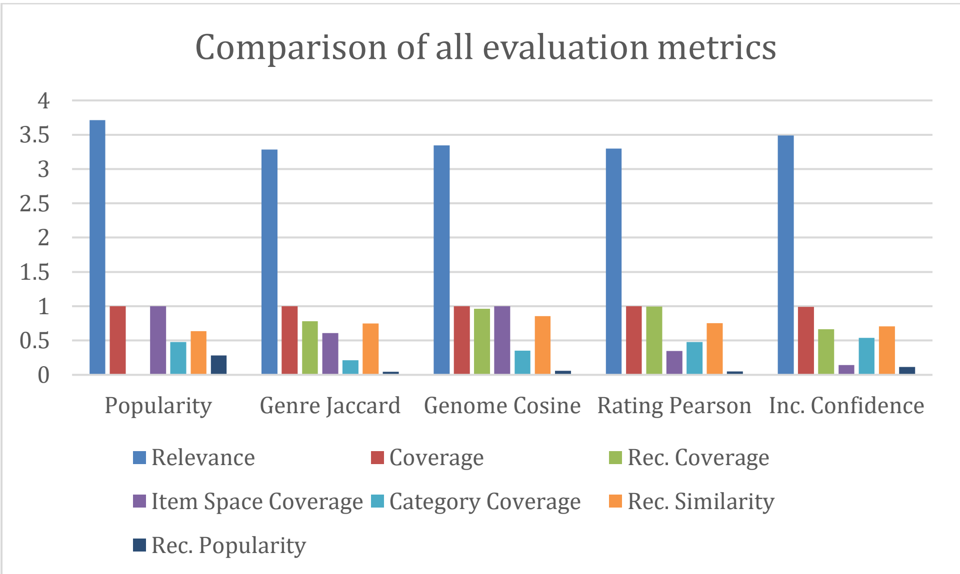

Figure 7: Comparison of all evaluation metrics across ranking approaches

As we can observe from Figure 6 and Figure 7, there is a clear trade-off between relevance and other metrics. For example, Popularity Ranker achieves the highest relevance metric but has the lowest recommendation coverage. While Inc. Confidence ranker shows a good balance, with the second highest relevance metric and having a reasonable diversity. It can be resulted, there is a clear inverse relationship between relevance and recommendation diversity, suggestion a trade-off between recommendation accuracy and diversity.

As it can be mentioned, all rankers perform excellent on user coverage, with most of them achieving the maximum value. However, for recommendation coverage, there is a significant variation, while Popularity Ranker performing extremely poor, Rating Pearson Ranker shows excellent performance. That suggest, popularity-based approaches recommend the same items to many users, while similarity-based approaches are more focused on personalized recommendations.

Talking about the balance and trade-off, between coverage metrics, it can be mentioned how different rankers balance various coverage metrics. For example, Genome Cosine Ranker achieves excellent balance across all coverage metrics (perfect item space coverage, high recommendation coverage, and moderate category coverage), making it a good approach for diverse recommendations.

1.7 Conclusions

Based on the analyse above, it is not possible to say which one approach is the most suitable one, as different rankers can suitable for different recommendation cases. For example:

Popularity ranker could be the best choice for new users, with now preference data. Or for the case where the mainstream content should be prioritized.

Genome Cosine Ranker is most well balance approach across all ranker, it could be suitable for general purpose recommendation systems, where already balanced performance is prioritized.

Rating Pearson Ranker is the best approach for cases, when it is necessary to expose users to a wide range of content. While Inc. Confidence Ranker provides the best metrics for category diversity, it could be suitable when it is necessary to expose users to diverse categories.

Genre Jaccard Ranker performs poorly on most metrics, so it might not be sufficient for effective recommendations.

It could be a good idea to use a hybrid approach combining Popularity Ranker (for new users) and Genome Cosine Ranker or Rating Pearson Ranker (for users with preference data) and would likely provide the best overall performance, balancing relevance with diversity.

2. Analysis of Collaborative Filtering Approaches

Lets start our second part of experiments with analyzing the effect of neighbourhood size on RMSE. Below is the representation, how RMSE changed over different k size.

