Multicriteria Recommendation System by Leveraging Predefined, Implicit, and Undefined Criteria

Emrul Hasan[1] and Chen Ding[2]

[1] Dept. of Comp. Science, TMU, & Vector Institute, Toronto, Canada
e1hasan@torontomu.ca
[2] Dept. of Comp. Science, Toronto Metropolitan University, Toronto, Canada
cding@torontomu.ca

Abstract. In this study, we propose a novel multi-criteria recommendation model that utilizes predefined, implicit, and undefined criteria. We use a semantic similarity-based sentence clustering method to identify the predefined and implicit criteria and a sentiment analyzer to estimate their ratings. Semantic similarity between each sentence in the review and the predefined criteria are calculated, and then the sentence is assigned to the most similar criteria. The criteria that are extracted from the review and are aligned with predefined criteria are referred to as implicit criteria. A sentence is considered as expressing opinions on an undefined criterion if the similarity score between this sentence and all the predefined criteria is lower than a predefined threshold. Ratings are computed for each extracted implicit criterion and the undefined criterion based on the review content. Finally, we use all three types of criteria and an aggregation model to make the final rating prediction for the recommendation system. Our proposed method demonstrates the superiority compared to several baselines on TripAdvisor and Beer Advocate datasets.

Keywords: Criteria rating, implicit criteria, undefined criteria, predefined criteria.

1 Introduction

Due to the rapid growth of online data, recommendation system has become a ubiquitous tool for e-commerce and content providers such as Amazon, Netflix, YouTube, etc. When users choose a product or service, they typically consider one or multiple attributes rather than relying solely on overall ratings. In the context of hotels, a user might prioritize factors such as location and service, or focus solely on price. The overall rating could be high due to satisfaction with other criteria, leading to a situation where the user may not receive appropriate recommendations. To address this issue, the researchers turned their attention to multi-criteria recommendation systems where multiple criteria ratings are utilized to learn user preferences [1–3]. Another direction is to build aspect-based systems where user preferences and item attributes are learned based on different aspects that users discussed in their reviews [3–5, 8].

While aspect-based [6] and multi-criteria rating-based [7] recommendations rely on a single data source (review or criteria) for user preference learning, we argue that a more accurate representation of user preferences can be learned when we consider both the criteria ratings and the sentiment expressed through reviews. It is observed that in addition to pre-defined criteria (domain-specific criteria that are defined by the business), users express their opinions on different aspects more freely through reviews.

To this end, we propose a novel multi-criteria recommendation model that utilizes both the predefined criteria and the criteria (aspects) extracted from the review. The predefined criteria may also be discussed in the review, however, the ratings (sentiment scores) associated with them may not be the same as the ratings users explicitly give to the predefined criteria. So, to differentiate the two, we call the ratings based on the review as implicit criteria ratings. If the aspects extracted from the review are not available in the predefined criteria list, they are named undefined criteria.

We assume that each sentence or sub-sentence in the review contains the user's opinion on one criterion. Some criteria are discussed in multiple sentences. To extract the criteria from the review, we split the review into sentences and clustered them by comparing the semantic similarity between predefined criteria and the split sentences. The sentences that have high similarity degrees with predefined criteria are clustered into different homogeneous groups, each of which corresponds to a predefined criterion. The following are the major contributions of our research:

1. We propose a novel multi-criteria recommendation model by leveraging both reviews and the predefined criteria. We show that the user's overall preference for an item not only depends on the predefined criteria ratings but also on the ratings extracted from reviews.
2. We leverage all three types of criteria ratings to learn the overall rating e.g. predefined, implicit, and undefined. To the best of our knowledge, our cluster-based aspect extraction method is new, and the process of aligning the criteria extracted from reviews with the predefined criteria to get undefined criteria ratings and include them in the overall rating prediction is also considered novel.

