

A Comprehensive Review of Recommendation System in Over-the- Top (OTT) Streaming Platform

Harshdeep Danidhariya¹, Ms Tosal Bhalodia²

1. (B Tech in Computer Science Engineering, Atmiya University, Rajkot, India

Email: harshdeep.danidhariya@gmail.com)

2. (Faculty of Engineering and Technology (CE), Atmiya University, Rajkot, India

Email: tosal.bhalodia@atmiyauni.ac.in

Abstract:

With the rise of digital media, audiences find the choices overwhelming, thereby making personalized recommendations a crucial aspect of streaming services today. OTT services, including Netflix, YouTube, Spotify, and Hotstar, have increasingly deployed recommendation systems to guide users toward content that they feel most attracted to. Going over how the systems have changed—from early days of pure collaborative and content-based filtering to hybrid and deep learning approaches—these systems represent widely used two-stages retrieval and ranking. The study inquires how some of the leading platforms fill in this architecture, solving their particular challenges; be it Netflix's foundation models, YouTube's large-scale neural networks, Spotify's hybrid ecosystem, or Hotstar's real-time, regionally aware personalization. It looks at key challenges, such as cold-start, data sparsity, algorithmic bias, privacy issues, and potential solutions such as diversity-aware ranking and federated learning. Finally, it attempts a look at potential futures—foundation with large language models, explainable AI, and causal inference—which are beginning to shape the recommendation system.

I. Introduction: Drowning in Content, Saved by the Algorithm

The digital media landscape has changed how we consume entertainment. Today, we are not just facing a lot of content; we are dealing with an overwhelming ocean of digital media. Imagine scrolling through a million movies, shows, and songs. It's paralyzing! This is where recommendation systems (RS) come in. These smart algorithms are no longer just features; they are essential for Over-the-Top (OTT) streaming giants like Netflix, YouTube, Spotify, and Hotstar. They predict what you'll

enjoy next, boosting not just engagement but also the crucial fight for user retention.

We will look at the basic techniques, uncover the complex architectures of major platforms, examine the toughest ongoing challenges, and look ahead at the exciting future, including the role of Large Language Models (LLMs).

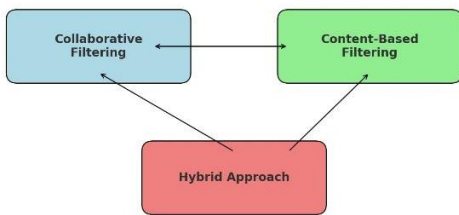
II. Core Philosophy: How Recommenders Learn Your Taste

The recommendation systems employ a few possible methods for connecting content to a user's preference.

Collaborative Filtering (CF): This is the classic "people like you also liked this" approach. The CF method identifies patterns amongst similar users. For example, if you and the other user both really loved Queen's Gambit, the system would go on to suggest you whatever that user watched next. It's a very strong method, often uncovering hidden connections in data through a series of mathematical procedures like matrix factorization. Think about the cold start problem though: How do you recommend anything to a new user with no history? How do you promote a new movie no one has seen yet?

Content-Based Filtering (CBF): It's simpler: recommend items that share characteristics with things you liked before. If you watched three space documentaries, the system might recommend a fourth one based on keywords and metadata. This method is very well applicable for solving cold start as it does not require tons of user history; all it needs is to have some idea about the new user or item.

Core Paradigms of Recommendation Systems



Hybrid Methods & Modern AI: Most of the time the most powerful systems employ a hybrid approach combining CF and CBF; usually Cs-based approaches dominate.

III. Industrial Architecture: The Two Stage Funnel

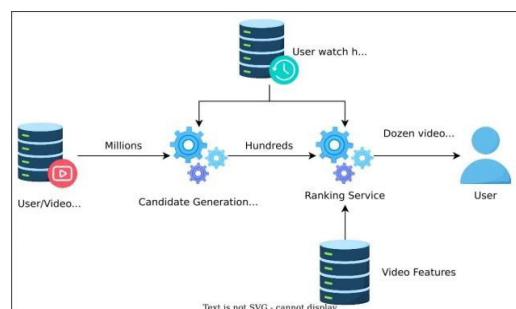
A single big algorithm is unfeasible when you have billions of items and millions of users. Instead, large-scale engines like YouTube and Netflix implement a stupidly simple two-stage algorithm so as to keep the cost and correct balance.

Stage 1: Retrieval (The Sieve)

This system initially sifts through millions of possible items in the entire catalog to narrow down the list efficiently into a much smaller set of a few hundred candidates. This stage uses embedding-based models, representing users and content as points in a high-dimensional space. The closer the points are, the more relevant the item. Like a sieve, it gives priority to creating a relevant list fast.

Stage 2: Ranking (The Prioritize)

The narrow set of candidates is now moved to the Ranking phase, where the system can apply more detailed and computationally expensive models considering very specific features about the user, such as time of day, device, and location, and details about the content and context, to reorder this subset. This final arrangement is the personalized list the user sees on the homepage. A very efficient funnel- powering.



IV. Platform Spotlights: Personalization in Practice

Major streaming platforms have customized these pipelines for their unique content and audience:

Netflix: Knowing matrix factorization techniques (Netflix) are some of the best ways to counter during the Netflix Prize competition, they had since

moved to this multi-model-based system (Netflix Tech Blog, 2023). Their challenge is no more predicting a

rating. It is about how probable, given a set of constraints, is it for a user to actually sit down and watch a title and also to make new titles recognized. This is termed content cold- start problem.

