Concept-Customizable Recommender System

by Langtian Ma

May 8, 2025

1 Introduction

In our previous work, we demonstrated that the Concept Activation Vector (CAV) [3] method is particularly well-suited for interpreting embedding-based recommender systems, offering both high computational efficiency and semantic clarity. Specifically, we showed that the conceptual sensitivity of a user u to a movie genre C can be computed as

$$S_C(\mathbf{e}_u) = \mathbf{e}_u^\top \mathbf{v}_C,$$

which enables efficient evaluation and directly reflects the user's affinity for genre C. A higher value of $S_C(\mathbf{e}_u)$ indicates a stronger preference for the genre.

To investigate this further, we trained a LightGCN model [2] on the MovieLens 1M dataset [1] and applied the CAV framework. Our results showed that while the model successfully learned meaningful genre-related semantics, the correlation between conceptual sensitivity and the average user ratings for movies in a given genre was weakBuilding on these findings, this project aims to explore two questions arises in the experiment:

1. To what extent does movie genre influence user preferences?

2. How can the CAV method be leveraged to generate customizable and explainable recommendations?

We show that, although some genres display notable correlations with user ratings, the overall relationship is generally weak. Motivated by the semantic interpretability of the CAV framework, we propose a simple yet effective method that allows users to adjust their genre-specific preferences, thereby enhancing recommendation transparency and user agency.

2 Movie Genres and User Preferences

To understand whether conceptual sensitivity aligns with user preferences, we computed the Pearson correlation coefficient between each user's conceptual sensitivity score and their average rating across movies of a given genre. Figure 1 summarizes these results, with bubble size indicating the number of users and color representing the p-value of the correlation.

As shown, over half of the genres exhibit statistically significant but weak positive correlations. However, several genres (e.g., *Western*, *Fantasy*, *Horror*) show either non-significant or near-zero correlation. These results suggest that while genre information does play a role in shaping user behavior for certain genres, it is not a strong universal predictor of user ratings across the board.

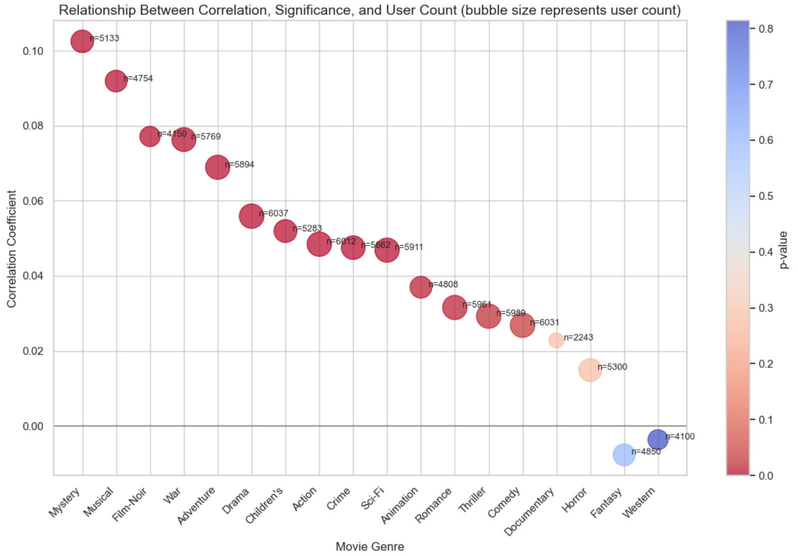

Figure 1. Relationship between conceptual sensitivity and average user ratings across genres. Bubble size indicates user count; color denotes statistical significance.

To further investigate the role of genre semantics, we conducted an ablation study by retraining a LightGCN model without access to genre features. We then fitted a logistic regression model to the learned item embeddings in both the full and reduced settings. The resulting classification performance, as visualized in Figures 3 and 4 in the appendix, indicates that genre classification from embeddings becomes substantially less accurate in the reduced model.