2 Related Work

Nassar et al. [1] introduce an aggregation-function-based recommendation consisting of two parts, criteria ratings prediction using a deep collaborative filtering model by leveraging DNN and MF, and learning an aggregation function using a DNN. Hong et al. [3] develop a multi-criteria tensor model for tourism recommender systems. They explore user preferences by incorporating cultural factors into the evaluation of multiple criteria ratings. Wang et al. [9] introduce a recommendation system with a Hybrid Deep Tensor Decomposition model, incorporating multiple criteria. It is observed that the consideration of multiple factors has an impact on recommendation performance. Hong et al. [10] introduce ClustPTF (Cluster-based Parallel Tensor Factorization), which employs sentiment analysis to mitigate data sparsity within the tensor model [11]. Next, K-means clustering is used to cluster similar user preferences, thereby improving recommendation diversity. Shambour [2] introduces a deep learning-based multi-criteria recommendation system. Although all the above-mentioned multi-criteria recommendation models have proven to be successful, they use predefined criteria while ignoring the user sentiment expressed through reviews.

On the other hand, aspect-based approaches concentrate on learning the aspect representations from reviews and utilizing them to predict overall preferences. ANR [4] introduces an Aspect-based Neural Recommendation model incorporating an attention mechanism to learn user and item representations simultaneously. Similarly, Li et al. [5] develop an aspect-based fashion recommendation system that extracts both local and global features to estimate overall ratings. Since relying on explicit criteria ratings, aspect-based recommendation systems face the challenge that inferred ratings from reviews may not always be accurate. Again, aspect-based recommendation considers only review for preference learning, ignoring the explicit multiple criteria ratings.

In a more recent development, Hasan et al. [12] propose a multi-criteria recommendation model that learns aspect representation from the review and then combines the aspect ratings with the explicit criteria ratings for the overall rating prediction. The most similar research to our proposed model are the ones by Nassar et al. [1] and by Hasan et al. [12]. The major difference between our work and the former is that they solely rely on criteria ratings without considering textual reviews. Although the latter considers both reviews and criteria, they ignore the undefined criteria. In contrast, our model considers predefined, implicit, and undefined criteria for a more comprehensive analysis.

3 Proposed Model

3.1 Preliminaries

Consider a user $u \epsilon U$ provided feedback to an item $i \epsilon I$ with three different modes: review $x \epsilon X$, criteria rating $r_K \epsilon R$, and overall rating $r_0 \epsilon R$ where K the number of predefined criteria. Our first aim is to extract the implicit and undefined criteria and corresponding ratings from the review. In the second step, our goal is to predict the criteria ratings using Neural collaborative filtering. In the third step, we aim to learn an aggregation function using all criteria (undefined, implicit, and predefined) and the overall rating. Finally, the overall rating for the unknown item is predicted using predicted criteria and the learned aggregation function. The architecture of the model is shown in Figure 1.

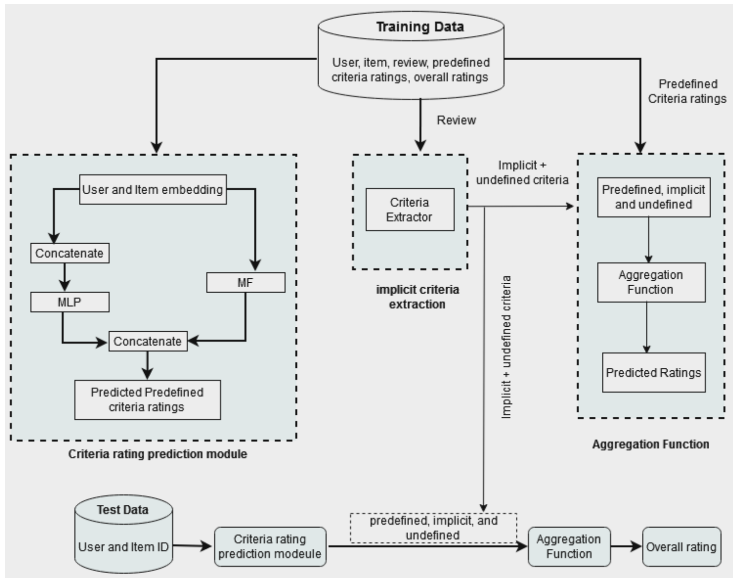

Fig. 1: The detailed architecture of the proposed model.

