Personalized Fashion Recommendation System

Oluwafemi Shobowale

I. Abstract

This project aims to design the best recommendation system that delivers personalized recommendations to H&M customers based on their transaction history. A series of techniques will be implemented and compared using the transaction history and other information like item information. The algorithms to be used include: a user-based KNN collaborative filter, an item-based KNN collaborative filter, a hybrid collaborative filter, and matrix factorization on the transaction data. Since product description is available, a content-based recommender is explored.
The initial step dealt with converting the transaction data into a utility matrix to create an implicit rating. Giving the data had product price; prices were converted to "1" for buy and "0" for not buy. The user-based CF and item-based CF uses KNN and distance measures to find similar users and items. After working on them separately any hybrid CF is applied to the data. In the hybrid CF KNN similarly users were found based on their purchases using (Jaccard Distance) then items base CF is done in this truncated KNN user data.
Evaluation of the systems was very critical due to sparsity. It is vital to measure all ranking of the systems using mean average precision (MAP)@12. This simply means k=12 and each user would have a recommendation score of 0-12 (12 being the best and 0 the least).

Keywords: recommender system, KNN, item-based, user-based, matrix factorization, evaluation, cluster, distance, similarities

II. Introduction

Recommendation systems solely predict a user's responses to options (products or services). These systems are widely used when users are faced with a wide range of products or services. Recommendation systems are divided into two main groups; content-based systems and collaborative filtering systems. Other techniques include: association rules, log-likelihood methods, and the hybrid method. The concept of a “long-tail” makes recommendations systems vital, not only popular items can be sold, and more products can be sold to people who need them. Content-based systems observe the properties of items to make recommendations, and collaborative filtering systems, on the other hand, recommend products based on the similarity measure between users or objects.

The dataset used is an H&M customer purchase history and item details. The data consists of three files; articles, customers, and transactions. The article file includes all the products available on the H&M website and 25 features about each product (product name, product type, product group name, department number, etc.) The customer section gives a complete insight into customers' information with seven features (customer id, FN, active, club member status, fashion news, frequency, age, and postal code). The transaction data consist of 5 features data, customer id, article id, price, and sales channel. This project aims to build a collaborative filter using both user-based and item-based architecture and a content base and matrix factorization recommendation system to help suggest, predict, and recommend more products to H&M users. The project mainly deals with implementing this system and comparing the results based on performance and evaluation.

III. Methods

A recommendation system consists of two vital classes of entities: the user and items. The users have some preference or want a specific item based on multiple reasons, and these reasons are modeled out of the data. The data is a utility matrix which is the relationship between each user and item. In this project the utility matrix is generated implicitly from the transaction data available. The transaction data shows what item a user buys and is hardcoded with “1” items purchased and “0” to items not purchased. This utility matrix is very sparse. It has more zeros than one.

[1]

[1] , Falk, K. (2019). *Practical recommender systems*. Manning Publications Company.

Similarities among users and products

Collaborative filtering is a process where items are recommended for users based on other users who exhibit similar tastes to the items. Similarity measures include: cosine similarity, Pearson correlation, Jaccard distance, and Euclidean distance. Ideally, for transactional data (unary data) or binary data, Jaccard distance captures the similarity better than other similarity measures.
In this project, the K-nearest neighbor (KNN) helps annihilate the nonsimilar users and items from the large data using the distance measures above.

User-based Collaborative Filtering

This approach finds the K-users similar to a selected user who made comparable purchases with a new user in the utility matrix. It computes a rating on products not bought by the user. In fig1. User 5 has similar purchase history with User 4, User 3, and User 1; these are KNN users to User 5, and recommendations are made on products this user bought.