Additionally, we visualized the Concept Activation Vectors (CAVs) for selected genres using PCA (Figure 2). In the full model, CAVs such as *Thriller* and *Horror* are closely aligned and nearly orthogonal to *Drama* and *Comedy*, reflecting intuitive genre similarities. In contrast, the reduced model yields CAVs that lack clear semantic structure.

These findings suggest that while genre does not linearly correlate with user ratings, it plays a critical role in shaping the geometry of the learned embedding space, thereby enhancing the model's ability to encode and differentiate meaningful concepts.

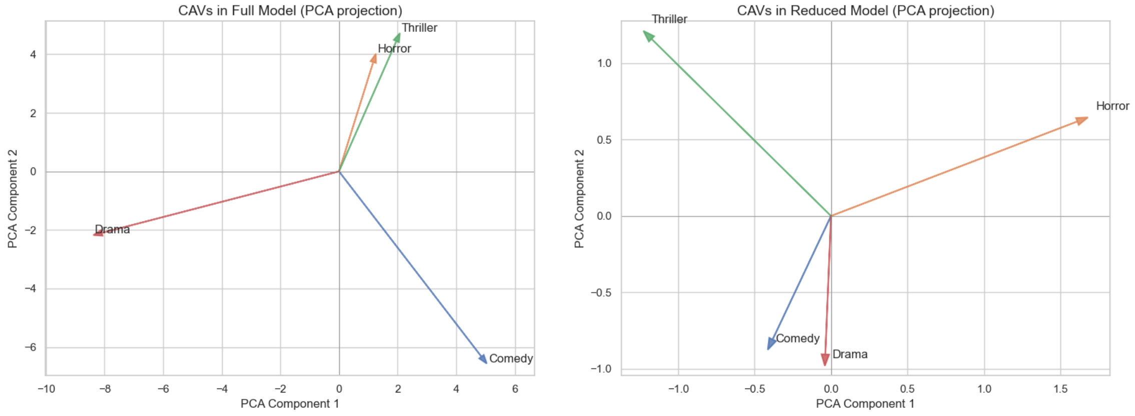

Figure 2. Conceptual activation vectors of full model and reduced model.

3 Concept Customizable Recommendation

Concept Activation Vectors (CAVs) in the learned embedding space provide an interpretable way to allow users to customize their genre-specific preferences. Traditionally, user customization might be achieved by ranking movies based on predicted scores and filtering by genre labels. However, this approach suffers from two key limitations: (1) genre labels are often incomplete or inconsistent—many movies possess characteristics of multiple genres but are labeled with only one; and (2) users are unable to flexibly specify the degree to which they prefer a given genre.

To address these limitations, we propose a CAV-based method that enables smooth and interpretable control over the user's preference vector in the embedding space.

Let $\mathbf{e}_u \in \mathbb{R}^d$ denote the embedding of user u, and let $\mathbf{v}_C \in \mathbb{R}^d$ be the CAV corresponding to genre C. Let $\mathbf{E}_m \in \mathbb{R}^{M \times d}$ be the item embedding matrix, where M is the total number of movies and d is the embedding dimension. We define the genre-adjusted prediction scores as:

$$\hat{\mathbf{y}}_{\text{adj}} = \mathbf{E}_m(\mathbf{e}_u + \gamma\,(\mathbf{v}_C - \mathbf{e}_u)),$$

where $\gamma \in (0,1)$ is a tuning parameter controlling the strength of the user's preference toward genre C.

Intuitively, this approach shifts the user embedding vector toward the direction of the genre CAV, promoting movies that are semantically aligned with that genre. Unlike simple genre filtering, this method leverages the geometry of the embedding space, which captures nuanced relationships among genres—even when explicit genre labels are missing or ambiguous. The parameter γ offers fine-grained control over the extent to which recommendations are biased toward the target genre.

