Content-Based Recommendation System Developed from the MovieLens 10M Dataset

Nicelle Sernadilla Macaspac

June 2023

WRITE YOUR ABSTRACT HERE

Introduction

The Big Tech companies use recommendation systems to provide customized suggestions of products to users (Koren 2008, 426; Rocca 2019). These algorithms aim to enhance user experience and build customer loyalty (Koren 2008, 426). Recommendation systems can be developed using content-based methods founded on user and product information, collaborative filtering methods based on user-product interactions, or a combination of both (Rocca 2019).

In 2006, the streaming service company Netflix offered USD 1,000,000 to any team that managed to come up with an algorithm that improved on their internal recommendation system, Cinematch, by at least 10%. The hybrid team of KorBell, Big Chaos and Pragmatic Chaos - Bell-Kor's Pragmatic Chaos - triumphed the challenge (Van Buskirk 2009). The team utilized advanced methods of collaborative filtering, restricted Boltzmann machine and gradient-boosted decision trees, yet Koren of team KorBell emphasized the significance of content information from the baseline predictors in capturing the main effects in the dataset and refining the prediction of the algorithm (Koren 2008, 434; Koren 2009, 9).

This project aims to develop a modest recommendation system based on the content information from the baseline predictors in a similar dataset of the movie recommender MovieLens as part of the capstone in the Professional Certificate Program in Data Science of Harvard Online. In the succeeding sections, we initially examine the dataset and user preferences then subsequently build algorithms to develop the recommendation system.

MovieLens 10M Dataset

MovieLens 10M Dataset (Harper and Konstan 2015)

"The dataset does not only tell us the rating values, but also which movies users rate, regardless of how they rated these movies." (Koren 2008, 428)

"All users selected had rated at least 20 movies." (GroupLens, n.d.)

User Preference

"Typical CF data exhibit large user and item effects - i.e, systematic tendencies for some users to give higher ratings than others, and for some items to receive higher ratings than others." (Koren 2008, 427)

Content-Based Algorithm

$E = mc^2$

<TABLE 1> (Madhugiri 2022)

to that predicts future user ratings with a root mean squared error rate of less than 0.86490

"In practice, we often report the root mean squared error (RMSE), because it is in the same units as the outcomes." (Irizarry 2022)

"if our recommender system is based on a model that outputs numeric values such as ratings predictions or matching probabilities, we can assess the quality of these outputs in a very classical manner using an error measurement metric such as, for example, mean square error (MSE)." (Rocca 2019)

Regularization

"Explainability is another key point of the success of recommendation algorithms. Indeed, it has been proven that if users do not understand why they had been recommended a specific item, they tend to lose confidence in the recommender system." (Rocca 2019)

"In order to combat overfitting the sparse rating data, models are regularized so estimates are shrunk towards baseline defaults. Regularization is controlled by constants, which are denoted as: lambda1, lambda2, … " (Koren 2008, 427)

"The regularizing term lambda avoid overfitting by penalizing magnitudes of the parameters." (Koren 2009, 2)

Recommendation System

Evaluation

Conclusion

"In this paper, we suggested methods that lower the RMSE to 0.8870."

"However, they are entered manually, so errors and inconsistencies may exist." (GroupLens, n.d.)

<LIMIT: RECURRING TIME>

"in others our success is restricted by the randomness of the process, with movie recommendations for example." (Irizarry 2022)

Reference

GroupLens, n.d. "README.txt." Accessed May 29, 2023. https://grouplens.org/datasets/movielens/10m/.

Harper, F. Maxwell, and Joseph A. Konstan. 2015. "The MovieLens Datasets: History and Context." *ACM Transactions on Interactive Intelligent Systems* 5, no. 4: 1-19. https://doi.org/10.1145/2827872.

Irizarry, Rafael A. 2002. *Introduction to Data Science: Data Analysis and Prediction Algorithms with R.* http://rafalab.dfci.harvard.edu/dsbook/.

Koren, Yehuda. 2008. "Factorization Meets the Neighborhood: a Multifaceted Collaborative Filtering Model." https://people.engr.tamu.edu/huangrh/Spring16/papers_course/matrix_factorization.pdf.

Koren, Yehuda. 2009. "The BellKor Solution to the Netflix Grand Prize." https://www2.seas.gwu.edu/~simhaweb/champalg/cf/papers/KorenBellKor2009.pdf.

Madhugiri, Devashree. 2022. "Top 7 Packages for Making Beautiful Tables in R." https://towardsdatascience.com/top-7-packages-for-making-beautiful-tables-in-r-7683d054e541.

Rocca, Baptiste. 2019. "Introduction to Recommender Systems." https://towardsdatascience.com/introduction-to-recommender-systems-6c66cf15ada.

Van Buskirk, Eliot. 2009. "BellKor's Pragmatic Chaos Wins $1 Million Netflix Prize by Mere Minutes." https://www.wired.com/2009/09/bellkors-pragmatic-chaos-wins-1-million-netflix-prize/.