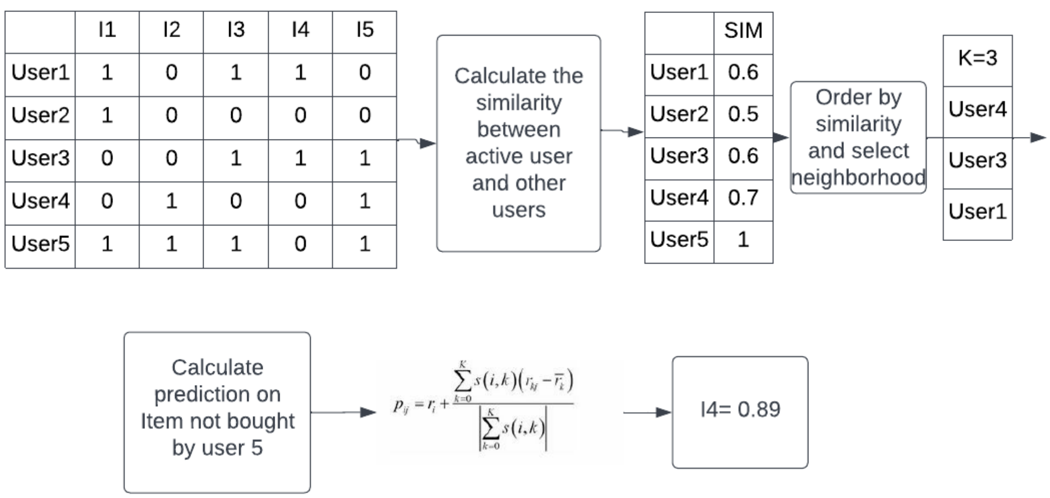

Fig1: Pipeline for user-based KNN collaborative filtering

[2]

[2] 2,Leskovec, J., Rajaraman, A., & Ullman, J. D. (2022). *Mining of massive datasets*. Cambridge University Press

3,Isoni, A., & Cervellin, D. (2016). *Machine learning for the web: Explore the web and make smarter predictions using Python*. Packt Publishing.

Item-based Collaborative Filtering

The item-based CF is similar to the user-based, but the only difference is that similarity is calculated on the items instead of the users. This implies that item similarities can be precomputed, allowing recommender systems to be scalable. The algorithm finds the most similar items and computes the rating on the k items.

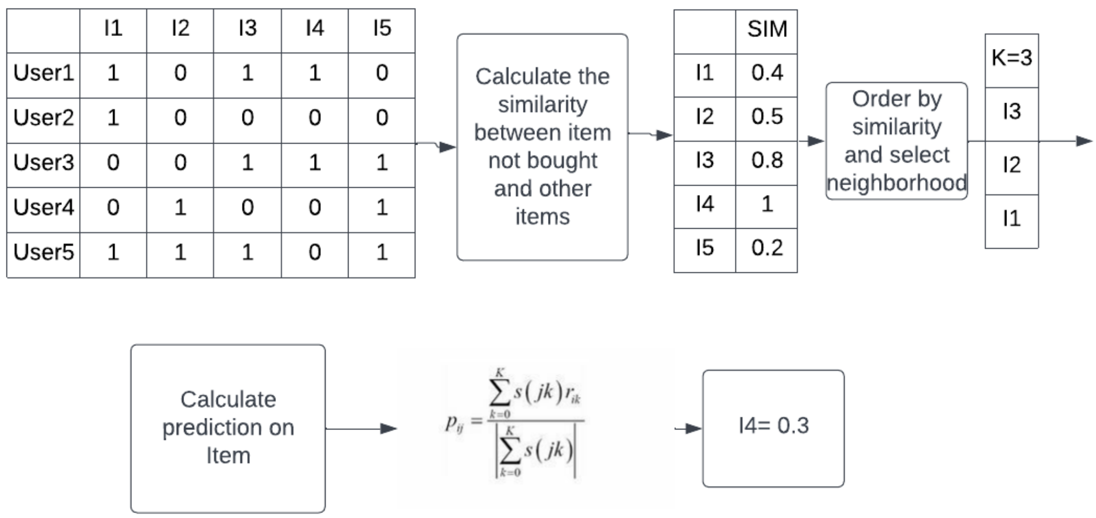

Fig2: Pipeline for user-based KNN collaborative filtering

Hybrid Collaborative Filtering

The hybrid CF uses the user KNN and prediction is made based on the item. This implies that for every KNN user (users with a similar purchase pattern as the active user or new user), we compute the similarities between this new subset and the similarities between the items they purchase. A recommendation is made on items the active user didn't buy. [3]

[3] Harrington, P. (2012). *Machine learning in action*. Manning.