To demonstrate this approach, we apply the adjustment with $\gamma = 0.1$ for a user aiming to receive more recommendations aligned with the "Horror" genre. Tables 1 and 2 display the top-5 recommended movies before and after the adjustment. Notably, while the original recommendations are preserved, additional "Horror" titles such as *House of Exorcism (1974)* emerge in the modified list, reflecting the model's flexibility and semantic awareness.

title	genres
American Beauty (1999)	Comedy|Drama
Sixth Sense(1999)	Thriller
Silence of the Lambs (1991)	Drama|Thriller
Matrix, The (1999)	Action|Sci-Fi|Thriller
Shawshank Redemption (1994)	Drama

Table 1. Original top-5 recommendations.

title	genres
American Beauty (1999)	Comedy|Drama
Sixth Sense, The (1999)	Thriller
Matrix, The (1999)	Action|Sci-Fi|Thriller
House of Exorcism (1974)	Horror
Silence of the Lambs, The (1991)	Drama|Thriller

Table 2. Modified top-5 recommendations with adjusted preference toward "Horror" ($\gamma = 0.1$).

4 Discussion

Our findings reveal that while conceptual sensitivity derived from Concept Activation Vectors (CAVs) shows only weak correlation with users' average ratings for specific genres, genre information nonetheless shapes the semantic structure of the embedding space. Building on this, we proposed a simple and effective method for concept-based recommendation customization, enabling users to flexibly adjust their genre preferences through interpretable embedding manipulation. This approach not only enhances personalization but also offers a transparent mechanism for user control, highlighting the broader utility of concept-level interpretability in recommender systems.

Bibliography

[1] F. Maxwell Harper and Joseph A. Konstan. The MovieLens Datasets: History and Context. 5(4):1-19.

[2] Xiangnan He, Kuan Deng, Xiang Wang, Yan Li, Yongdong Zhang, and Meng Wang. LightGCN: Simplifying and Powering Graph Convolution Network for Recommendation.

[3] Been Kim, Martin Wattenberg, Justin Gilmer, Carrie Cai, James Wexler, Fernanda Viegas, and Rory Sayres. Interpretability Beyond Feature Attribution: Quantitative Testing with Concept Activation Vectors (TCAV). In *Proceedings of the 35th International Conference on Machine Learning*, pages 2668-2677. PMLR.

