|  |
| --- |
| INTERNATIONAL UNIVERSITY  PRE-THESIS  REPORT 3: |
| Collaborative Filtering in Social Network |
| **Advisor**: MSc. Vo Duy Khoi |
|  |
|  |
|  |

**Student:** Nguyen Tien Thanh

**ID:** ITIU09076

1. Abstract:

Predicting people who other people may like has recently become an important task in many online social networks. The purpose of this paper is to research and understand about one of the importance techniques that can help online social network deal with this work. And what I want to research about is Collaborative Filtering technique. In this paper, I will come to discuss the primary use of collaborative filtering for online social network as well as the way it works.

1. Introduction:

Currently the growth of the online social network (Facebook, Twitter …) has made it much more difficult to effectively extract useful information from all the various available online information. The overwhelming amount of data necessitates mechanisms for efficient information filtering. One of the techniques used for dealing with this problem is called collaborative filtering.

The motivation for collaborative filtering comes from the idea that people often get the best recommendation from someone with similar taste, similar interest. Collaborative filtering explores techniques for matching people with similar interests and making recommendations on this basis. The basic idea is: If I have a lot of in common with you, I am likely to like what you like.

In detail, Collaborative Filtering is a mechanism used to filter large amounts of information by spreading the process of filtering among a large group of people. In general collaborative filtering is composed out of the steps of filtering and collaborating. Filtering describes the process of performing automated predictions. Input data for filtering is generated by the collaborating process where preference information from many users is put together. Collaborative filtering is characterized by the option to choose from many options, specific user preferences and the assumption that users who agreed on certain items in the past will continue to do so in the future.

Collaborative filtering algorithms often require (1) users’ active participation, (2) an easy way to represent users’ interests to the system, 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 CDs) 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.

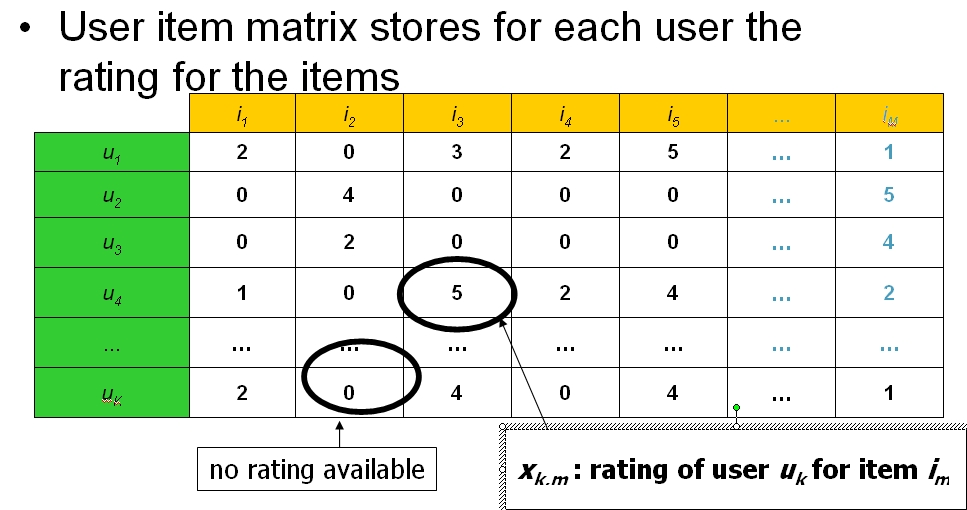
1. Collaborative filtering algorithm:

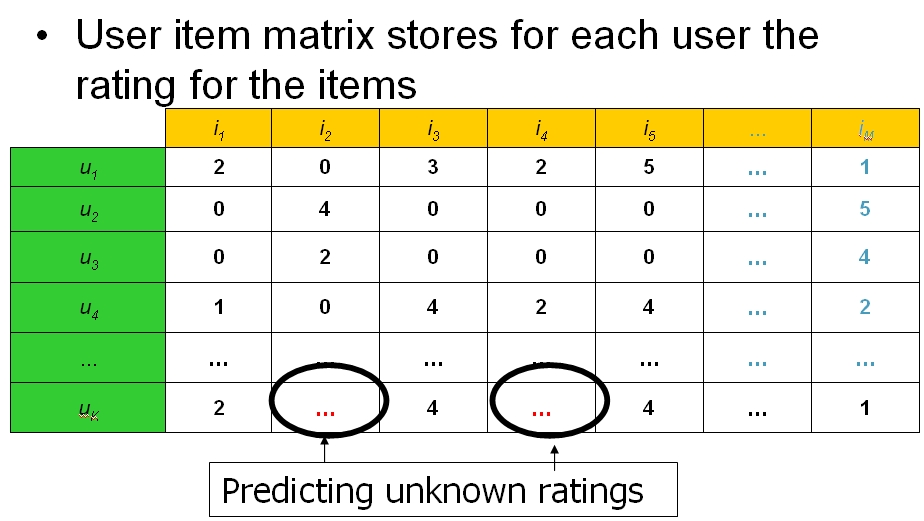
There are numerous ways that collaborative filtering technique uses to filter the information from a large amount of users data.