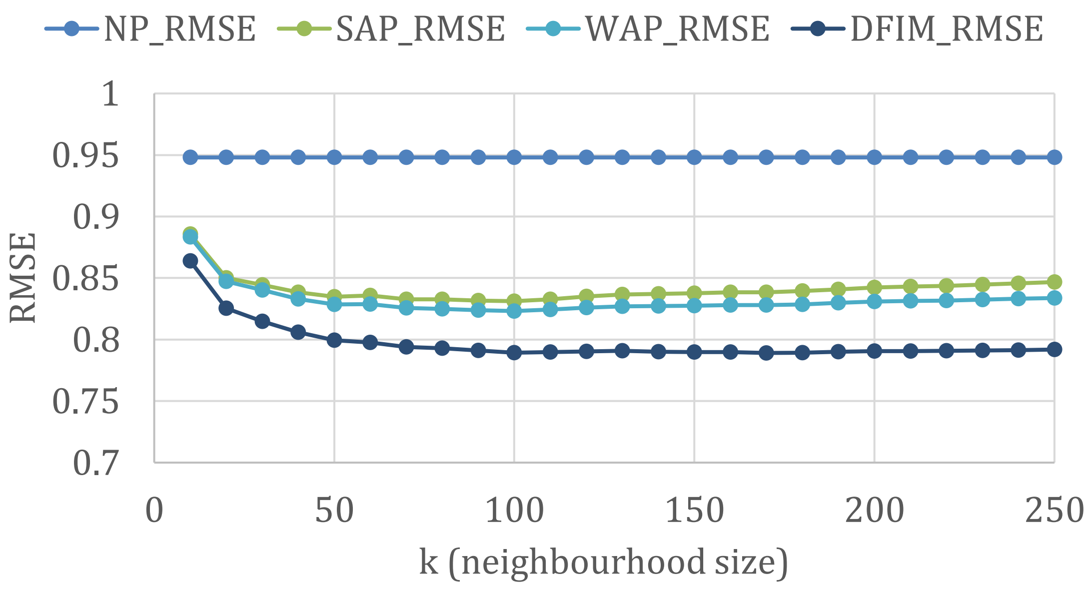

Figure 2: Effect of Neighborhood Size on RMSE

Based on the correlation above, the non-personalized approach maintains a constant RMSE of approximately 0.95, and was not affected at all by changing the k. Which was expected as that approach does not use the neighbourhood size.
Talking about tree personalized predictors (Simple Average, Weighted Average, and Deviation from Item Mean), all of them show improvement, which represented in decreasing in RMSE, with increasing in number of neighbourhoods being used. With the most significant improvements occurring in the range of 5-75 neighbours.
The Deviation from Item Mean predictor shows the best performance over the other approaches, reaching the lowest RMSE (around 0.85) across all neighbourhood sizes.
The weighted average predictor, performs slightly better than the Simple average predictor, especially represent as number of neighbourhood size increases
All their personalised approach tends to decrease their RMSE value significant, till they reach neighbourhood size around 100. After that there is not significant improvements, suggesting additional neighbourhoods does not affect the accuracy anymore.

Most of that observations are expected as large neighbourhoods provides more data for predictions, which is necessary for accuracy improvement. The Deviation from Item Mean predictor considered both uses and item biases, which leads for a better prediction. As we observe, after around 100 neighbourhoods, there is no changes on RMSE almost, which can occur, because of the less information from neighbourhoods that a far (low similarity).

2.1 Analysis of Coverage Results

Taking about coverage and its dependences on neighbourhoods' size, which is represented on figure below, also illustrate some important patterns.

