1. Introduction

Nowadays, along with the popularity of web-based e-commerce, collaborative filtering-based recommender systems have become an indispensable part of e-commerce websites. Collaborative filtering is a framework for filtering information based on the preferences of users, that is, CF can predict a user’s preferred products by using the user’s known history data as well as other users’ known history data, and then recommends products to the user. The history data includes users’ previous transaction histories, feedback, and browsing histories. Therefore, a collaborative filtering-based recommender system not only reduces customers’ searching time, it also improves the customers’ satisfaction and loyalty and increases sales.

However, it has some challenging issues such as data sparsity, cold start, new user, synonymy, gray sheep, and shilling attacks [1]. This paper is a survey paper that is going to provide a comprehensive study on collaborative filtering, and its methodology. It will also present CF challenges and methods to overcome these challenges.

2. Methodology

Collaborative filtering techniques are divided into three main categories, each of them use some algorithms for prediction [1].

2.1. Memory-based Collaborative Filtering Technique.

Memory-based collaborative filtering uses some algorithms to utilize the entire user-item database in order to generate a prediction. It uses some statistical techniques to find a set of neighbors who share a similar history. In other words, it finds the users who either rated similar items or purchased similar sets of merchandise and by employing some algorithms to combine the preferences of the neighbors it produces the recommendation for the active user. The nearest-neighbor algorithm is a well-known algorithm that is used with memory-based CF to generate a prediction [2]. This technique and its algorithm will be explained in detail later in the paper.

2.2. Model-based Collaborative Filtering Technique.

Model-based collaborative filtering algorithms provide user recommendations based on learned models. This technique depends on learning concept, that is, the system that can analyze the training data, summarize the complicated patterns into the learned models, and then make predictions based on the learned models. The model building process can be performed by different machine learning algorithms such as Bayesian network, clustering, and rule-based approaches [2]. This technique and its algorithm will be explained in detail in the paper.

2.3. Hybrid Collaborative Filtering Technique

Hybrid CF systems technique combines CF with other recommendation techniques, such
as content-based recommended systems, and demographic-based recommender systems in order to make recommendations. This technique and its algorithm will be explained in detail in the paper [3].

3. Related work or Discussion

3.1 Memory-based CF Related work - recommending music

In this example we will see how the website uses the CF system to recommend music to its user by history data. The KNN algorithm first looks into the training set of data to find k users $U_1, U_2, \ldots, U_k$ which are the closest to the preference $U$. Ties are broken by preferring neighbors $U[i]$ with fewer positive ratings. Artists are then scored according to their popularity in this set of neighboring users[8]:

$$\text{SCORE}(A) = \text{rating}(U[1], A) + \ldots + \text{rating}(U[k], A)$$

3.2 Hybrid CF & Model-based CF Related work - Tivo.

TiVo uses two algorithms for its predicting system: a Bayesian content-based filtering algorithm and a model-based collaborative-filtering algorithm. First the TiVo system will analyze the data volume and then correlates the data which is computed between users using the nearest-neighbor algorithm. However TiVo’s enormous data volume leads to the computational limitation of just using the content-based CF algorithm; therefore the model-based CF is used to mitigate the computational limitation by learning.

4. Conclusion

In this paper, we first introduce CF tasks and their main challenges, and their possible solutions. We then present three main categories of CF techniques: memory-based, model-based, and hybrid CF algorithms with examples for representative algorithms of each category, and analysis of their predictive performance and their ability to address the challenges. From basic techniques to the state-of-the-art, we attempt to present a comprehensive survey for CF techniques, which can be served as a roadmap for research and practice in this area. There are many challenges for collaborative filtering tasks. CF algorithms are required to have the ability to deal with highly sparse data, to scale with the increasing numbers of users and items, to make satisfactory recommendations in a short time period, and to deal with other problems like synonymy (the tendency of the same or similar items to have different names), shilling attacks, data noise, and privacy protection problems. As drawing convincing conclusions from artificial data is risky, data from live experiments are more desirable for CF research. The commonly used CF databases are MovieLens [9], Jester [10], and Netflix prize data [11].

References:


