Technology Review: Collaborative Filtering: Applications and Main Challenges

Ji Wu

jw124@illinois.edu
NetID: jw124

November 3, 2022

1. Introduction

Collaborative filtering (**CF**) is a technique used by recommender systems. Collaborative filtering is to use a certain interests, share a common experience of the group's preferences to recommend user information of interest, personal information through cooperation mechanism to give a fair degree of response (e.g., grade) and recorded in order to achieve the purpose of filtering and screening information to help others.(The response “interested” and “not interested” both are important and will impact the recommendation.[3])

2. Overview

The motivation for collaborative filtering comes from the idea that people often get the best recommendations from someone with tastes similar to themselves. Collaborative filtering encompasses techniques for matching people with similar interests and making recommendations on this basis.

Collaborative filtering algorithms often require (1) users' active participation, (2) an easy way to represent users' interests, and (3) algorithms that are able to match people with similar interests.

Typically, the workflow of a collaborative filtering system is:

1. A user expresses his or her preferences by rating items (e.g. books, movies, or music recordings) of the system. These ratings can be viewed as an approximate representation of the user's interest in the corresponding domain.
2. The system matches this user's ratings against other users' and finds the people with most "similar" tastes.
3. With similar users, the system recommends items that the similar users have rated highly but not yet being rated by this user (presumably the absence of rating is often considered as the unfamiliarity of an item)

A key problem of collaborative filtering is how to combine and weight the preferences of user neighbors. Sometimes, users can immediately rate the recommended items. As a result, the system gains an increasingly accurate representation of user preferences over time.[1]

3. Methodology

Collaborative filtering systems have many forms, but many common systems can be reduced to two steps:

1. Look for users who share the same rating patterns with the active user (the user whom the prediction is for).
2. Use the ratings from those like-minded users found in step 1 to calculate a prediction for the active user

This falls under the category of user-based collaborative filtering. A specific application of this is the user-based Nearest Neighbor algorithm.

Alternatively, item-based collaborative filtering (users who bought x also bought y), proceeds in an item-centric manner:

1. Build an item-item matrix determining relationships between pairs of items
2. Infer the tastes of the current user by examining the matrix and matching that user's data

See, for example, the Slope One item-based collaborative filtering family.

Another form of collaborative filtering can be based on implicit observations of normal user behavior (as opposed to the artificial behavior imposed by a rating task). These systems observe what a user has done together with what all users have done (what music they have listened to, what items they have bought) and use that data to predict the user's behavior in the future, or to predict how a user might like to behave given the chance. These predictions then have to be filtered through business logic to determine how they might affect the actions of a business system. For example, it is not useful to offer to sell somebody a particular album of music if they already have demonstrated that they own that music.

Relying on a scoring or rating system which is averaged across all users ignores specific demands of a user, and is particularly poor in tasks where there is large variation in interest (as in the recommendation of music). However, there are other methods to combat information explosion, such as web search and data clustering.[1]

4. Types

1.Memory-based

UserCF

This recommendation process can be roughly divided into six steps:

(1) There are four items in total.

(2) Historically, users A,B,C,D,X have made some visits to the product and have left positive and negative marks (corresponding to green and red). Now it is necessary to use user X 's historical evaluation of the product and other users' historical evaluation of the product to predict whether to recommend one item to user X.

(3)To facilitate calculation, put the user and product into the matrix (called "co-occurrence matrix"), and set the good rating as 1, the bad rating as -1, and the unrated rating as 0 (if there is a specific score, such as 1-5 stars, the score can be used as the element value of the matrix).

(4)Now convert the question about whether the item is recommended to the corresponding value in the matrix. Since it is collaborative filtering, it is necessary to consider n users whose interests are most similar to user X, and then synthesize the evaluation of these n users on the item to obtain the prediction of the interaction between X and the item. (Top n user problem, n is a hyperparameter)

The similarity of two users, that is, in the co-occurrence matrix, the similarity of the corresponding vector of two users can be calculated by the following methods:

(1)Cosine measure

$$\mathrm{sim}(i,j) = \cos(i,j) = \frac{\boldsymbol{i} \cdot \boldsymbol{j}}{\|\boldsymbol{i}\| \cdot \|\boldsymbol{j}\|}$$

The larger the value, the more similar the two users are

(2)Pearson's correlation coefficient

$$\mathrm{sim}(i,j) = \frac{\sum_{p \epsilon P} \left(R_{i,p} - \bar{R}_i\right)\left(R_{j,p} - \bar{R}_j\right)}{\sqrt{\sum_{p \epsilon P} \left(R_{i,p} - \bar{R}_i\right)^2} \sqrt{\sum_{p \epsilon P} \left(R_{j,p} - \bar{R}_j\right)^2}}$$

The larger the value, the more similar the two users are

After obtaining Top n similar users, the weighted average of user similarity and similar users' evaluations is most commonly used to obtain the final prediction

$$R_{u,p} = \frac{\sum_{s \epsilon S} \left(w_{u,s} \cdot R_{s,p}\right)}{\sum_{s \epsilon S} w_{u,s}}$$

$w_{u\text{-}s}$ is the similarity between user u and user s.