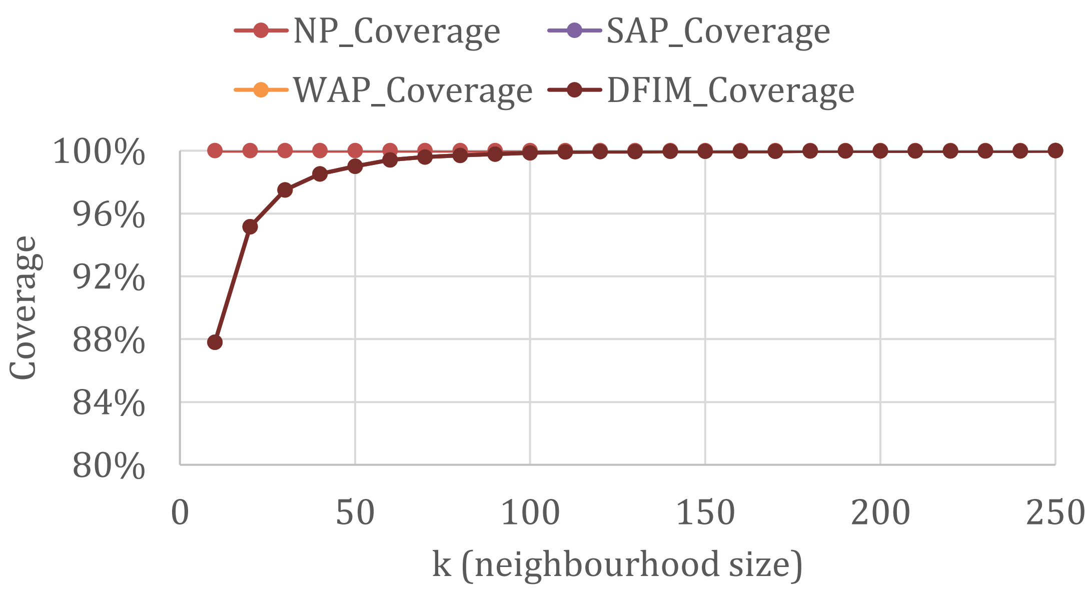

Figure 2: Effect of Neighbourhood Size on coverage

The Non-personalised approach maintains 100% coverage and does not change at all, which is also expected as it generate predictions for all user-item pairs and does not take information from neighbourhoods.

The deviation from Item Mean predictor, shows the most significant improvements as number of neighbourhoods increasing, (from around 88% with 5 neighbourhoods and reaching almost 98% with 250 neighbourhoods).

The simple average and weighted average does not appear on a coverage graphs, because they also maintain consistent coverage (same as Non-Personalized approach), so they are represented behind it.
The coverage improvement for the Deviating from Item Mean approach is most dramatic in the range of 5-50 neighbourhoods, with sharp increasing to 100% point.

Most of these results are expected as large neighbourhoods increasing the potential if finding similar users who have rated an item. The Deviation from Item Mean approach, are depend on neighbourhoods information, which makes it more sensitive to neighbourhood size. There is also a point, at which the number of neighbourhoods does not make any changes for predictor, indicating that user-item pairs probably represent unusual preferences, which are difficult to consider.

2.2 Effect of Neighborhood Threshold on RMSE and Coverage

To figure out the trade-off between prediction accuracy (inverse of RMSE) and coverage, figure 3 will be used.