3.2 Criteria extraction from the review

In this step of the process, we split the review into sentences on punctuation symbol "." , and words such as "but", "while", "however", "nonetheless". Here we assume that when using these conjunctions users start to speak on different aspects. Next, we compute the semantic similarity between each sentence and the predefined criteria using cosine similarity measure. We utilize the pretrained sentence transformer (e.g. SBERT [15]) to generate embeddings for both sentences and the predefined criteria. Sentences are assigned to one of the predefined criteria if the similarity score with that criterion is bigger than a threshold value. Sentences whose cosine similarity scores with all the predefined criteria are less than this threshold value are assigned to one cluster, corresponding to the undefined criteria. We consider these sentences to express users' opinions on aspects that are not covered by predefined criteria. Once the clustering is done, all the sentences in the cluster for each criterion are concatenated, and finally, a sentiment analyzer is used to estimate the rating for each of the clusters.

3.3 Criteria Rating Prediction

Inspired by Nassar et al. [1], we employ Neural Multi-criteria Collaborative Filtering for criteria rating prediction. Since the IDs are categorical features, motivated by He et al. [13], the user ID and item ID are vectorized by leveraging the embedding technique and initialized with random values. Consider the embedding for user u and item i are e_u and e_i respectively. The MF and MLP are fused to learn the user-item interaction.

$$\hat{y}_{ui} = \sigma\left(W^T(e_u \odot e_i) + h^T \begin{bmatrix} e_u \\ e_i \end{bmatrix} + b\right) \tag{1}$$

Rectified Linear Unit (ReLU)[14] and sigmoid σ are the activation function for hidden and final layers respectively.

3.4 Learning Aggregation Function

We employ a DNN-based aggregation function to capture the relationship between the criteria and the overall ratings. Mathematically, overall rating is:

$$r_0 = f(r_1, r_2, ..., r_K) \tag{2}$$

where k is the number of criteria, and r_0 is the overall rating. The final input vector is then passed through several dense layers followed by a *ReLU* layer.

3.5 Overall Rating prediction

The predicted ratings are used as input to the learned function $f(r)$ to predict the final output and are defined as

$$\hat{r}_o = f(\hat{r}_1, \hat{r}_2, ..., \hat{r}_K) \tag{3}$$

where $\hat{r}_o$ is the predicted final overall rating and $\hat{r}_1, \hat{r}_2, ..., \hat{r}_K$ are the predicted criteria ratings. We exploit the regression with squared loss as the objective function

$$J = \sum(\hat{r}_0 - r_0)^2 + \lambda_\theta \|\Theta\|^2 \tag{4}$$

r_0 and $\hat{r}_0$ are true and the predicted ratings respectively, and $\lambda_\theta \|\Theta\|^2$ is used as regularization to prevent the model from overfitting.

4 Experiments

Datasets: We use two publicly available multicriteria datasets: Tripadvisor[3] and BeerAdvocate[4]. The reason we choose these datasets is that they come with three forms of feedback, e.g. textual review, criteria ratings, and overall rating. Tripadvisor.com is a famous hotel and tourism management platform where users can book hotels, flights, and tourist attractions. The TripAdvisor dataset was

[3] https://www.cs.virginia.edu/ hw5x/Data/LARA/TripAdvisor/
[4] https://cseweb.ucsd.edu/ jmcauley/datasets.html#multi_aspect

gathered in 8 years, from 2004 to 2012. The scores for both individual criteria and overall ratings span a scale from -1 to 5.

Baselines: Our comparison is restricted to papers that have openly shared their code. **MRRRec** [12] is a multi-criteria recommendation model that leverages both the criteria ratings and the review to learn the user preferences, demonstrating the performance improvement over several baselines. **ANR**[4] leverages user and item reviews to learn user preferences and item features, increasing performance and explainability. **Deep Cooperative Neural Network (DeepCoNN)** [13] is a state-of-the-art recommendation model that incorporates reviews for predicting ratings. In contrast, **Multi-criteria RS by Nassar et al.** [17] represents a multi-criteria recommendation system based on deep learning. The model predicts overall criteria ratings through a two-step process. **NARRE** [18] employs attention mechanism to capture the representation of the user and item while predicting the overall ratings for the user. **DAML** [19] introduces a dual attention mutual learning method for recommendation by leveraging ratings and reviews. Review features are jointly learned by local and mutual attention, which can increase the interpretability. **MPCN** [13] tries to improve the performance over the state-of-the-art D-ATTN [20] and DeepCoNN model [21].

