CS410: Text Information Systems

Technology Review
Discussion of Hybrid Collaborative Filtering Techniques

Date: 02-November-2020

Prepared by:

Rasbihari Pal(pal9@illinois.edu)

Contents

Introduction

With the grow of internet in past decade digital modernization is everywhere. Current pandemic situation also aided that paradigm shift. Now the internet is part and parcel of Human's life. But there is always too much information available in the internet. This is an era of digital information explosion. To aid the customers in their decision making, business is deploying recommended systems to guide. From online retailers to online movie stores, from travel choices to investment options, the recommender systems are pretty much everywhere. This is an active research area today as new business wants to implement recommender systems as well as existing business wants to improve the recommender system to increase the customer satisfaction.

To build a recommender system, the most two popular approaches are Content based (CB) and Collaborative Filtering (CF). Content-based approach requires a good amount of information of items' own features, rather than using users' interactions and feedbacks. For example, it can be movie attributes such as genre, year, director, actor etc., or textual content of articles that can extracted by applying Natural Language Processing. Collaborative Filtering, on the other hand, doesn't need anything else except users' historical preference on a set of items. Because it's based on historical data, the core assumption here is that the users who have agreed in the past tend to also agree in the future. Content-based techniques have the start-up problem (cold start), in which they must have enough information to build a reliable classifier. Also, they are limited by the features explicitly associated with the objects they recommend (sometimes these features are hard to extract), while collaborative filtering can make recommendations without any descriptive data. Also, content-based techniques have the overspecialization problem, that is, they can only recommend items that score highly against a user's profile or his/her rating history.

While both methods have their own advantages, individually they fail to provide good recommendations in many situations. Incorporating components from both methods, a hybrid recommender system can overcome these shortcomings.

Details

Typically, hybrid CF recommenders are combined by adding content-based characteristics to CF models, adding CF characteristics to content-based models, combining CF with content-based or other systems, or combining different CF algorithms [1]

In this paper, the following hybrid algorithms will be discussed:

1. Hybrid Recommenders - Content-Boosted Collaborative Filtering (CBCF)

The CBCF algorithm [2] uses naive Bayesian text classifier as the content classifier, it then fills in the missing values of the rating matrix with the predictions of the content predictor to form a pseudo rating matrix, in which observed ratings are kept untouched and missing ratings are replaced by the predictions of a content predictor. It then makes predictions over the resulting pseudo ratings matrix using a weighted Pearson correlation-based CF algorithm, which gives a higher weight for the item that more users rated and gives a higher weight for the active user.

This algorithm overcomes the cold start Problem as it can be seen in CF. In pure CF a prediction cannot be made for an item, for the active user, unless it was previously rated by other users. However, we can make such a prediction using a content-based predictor for the user. Using CBCF we can further improve the CB predictions by utilizing the content-based predictions of other users as well. If the neighbors of the active user are highly correlated to it, then their CB predictions should also be very relevant to the user.

This also eliminates the root of the sparsity problem as a pseudo ratings matrix, which is a full matrix, is used. Pseudo user-ratings vectors contain ratings for all items; and hence all users will be considered as potential neighbors. This increases the chances of finding similar users. Thus, the sparsity of the user-ratings matrix affects CBCF to a smaller degree than CF.

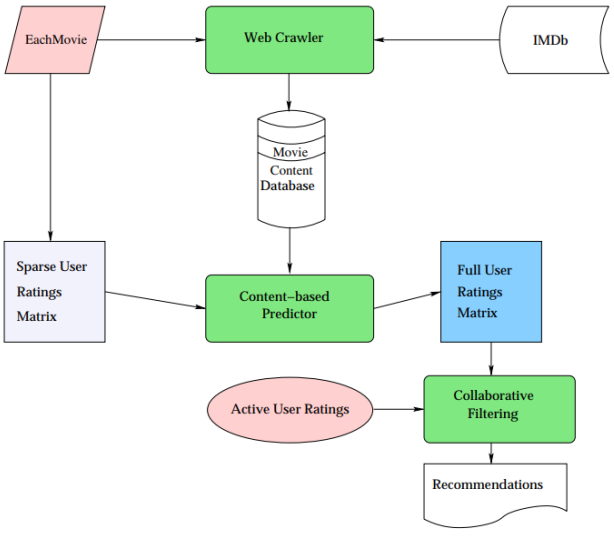

Figure: An example of CBCF for movie recommendation. Reference [Link]

2. Collaborative Filtering by Personality Diagnosis (PD): A Hybrid Memory- and Model-Based Approach

Personality diagnosis (PD) [3] is a representative hybrid CF approach that combines memory-based and model-based CF algorithms and retains some advantages of both algorithms. In PD, the active user is assumingly generated by choosing one of the other users uniformly at random and adding Gaussian noise to his or her ratings. Given the active user's known ratings, one can calculate the probability that he or she is the same "personality type" as other users, and the probability he or she will like the new items. PD retains some of the advantages of traditional similarity-weighting techniques in that all data is brought to bear on each prediction and new data can be added easily and incrementally. Additionally, PD has a meaningful probabilistic interpretation, which may be leveraged to justify, explain, and augment results.