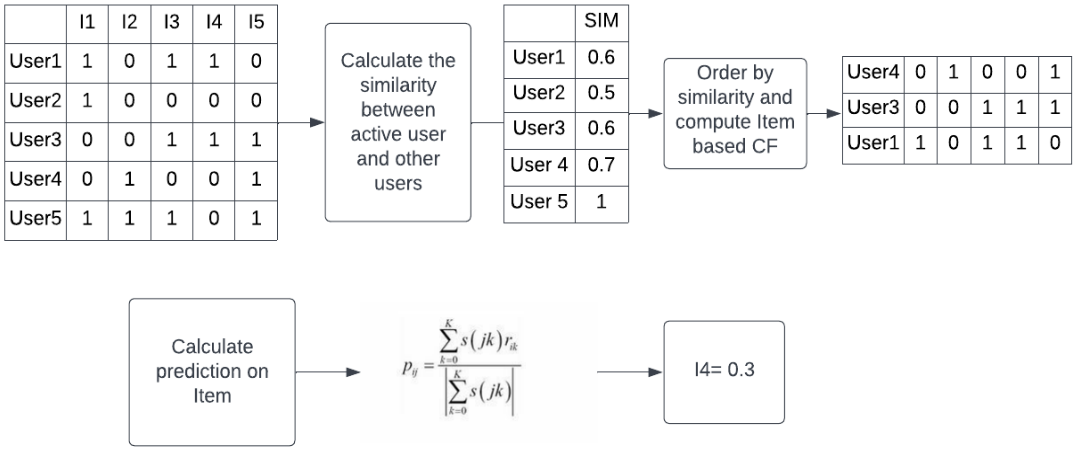

Fig3: Pipeline for Hybrid collaborative filtering

IV. Evaluation and Results

Initial evaluation is done on the nearest neighbors to make sure the neighbors are useful and meaningful, since the official document stated 12 recommendations (k= 12) should be made per user hence, optimizing for K is not an option. Instead the information also states that evaluation done using mean average precision @ 12. Average precision measures how good the rank is by running the precision from 1 to m, where m is the recommended number of referred items. The MAP @ 12 implies that k=12 and each user would have a recommendation score of 0-12 based on relevant recommendation suggested to each user (12 being the best and 0 the least).

MAP is computed for all three algorithms using all similarity measures listed above. KNN- user based with jaccard measure had a more precion score, followed by pearson and euclid. Based on accuracy and computation expense a user based recommender is better.

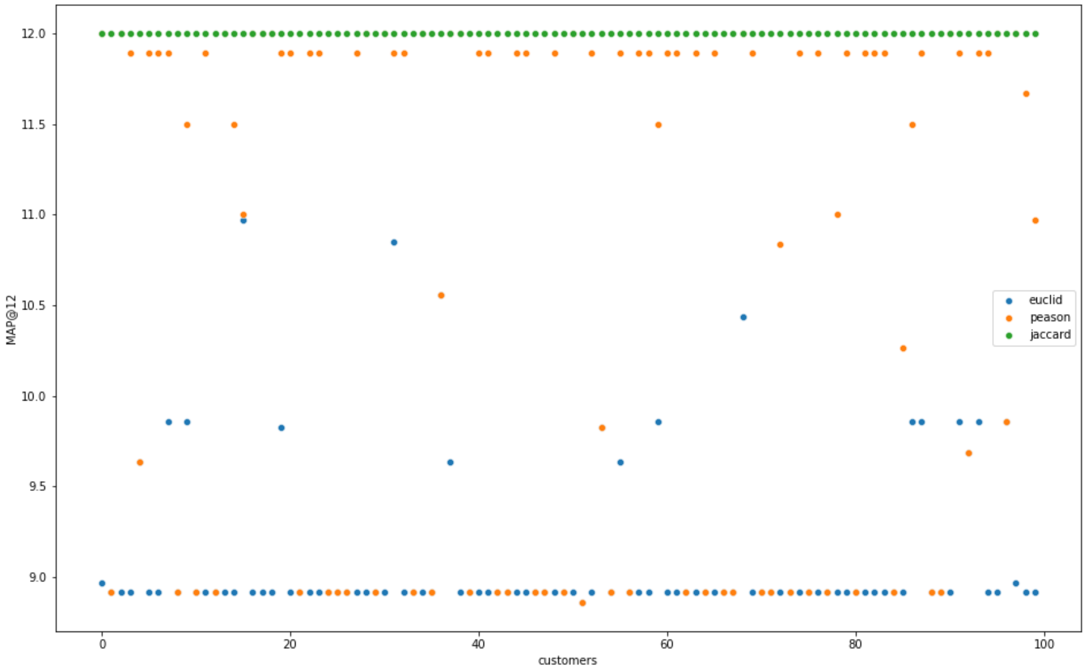

Fig4: MAP@12 evaluation for usee based CF with euclid,peason and jaccard measures