We evaluate the performance of our model with two error metrics: MSE (Means Squared Error) and MAE (Mean Absolute Error). The reason behind choosing these two metrics is due to their widespread adoption in state-of-the-art rating prediction recommendation models [4, 19, 22, 23]. We further extend our evaluation to precision, recall, and F1 accuracy measures. To compute the precision, recall, and F1, we formulate the problem as a binary classification problem by converting the ratings to 0 and 1 based on a threshold score. In our case, we set this threshold to 3.5. A rating above 3.5 is converted to 1; if it is less than 3.5, it is converted to 0.

4.1 Experimental Settings

We partitioned the datasets into three subsets: training, validation, and test sets, using an 80, 10, 10 ratio. We implemented our model and the baselines using the Pytorch library and T4 GPU from Google Colab Pro. For a fair comparison, we kept the hyperparameters the same as the baseline papers. In our experiments, for clustering tasks, we used different similarity threshold scores, ranging from 0.10 to 0.40 with an increment of 0.05.

5 Results

The model obtains 3.08% and 2.02% lower MSE and MAE respectively compared to the best-performing baseline, i.e., MRRRec, on TripAdvisor dataset (shown in Table 1). For the Beer Advocate dataset, the model achieves a 3.65% and 2.97% lower MSE and MAE respectively. The existing state-of-the-art recommendation models use only MSE and MAE for performance evaluation [4,

Table 1: Performance comparison in terms of MSE and MAE

Performance Comparison				
Model	TripAdvisor		BeerAdvocate	
	MSE	MAE	MSE	MAE
Proposed	**0.3912**	**0.3721**	**0.2921**	**0.3024**
MRRRec	0.4220	0.3923	0.3286	0.3762
NARRE	0.5798	0.5167	0.5165	0.4602
DAML	0.6198	0.5722	0.4918	0.5013
MPCN	0.5526	0.5282	0.4526	0.5282
ANR	0.4431	0.4110	0.4452	0.3921
DeepCoNN	0.6002	0.5535	0.5006	0.4654
Nassar et al.[1]	0.4670	0.3945	0.3979	0.3321

Table 2: Performance comparison in terms of precision, recall, and F1

Performance Comparison						
Model	TripAdvisor			BeerAdvocate		
	Precision	Recall	F1	Precision	Recall	F1
Proposed	**0.9511**	**0.9596**	**0.9553**	0.9095	**0.9421**	**0.9255**
MRRRec	0.9301	0.9490	0.94901	0.8938	0.9110	0.9023
NARRE	0.8521	0.9116	0.8808	0.8935	0.9257	0.9093
DAML	0.9010	0.8793	0.8793	0.8968	0.8813	0.8889
MPCN	0.8323	0.9021	0.9021	0.8776	0.9317	0.9038
ANR	0.9092	0.9521	0.9302	0.9121	0.9001	0.9060
DeepCoNN	0.8725	0.9295	0.9000	0.8582	0.9176	0.8869
Nassar et al.[1]	0.8911	0.9201	0.9053	0.7454	0.9247	0.8254

19, 24, 25] because they use their models for rating prediction tasks. We further extend our evaluations to precision, recall, and F1 scores to understand the recommendation accuracy. Table 2 depicts the accuracy measures of our proposed model and compares them with the baselines. On the Tripadvisor dataset, our model achieves higher accuracy compared to all the baselines in terms of precision, recall, and F1 measures. In comparison to the best-performing baseline, i.e., MRRRec, precision, recall, and F1 are increased by 0.69%, 0.35%, and 0.93% respectively. For the BeerAdvocate dataset, recall and F1 scores are increased by 1.7% and 1.95% respectively. However, precision sees a decrease of 0.25%, indicating a slightly lower performance compared to the ANR model.