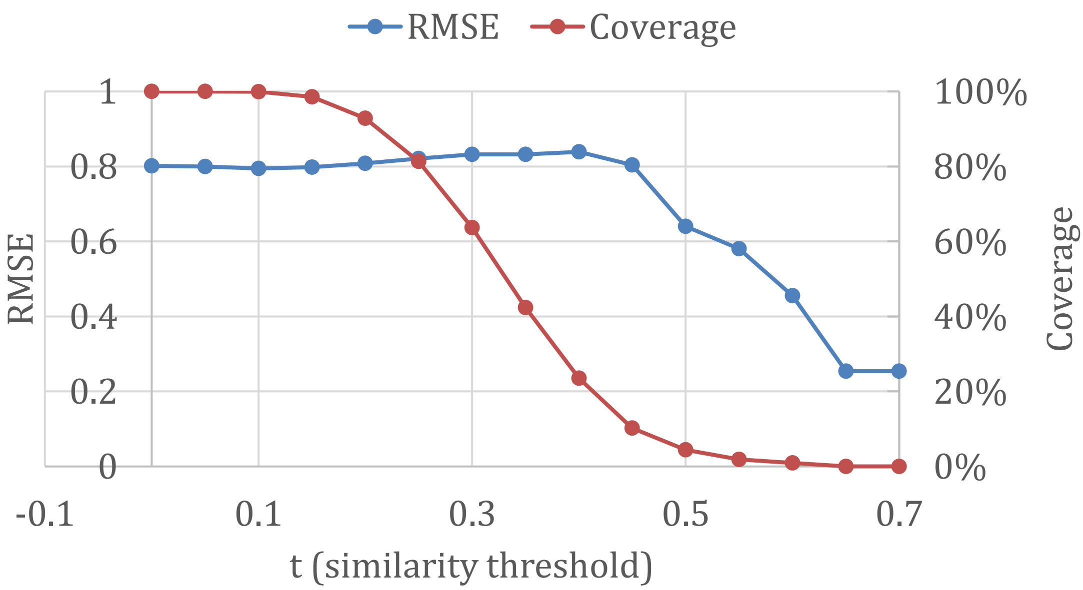

Figure 3: Correlation between RMSE and Coverage

As it can be observed, from the figure above, RMSE initially remains stable (around 0.85) until similarity threshold reaches about 0.3, after that RMSE decreasing significantly, reaching almost 0.2, when similarity threshold reaches 0.7. Almost at the same time, coverage decreases as threshold increases, starting from almost 100% at threshold equals 0 and dropping dramatically to almost 0% as threshold reaches 0.7. It can also be mentioned, that both curves intersect at approximately threshold equal to 0.25 where both RMSE and Coverage are equals to 0.8 and 80% respectively.

Most of that happened, as higher threshold tend to decrease the number of neighbourhoods to only very similar users, which improve the accuracy by excluding dissimilar users. However, as threshold increases more but the data and number of users are limited, many items do not have enough similar users who have rated them, which tend to decreasing accuracy and increasing RMSE.

Well balance of moderate threshold (around 0.2 – 0.3) will maintain the high coverage rate (around 90%) and will not significantly decreasing prediction accuracy.

2.3 Effect of Similarity Metric on RMSE and Coverage

During this observation we will compare two similarity metrics: Cosine similarity and Mean Squared Difference (MSD).

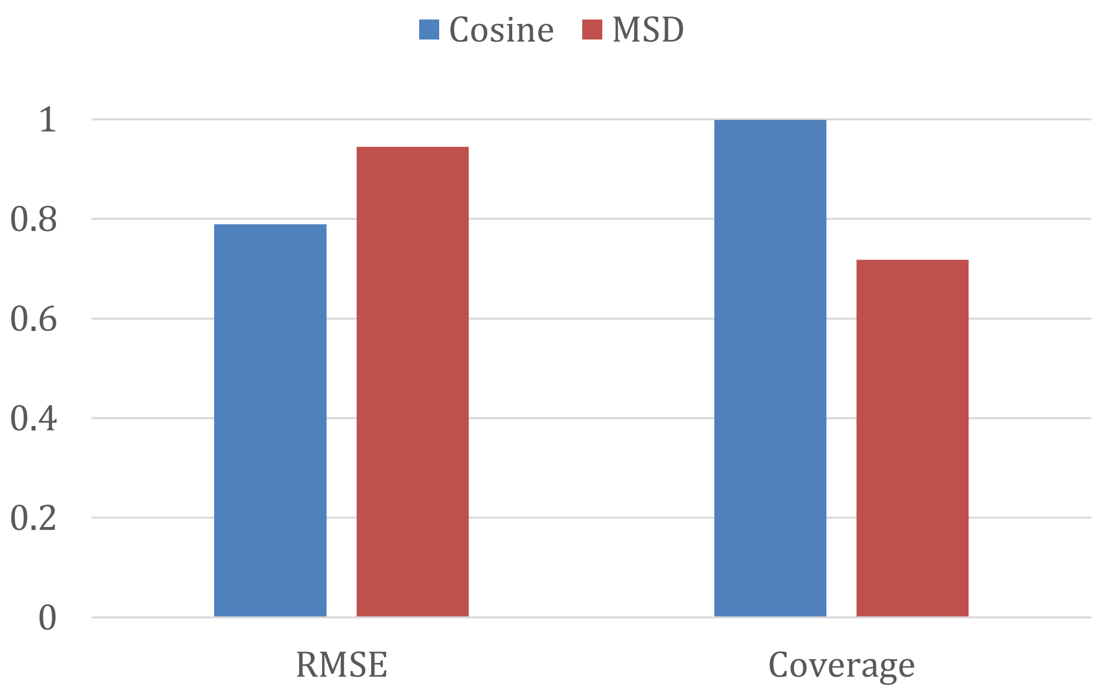

Figure 4: Similarity metrics

These bar charts represent that Cosine similarity achieves an RMSE of almost 0.8 while MSD achieves slightly higher RMSE (around 0.86), indicating that Cosine similarity provides better prediction accuracy.
In terms of coverage, Cosine similarity achieves almost 100%, while MSD achieves around 70%. Representing, that Cosine similarity also provide much better coverage.

It is possible to figure out some reasons of that patterns: Cosine similarity focuses on direction of rating vectors rather than absolute values, making it difficult to differ rating scales among users. While MSD is more sensitive to the absolute differences in rating, which can be affected by noise ratings. Overall Cosine similarity appears to be a better choice, providing better prediction and higher coverage.