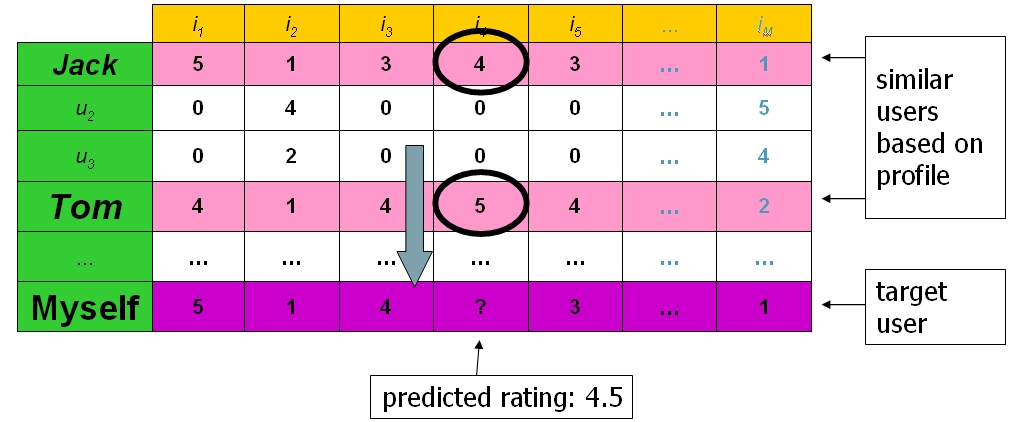
* 1. A matrix or a network.

Collaborating Filtering techniques use a database of preferences for items by users to predict additional topics or products a new user might like. Depend on the user-item matrix it will predict the missing values.

For example: We have a list of m item {i1, i2, …, im} and a list of k users {u1, u2, …, um}. For each user and item, xk.m is the rating of user uk for item im.





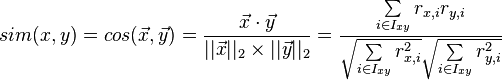


The images above show tables representing the user-item matrix, i.e. each row representing a user and each column representing the user rating for a particular movie. The goal is to predict those ratings the customer would assign to the movie once he has watched and evaluated it. Prediction is based on user profiles which indicate similiar interests, i.e. as inferred by the movies which were rated and which movies actually were watched. In the example above Jack and Tom share similiar movie preferences and we try to predict our own rating for movie I4 which turns out to be 4.5 for this particular case given that Jack and Tom voted 4 and 5 for this movie, respectively.

* 1. Traditional collaborative filtering technique:

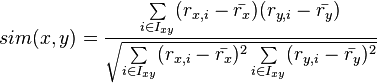
A traditional collaborative filtering algorithm represents a customer as an N-dimensional vector of items, where Nis the number of distinct catalog items. The components of the vector are positive for purchased or positively rated items and negative for negatively rated items. To compensate for best-selling items, the algorithm typically multiplies the vector components by the inverse frequency (the inverse of the number of customers who have purchased or rated the item), making less well-known items much more relevant.

The algorithm generates recommendations based on a few customers who are most similar to the user. It can measure the similarity of two customers, A and B, in various ways; a common method is to measure the cosine of the angle between the two vectors:



Besides, there is also a mechanism that can use to calculate the similarity between two users and items then produces a prediction for the user taking the weighted average of all the ratings. This is Pearson correlation:

The Pearson correlation similarity of two users x, y is defined as:



where Ixy is the set of items rated by both user x and user y.

* 1. K-nearest neighbor:

One of the most common forms of Collaborative Filtering is the nearest neighbor approach. The k-nearest neighbor algorithm is a method of classification or regression that is based on finding the k-closest training data neighbors in the feature space nearest to a target point and combining the information from these neighbor - perhaps in a weighted manner - to determine the classification or regression value for the target point.

There are two main variants of nearest neighbors for collaborative recommendation, user-based and item-based - both methods generally assume that no user or item features are provided, so here x and y are simply respective user and item indices. Given a user x and an item y, let N (x, y) be the set of user nearest neighbors of x that have also given a rating for y, let N (y,x) be the set of item nearest neighbors of y that have also been rated by x, let Sx,z some measure of similarity rating between users x and z, and let Sy,y’ be some measure of similarity rating for items y and y’. So, the predicted rating  that the user x gives item y can then be calculated in one of two ways:

* User-based similarity:



* Item-based similarity:



The question of which approach to use depends on the dataset. When the number of items is far fewer than the number of users, it has been found that the item-based approach usually provides better predictions as well as being more efficient in computations.

Reference:

<http://en.wikipedia.org/wiki/Collaborative_filtering>

<http://en.wikipedia.org/wiki/Pearson_product-moment_correlation_coefficient>

<http://en.wikipedia.org/wiki/Cosine_similarity>

Social Collaborative Filtering Algorithms for Recommendation Report.