$R_{s,p}$ is the rate of user s to user p.

ItemCF

The idea is similar to UserCF, in simple terms, if user A buys both item 1 and item 2, the correlation between item 1 and item 2 is high. When user B also buys item 1, it can be inferred that he also has a need to buy item 2.

Predict to recommend Top k items to user X .

(1) Based on historical data, construct the m×n dimensional co-occurrence matrix of users and items.

(2) Calculate the similarity between each item vector and construct the item similarity matrix with dimension n×n.

(3)Obtain the list of positive feedback items in user X 's historical behavior data.

(4)Use item similarity matrix to find similar Top k items according to positive feedback item list.

(5)The Top k items are sorted by similarity scores to generate the final recommendation list.

[2]

2.Model-based

In this approach, models are developed using different data mining, machine learning algorithms to predict users' rating of unrated items. There are many model-based CF algorithms. Bayesian networks, clustering models, latent semantic models such as singular value decomposition, probabilistic latent semantic analysis, multiple multiplicative factor, latent Dirichlet allocation and Markov decision process based models.

Through this approach, dimensionality reduction methods are mostly being used as complementary technique to improve robustness and accuracy of memory-based approach. In this sense, methods like singular value decomposition, principal component analysis, known as latent factor models, compress user-item matrix into a low-dimensional representation in terms of latent factors. One advantage of using this approach is that instead of having a high dimensional matrix containing abundant number of missing values we will be dealing with a much smaller matrix in lower-dimensional space. A reduced presentation could be utilized for either user-based or item-based neighborhood algorithms that are presented in the previous section. There are several advantages with this paradigm. It handles the sparsity of the original matrix better than memory based ones. Also comparing similarity on the resulting matrix is much more scalable especially in dealing with large sparse datasets.[1]

3.Hybrid

A number of applications combine the memory-based and the model-based CF algorithms. These overcome the limitations of native CF approaches and improve prediction performance. Importantly, they overcome the CF problems such as sparsity and loss of information. However, they have increased complexity and are expensive to implement.Usually most commercial recommender systems are hybrid, for example, the Google news recommender system.[1]

4.Deep-Learning

In recent years a number of neural and deep-learning techniques have been proposed. Some

generalize traditional Matrix factorization algorithms via a non-linear neural architecture, or leverage new model types like Variational Autoencoders. While deep learning has been applied to many different scenarios: context-aware, sequence-aware, social tagging etc. its real effectiveness when used in a simple collaborative recommendation scenario has been put into question. A systematic analysis of publications applying deep learning or neural methods to the top-k recommendation problem, published in top conferences (SIGIR, KDD, WWW, RecSys), has shown that on average less than 40% of articles are reproducible, with as little as 14% in some conferences. Overall the study identifies 18 articles, only 7 of them could be reproduced and 6 of them could be outperformed by much older and simpler properly tuned baselines. The article also highlights a number of potential problems in today's research scholarship and calls for improved scientific practices in that area. Similar issues have been spotted also in sequence-aware recommender systems.[1]

5. Applications in Different Places

Collaboration Filtering is the most common algorithm which is used in various kinds of places. When we see recommendation system in different applications, they always use collaboration filtering to build it. We can see it is used in shopping site like Amazon, eBay, Taobao, JingDong. We can see it used in video websites like YouTube, Netflix. We can also see it in some other social medias like Reddit, twitter, etc.

Specific application
1.Amazon online bookstore:
At first, Amazon use UserCF as their main way to recommend items. However, it will take a really long time if we have too many users and items to build the matrix. When people want to use less data to make the collaborative filtering faster, it will lower the quality of the recommendation. So people find the best way is to use model. To find customers who are similar to the user, cluster models divide the customer base into many segments and treat the task as a classification problem.[5]
But the cluster models will still lower the recommendation quality. As UserCF relies too much on users, Amazon also becomes to user ItemCF. It can compute offline to prepare a list of recommendation for the users. The combination of CF all contributes to the recommendation system.

2.YouTube:
Youtube was collaborative filtering. Collaborative filtering makes predictions for one user based on a collection of data from users with a similar watch history.
For example, if user A and B both watch videos about baking cookies, and user A also watches a video about magic tricks, Youtube’s algorithm may recommend user B videos on magic tricks even if user B hasn’t watched any before.
Youtube’s algorithm also recommends videos based on a user's viewing history, and over-time,

typically becomes very accurate at predicting what a user wants to watch. Youtube does this by comparing videos they had previously recommended with videos that were watched. Watching and ignoring a video tells the algorithm to serve up more or less of that type of content. Your watch history then dictates how high a recommendation appears on your feed through the use of an ML ranking system. You can read more about ranking here.
The first step to Youtube recommending videos that held value to their users was deploying a survey after a user completed their video. Surveys asked users to rate the video from 1 to 5 stars, with only videos rated 4 or 5 stars being deemed valuable. Youtube used machine learning to predict survey results for those who didn't fill out the survey. This was based on collaborative filtering and watch history and engagement.[4]

3.Twitter:
In Twitter, the recommendation for users are different topics. As we all know, users will get the blogs of the people who they do not follow by a few ways. First, if the people like this blog or follow another one, the other users who follow this people will also see that blog. But this blog also has two features. First, it may have more than 10 thousand likes. Second, this people have the same tags as other users that you have followed. In this part, we will see UserCF is one part of the recommendation system. And twitter will show the topics that you may be interested in. This part is always prepared offline and people can directly see it even if they do not post anything or search anyone. Some of the topics will be related to the people you have followed. ItemCF can be the main way to tackle this part.

4.Mandarake:
Mandarake is a big second hand market in Japan. It has specific category and tags for every item. And they also use ItemCF to give recommendation. When the user first log in this site, there will be no recommendation. Once the user begin to find one item and open the specific page, the recommendation bar will appear at the bottom of the page. Clearly, it is from item to item to show other related items by CF. And when people open some items' pages which have many same tags, the recommendation system will shrink the range of the items. So users will be much easier to find the items which are produced by the same people or company and they are in the same kind. When the user leave the page, the website maintain parts of the memory as reference to give the user recommendation next. But the main reference will become the user's cart. When the user open the page again, it will preferentially give recommendations based on user's cart. And a small number of items will be related to the items that user searched last time. As there are more tags for every item. As a user of this website, the experience is better than Amazon.

6. Main Challenges

Here we list some defects of CF
(1) The recommended quality is poor at the beginning of the system
(2) Quality depends on the historical data set
(3) Data Sparsity

(4) System Scalability
(5) The deviation of system caused by malicious evaluation provided by users[2]

Challenge1:
When the user first log in the website, there will have nothing to recommend. And the recommendation quality needs to improve as the user gives more history in this website as a reference. It is impossible to give recommendation without any history. But the main part we still can improve is to use fewer histories to give high quality recommendations. It may depend on every history of the user will have more tags and specific categories. This challenge is related to the first three defects of CF. It requires more pretreatment to improve the quality. And it will result in the challenges in another field.

Challenge2:
As collaborative filtering relies on histories, the data set will enlarge every time. The matrix of algorithm will become bigger and bigger. Apparently, the time to analyze the data will be longer. This is related to the fourth defect. We may use dimensionality reduction to solve or directly discard part of old histories. Both methods may result in the quality's lowering. So how to maintain the user's data and contain a high quality of the recommendation is a big challenge.

Challenge3:
The human factor results in some products have too low grade will also impact the recommendation system. It is not related to CF but the data is not reliable. UserCF will have more negative effects. We may solve it in different ways such as ignore some low grade by the users who only give low grade to some items or never leave the reason for why they give the low grade. Or we can change the ratio of the grade. The user who have high similarity will impact the algorithm more. But we still can not solve this problem completely. This challenge is also related to many other filed including network supervision, network environment, etc.

7. Conclusion

In summary, collaborative filtering is used in different recommendation systems. No matter in market website, video websites or social media, we will have recommendation systems and use Collaborative Filtering. The challenges are related to data, user's histories and the users themselves. It may need more different algorithms to work together to solve. And Collaborative Filtering still have many places to improve.

Reference

[1] https://en.wikipedia.org/wiki/Collaborative_filtering
[2] https://blog.csdn.net/weixin_45884316/article/details/119038148
[3] https://baike.baidu.com/item/%E5%8D%8F%E5%90%8C%E8%BF%87%E6%BB%A4/4732213?fr=aladdin

[4] https://dev.to/mage_ai/youtubes-machine-learning-ml-algorithm-ej0
[5] https://cseweb.ucsd.edu//classes/fa17/cse291-b/reading/Amazon-Recommendations.pdf