3. Probabilistic Memory-based Collaborative Filtering (PMCF)

PMCF [4] uses a mixture model built based on a set of stored user profiles and use the posterior distribution of user ratings to make predictions. To address the new user problem, an active learning extension to the PMCF system can be used to actively query a user for additional information when insufficient information is available. To reduce the computation time, PMCF selects a small subset called profile space from the entire database of user ratings and gives predictions from the small profile space instead of the whole database. PMCF has better accuracy than the Pearson correlation-based CF and the model-based CF using naive Bayes. [1]

Within the PMCF model, predictions for the active user are thus made by combining the predictions based on other prototype users xi, weighted by their degree of like-mindedness to user a. This puts the key idea of memory-based collaborative filtering into a probabilistic framework.

A schematic drawing of the components of PMCF is shown in following figure.

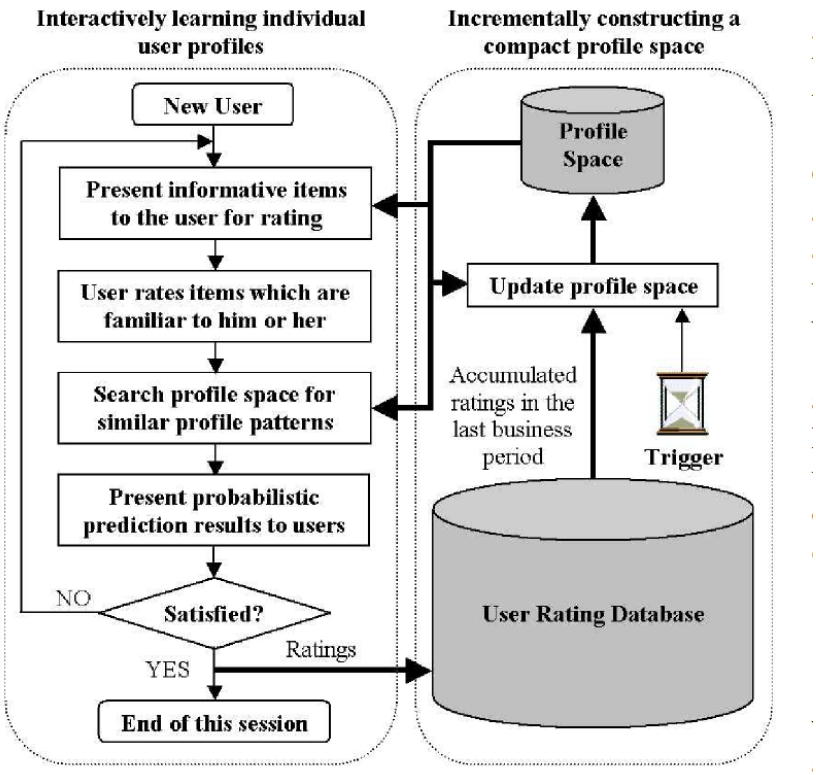

Figure: A schematic drawing of the components of probabilistic memory-based collaborative filtering (PMCF). Reference [Link-Page 3]

Conclusion:

If one compares hybrid recommender systems with collaborative or content-based systems, the recommendation accuracy is usually higher in hybrid systems. The combination of both leads to common knowledge increase, which contributes to better recommendations. The knowledge increase makes it especially promising to explore new ways to extend underlying collaborative filtering algorithms with content data and content-based algorithms with the user behavior data. Techniques like Clustering, Similarity and Classification are used to get better recommendations thus increasing precision and accuracy. This is an active research area.

Although there is no cure-all solution available yet, researchers are working out solutions for each of the problems by coming up with new hybrid models or the extension of existing hybrid algoritms.

References:

[1] Xiaoyuan Su, Taghi M. Khoshgoftaar, "A Survey of Collaborative Filtering Techniques", Advances in Artificial Intelligence, vol. 2009, Article ID 421425, 19 pages, 2009. Link

[2]P. Melville, R. J. Mooney, and R. Nagarajan, “Content-boosted collaborative filtering for improved recommendations” Link

[3] D. M. Pennock, E. Horvitz, S. Lawrence, and C. L. Giles, “Collaborative filtering by personality diagnosis: a hybrid memory- and model-based approach” Link

[4] Yu, Kai & Schwaighofer, Anton & Tresp, Volker & Xu, Xiaowei & Kriegel, H.-P. (2004). Probabilistic Memory-Based Collaborative Filtering. Knowledge and Data Engineering, IEEE Transactions on. 16. 56- 69. 10.1109/TKDE.2004.1264822.