2.3 Exploring the Relationship Between RMSE and Coverage

As this section we will investigate the relationship between RMSE and coverage. In previous section we already faced some similar, by increasing the threshold initially improves RMSE slightly, while reduce coverage. This trade-off can be approximately calculated: in previous section, increasing the threshold from 0.0 to 0.3 results in approximately a 1% improvement for RMSE, but a 7% decreasing for coverage. So we can assume a ratio around 1:7. Moreover, the figure x also demonstrates the same inverse correlation, as it can be observed: decreasing RMSE tends to increase Coverage and vice versa.

This ratio, could be really useful for the large datasets in real-world recommendation systems.

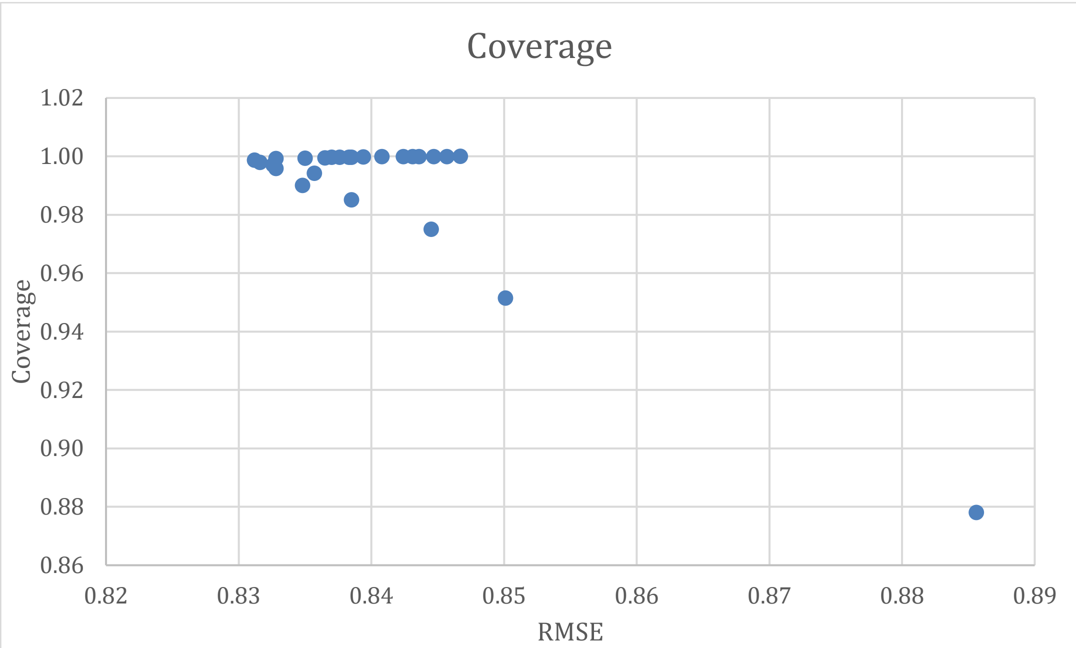

Figure 5: Relationship Between RMSE and Coverage

3. Conclusions

Taking into consideration all that is mentioned above, the Deviation from Item Mean predictor consistently performs better than other approaches in terms of RMSE. Which can be the best option for cases where the prediction accuracy is the most important aspect.
Talking about neighbourhood size, we can conclude that neighbourhood size of around 75-100 elements provides a good balance between prediction accuracy and efficiency. While large number of neighbourhood size does not really effect the RMSE and coverage.
For similarity threshold, the most appropriate is 0.2-0.3, that amount offers a good balance between accuracy and coverage. While higher threshold significantly reduce coverage, without improvements in accuracy.

To conclude regarding similarity metrics, cosine similarity performs slightly better than MSD in terms of RMSE and coverage, Which can be an efficient approach for the similarity based datasets.

As it was observed, there is a clear trade-off process between prediction accuracy and coverage, that is necessary to consider, when designing a recommendation systems.
The best approach for a recommendation system is a hybrid one, as it is possible to take all necessary advantages from each approach, the Deviation from Item Mean predictor with cosine similarity can give the best metric performance, with the neighbourhood size to approximately 100 and threshold around 0.2-0.3. That configuration will provide the best balance between prediction accuracy and coverage.