5.1 Ablation Study

We analyze the rating prediction performance for five different scenarios: 1) implicit criteria, undefined criteria, and predefined criteria, 2) implicit and predefined criteria, 3) undefined and predefined criteria, 4) undefined criteria and implicit criteria, 5) predefined criteria. The focus in this set of experiments is to study the impact of different criteria ratings on overall ratings. Once we identify the scenario that gives the best performance on predicting overall ratings, we

Table 3: Performance with different scenarios of criteria ratings

Criteria	MSE	MAE
implicit+undefined+predefined	0.2053	0.3226
implicit+predefined	0.2119	0.3251
undefined+predefined	0.2068	0.3226
undefined+implicit	0.7854	0.6942
Only predefined	0.2121	0.3237

use that combination of criteria in our final model. The result is shown in Table 3. It is important to note that we conduct the ablation study on TripAdvisor datasets only, because TripAdvisor datasets are the most widely used datasets in multi-criteria rating research [3, 17, 24]. For each scenario, we predict criteria ratings and learn the aggregation function for overall rating prediction. Among the five cases, we obtain the best performance when we use all the criteria ratings, e.g. implicit, undefined, and predefined criteria ratings shown in Table 3. For all criteria, MSE and MAE are decreased by 0.678% and 0.11% respectively. The worst performance is noted from the undefined and implicit criteria, in which we obtain the MSE value of 0.7854 and the MAE value of 0.6942, respectively. This suggests that relying solely on undefined and implicit criteria extracted from the review is insufficient for capturing the user's overall preferences. Likewise, relying solely on predefined criteria is inadequate, as indicated by MSE and MAE values of 0.2121 and 0.3237, respectively. Therefore, there is a need to combine predefined, undefined, and implicit criteria to learn user preferences.

5.2 Discussion

It is challenging to cluster free-form reviews because of the length and structure of the content. The reason we chose the sentence clustering method using the semantic search technique is that other clustering algorithms (e.g. k-means) require setting a predefined number of clusters which is infeasible due to the variable review lengths. We tried to find a similarity score between sentences and criteria at which model gives the best clusters. We obtained the best performance at a similarity threshold value of 0.30. Each review contains multiple sentences and it is computationally expensive and time-consuming.

Existing approaches utilize either multiple criteria or reviews to capture the relationship between features (e.g. aspect, and criteria) and ratings. We leverage a sentiment analyzer to estimate the criteria ratings from the review. The ratings on predefined criteria and implicit criteria are different. From our results, it is evident that the criteria ratings extracted from reviews have an impact on user preference learning. The results also reveal that it is important to consider users' sentiments on criteria in addition to predefined criteria.

6 Conclusion

We introduce a novel multicriteria recommendation system that utilizes implicit, undefined, and predefined criteria ratings. We show that users' preferences for an item not only depend on explicit ratings given on predefined criteria but also on the implicit and undefined criteria that users express in the review. We employ an aggregation-based method to develop a multi-criteria recommendation system. Predefined, implicit, and undefined criteria are utilized to learn the aggregation function and predict the overall ratings. We show that our proposed method outperforms the existing baselines in terms of both error and accuracy measures on the TripAdvisor dataset and Beer Advocate dataset.

7 Acknowledgments.

This work is partially sponsored by Natural Science and Engineering Research Council of Canada (grant 2020-04760). The authors would like to thank our reviewers for their tremendous efforts in review and the constructive feedback.