YouTube: Running at an extreme scale, the deep-learning pipeline of YouTube focuses the ranking stage at predicting the watch time, in contrast to clicks like thumbnail view. This small change essentially treasures the system to recommend content which leads to long- term user satisfaction instead of a short engagement trick.

Spotify: This music giant uses a truly hybrid system to power discovery. Collaborative filtering is paired with natural-language-processing (NLP) techniques applied to text found online to get an understanding of a song's vibe, while CNNs are used for an analysis of the audio itself (Spotify Newsroom, 2023). Their goal is to find a balance between chart-toppers and off-the-beaten-path discoveries, so you discover amazing things.

V. Persistent Roadblocks

Yet all recommenders contend with fundamental and ethical hurdles:

The Cold-Start and Sparsity Trap:

Still, this is the major obstacle. With little data deposited with a new user or item, the system must either guess. Hybrid modelling and, to an extent, leverage auxiliary features (like device type or low-level demographic

Bias and The Echo Chamber Effect:

A majority of recommenders suffer from popularity bias—they tend to recommend popular items, thereby creating fewer chances for diversity and forming what is called a "filter bubble." Basically, the user will see content that is only very similar to what they already know. Platforms are, hence, working on diversity- aware ranking and also investigating methodologies to measure the influence of recommendations onto consumption diversity (Ahmed, 2022).

Scaling and Speed:

The delivery of a complex list, based on personalization, in real time, to a global level, demands ultra-efficient pipelines. A constant battle thus gets fought on the engineering side, with one side pushing for ultra-low latency and the other pushing for highly personalized results.

VI. From the next century: transparency and intelligence

The next century will see fields expanding beyond prediction toward intelligent human- centred systems.

Large Language Models (LLMs): Large Language Models set to be powerful "foundation models" of personalization should revolutionize the UI. Imagine a conversational recommender that is able to answer a question such as "Recommend a sci- fi comedy like Guardians of the Galaxy but with a stronger female lead" and justify that answer (Manipal Global Media Journal, 2025).

Explainability and Causal Inference: Users are losing trust in black-box algorithms. Future systems must explain their choices in order to gain trust ("We suggested this because you watched..."). Also, causal

models are being developed so as to measure the long-term impact of recommendations on a user's overall satisfaction and retention, the causes being beyond clicks (Li, 2024).

VII. Conclusion

From the discussion above, it can be concluded that recommendation systems have turned into advanced and critical infrastructure since they enable huge personalization, without which a streaming platform would be unable to survive. Persistent challenges like cold-start, algorithmic bias, and privacy still bother the engineer, but the fast-paced land of deep learning, GNNs, and specially-favoured profile generation systems or emerging powers of LLMs are slowly taking us toward the next generation of recommenders-transparent, contextual, hence more human-centred.

VIII. References

- [1] Li, Y., et al. (2024). Recent developments in recommender systems: A technical survey. *IEEE Transactions on Multimedia Computing, Communications, and Applications*.
- [2] Wu, L., Sun, P., Fu, Y., Hong, R., & Wang, M. (2022). A survey on graph neural networks for recommender systems. *IEEE Transactions on Knowledge and Data Engineering*.
- [3] Gao, C., et al. (2022). Recommender systems based on deep learning: A survey. *IEEE Access*.
- [4] Pourashraf, S., & Mobasher, B. (2023). Culturally-aware recommender systems. *arXiv preprint*.
- [5] Ahmed, A., et al. (2022). Exploring the impact of personalized music recommendations on music consumption diversity: Evidence from a randomized field experiment. *arXiv preprint*.
- [6] Netflix Tech Blog. (2023). Foundation model for personalized recommendation. *Netflix Research*.
- [7] Spotify Newsroom. (2023). Responsibly balancing what goes into your personalized recommendations. *Spotify Research Blog*.
- [8] Manipal Global Media Journal. (2025). AI-driven recommendation systems on OTT platforms.
- [9] Zhang, H., et al. (2023). A comprehensive review of fairness in recommendation systems. *ACM Computing Surveys*.
- [10] He, X., et al. (2022). Advances in neural collaborative filtering. *IEEE Transactions on Knowledge and Data Engineering*.
- [11] Zhao, Y., et al. (2024). Conversational recommender systems: Progress and open challenges. *Information Fusion*.
- [12] Li, Z., et al. (2022). Differential privacy in recommendation systems: Methods and applications. *ACM Transactions on Information Systems*.
- [13] Sun, M., et al. (2023). Graph neural networks for personalized recommendation: A practical perspective. *Expert Systems with Application*

[14] Wang, S., et al. (2023). Large language models for recommendation: Opportunities and challenges. arXiv preprint.

[15] Kumar, A., et al. (2022). Regional personalization in OTT platforms: A case study on Indian audiences. International Journal of Multimedia Information Retrieval.

[16] Spotify Research. (2024). Machine learning for music discovery at scale. Spotify Engineering Blog.

[17] YouTube Research. (2023). Deep neural networks for scalable recommendations.

[18] Zhang, Y., et al. (2025). Mitigating popularity bias in recommendation systems: A causal inference approach. ACM Transactions on Recommender Systems.

[19] Hotstar Engineering. (2023). Real-time personalization at scale: Lessons from live sports. Hotstar Tech Blog