A Confusion Matrices of Logistic Regression Models

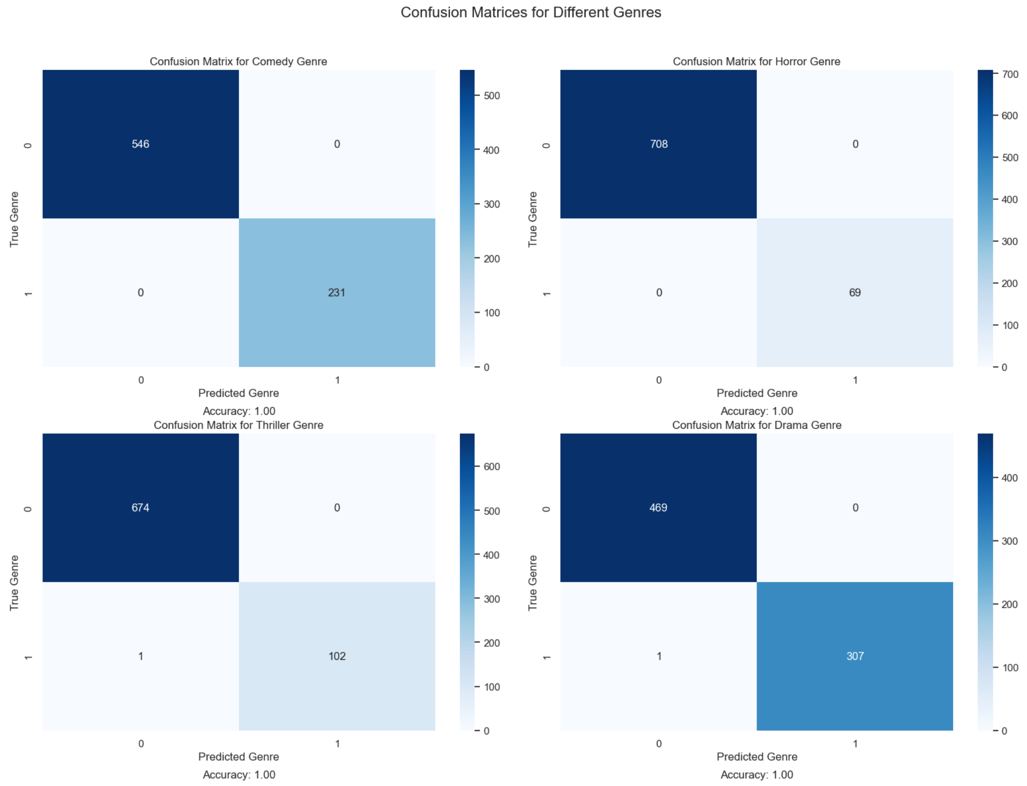

Figure 3. Confusion matrices of logistic regresion trained on the embedding space of the full model.

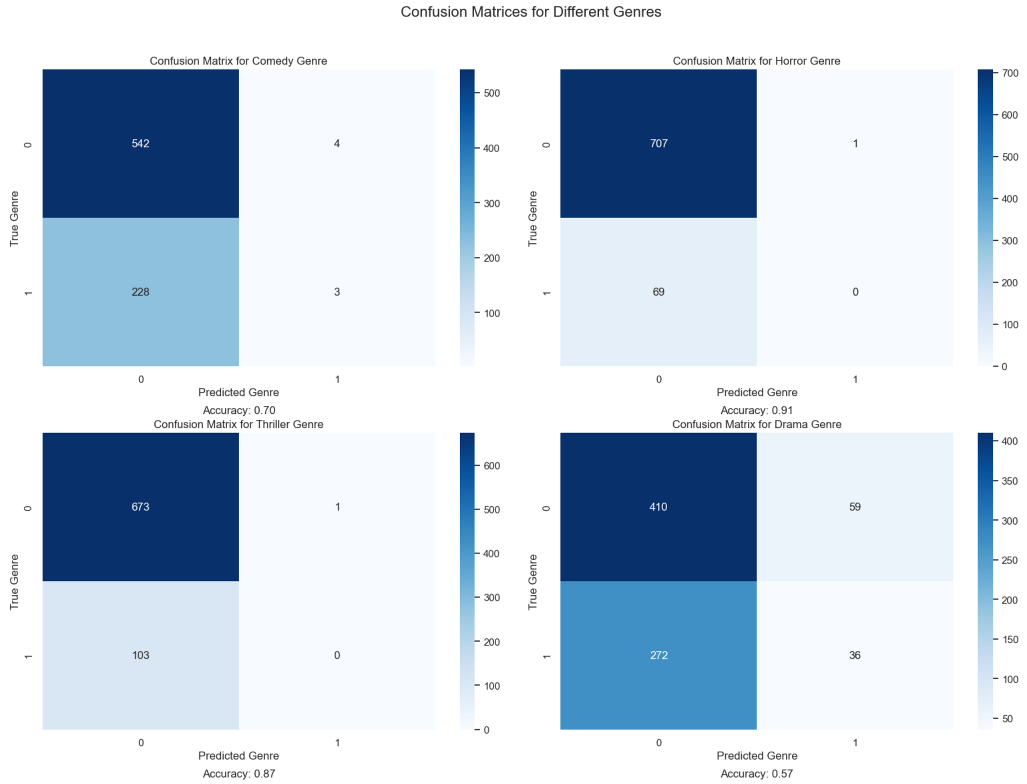

Figure 4. Confusion matrices of logistic regresion trained on the embedding space of the reduced model.