References

1. N. Nassar, A. Jafar, and Y. Rahhal, "Multi-criteria collaborative filtering recommender by fusing deep neural network and matrix factorization," Journal of Big Data, vol. 7, no. 1, pp. 1–12, 2020.
2. Q. Shambour, "A deep learning based algorithm for multi-criteria recommender systems," Knowledge-Based Systems, vol. 211, p. 106545, 2021.
3. M. Hong and J. J. Jung, "Multi-criteria tensor model for tourism recommender systems," Expert Systems with Applications, vol. 170, p. 114537, 2021.
4. J. Y. Chin, K. Zhao, S. Joty, and G. Cong, "Anr: Aspect-based neural recommender," in Proceedings of the 27th ACM International conference on information and knowledge management, 2018, pp. 147–156.
5. W. Li and B. Xu, "Aspect-based recommendation model for fashion merchandising," in Digital Marketing & eCommerce Conference. Springer, 2021, pp. 243–250
6. Hasan, E., Rahman, M., Ding, C., Huang, J. X., & Raza, S. (2024). based Recommender Systems: A Survey of Approaches, Challenges and Future Perspectives. arXiv preprint arXiv:2405.05562
7. Hasan, E., & Ding, C. (2023, October). Aspect-Aware Multi-Criteria Recommendation Model with Aspect Representation Learning. In 2023 IEEE International Conference on Web Intelligence and Intelligent Agent Technology (WI-IAT) (pp. 268-272). IEEE
8. C. Wu, F. Wu, J. Liu, Y. Huang, and X. Xie, "Arp: Aspect-aware neural review rating prediction," in Proceedings of the 28th ACM International Conference on Information and Knowledge Management, 2019, pp. 2169–2172
9. W. Wang, D. Liu, X. Liu, and L. Pan, "Fuzzy overlapping community detection based on local random walk and multidimensional scaling," Physica A., vol. 392, pp. 6578–6586, 2013.
10. M. Hong and J. J. Jung, "Clustptf: Clustering-based parallel tensor factorization for the diverse multi-criteria recommendation," Electronic Commerce Research and Applications, vol. 47, p. 101041, 2021.

11. Y.-D. Kim and S. Choi, "Nonnegative tucker decomposition," in 2007 IEEE conference on computer vision and pattern recognition. IEEE, 2007, pp. 1–8.
12. H. Emrul, D. Chen, and C. Alfredo, "Multi-criteria rating and review based recommendation model," Proceedings of the 33 IEEE international conference on Big Data, 2022
13. X. He, L. Liao, H. Zhang, L. Nie, X. Hu, and T.-S. Chua, "Neural collaborative filtering," in Proceedings of the 26th international conference on world wide web, 2017, pp. 173–182.
14. V. Nair and G. E. Hinton, "Rectified linear units improve restricted boltzmann machines," in Icml, 2010.
15. Reimers, N., & Gurevych, I. (2019). Sentence-bert: Sentence embeddings using siamese bert-networks. arXiv preprint arXiv:1908.10084
16. S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in International conference on machine learning. PMLR, 2015, pp. 448–456.
17. N. Nassar, A. Jafar, and Y. Rahhal, "A novel deep multi-criteria collaborative filtering model for recommendation system," Knowledge-Based Systems, vol. 187, p. 104811, 2020.
18. C. Chen, M. Zhang, Y. Liu, and S. Ma, "Neural attentional rating regression with review-level explanations.(2018)," 2018.
19. D. Liu, J. Li, B. Du, J. Chang, and R. Gao, "Daml: Dual attention mutual learning between ratings and reviews for item recommendation," in Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining, 2019, pp. 344–352.
20. S. Seo, J. Huang, H. Yang, and Y. Liu, "Interpretable convolutional neural networks with dual local and global attention for review rating prediction," in Proceedings of the eleventh ACM conference on recom- mender systems, 2017, pp. 297–305.
21. L. Zheng, V. Noroozi, and P. S. Yu, "Joint deep modeling of users and items using reviews for recommendation," in Proceedings of the tenth ACM international conference on web search and data mining, 2017, pp. 425–434.
22. D. Deng, L. Jing, J. Yu, S. Sun, and H. Zhou, "Neural gaussian mixture model for review-based rating prediction," in Proceedings of the 12th ACM conference on recommender systems, 2018, pp. 113–121.
23. D. Rafailidis and F. Crestani, "Adversarial training for review-based recommendations," in Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval, 2019, pp. 1057–1060.
24. N. Nassar, A. Jafar, and Y. Rahhal, "Multi-criteria collaborative filtering recommender by fusing deep neural network and matrix factorization. j big data. 2020; 7: 34."
25. C. Shi, X. Han, L. Song, X. Wang, S. Wang, J. Du, and S. Y. Philip, "Deep collaborative filtering with multi-aspect information in heterogeneous networks," IEEE transactions on knowledge and data engineering, vol. 33, no. 4, pp. 1413–1425, 2019.