Social Network Movie Recommendation System

A BACHELOR’S MINI PROJECT

Submitted in partial fulfillment

of the requirements for the completion of the VII semester

of the

UNDER GRADUATE PROGRAM

in

INFORMATION TECHNOLOGY

(B.Tech in IT)

Submitted by

1. Sourabh Jain (IIT2010038)

2. Aman Kumar (IIT2010043)

3. Ishant (IIT2010024)

Under the Guidance of:

Dr. Shirshu Varma

IIIT-Allahabad



INDIAN INSTITUTE OF INFORMATION TECHNOLOGY ALLAHABAD – 211012 (INDIA)

Oct, 2013

**CANDIDATE’S DECLARATION**

I hereby declare that the work presented in this project entitled “Social Network Movie Recommendation System”, submitted in the partial fulfillment of the completion of the semester VII of Bachelor of Technology (B.Tech) program, in Information Technology at Indian Institute of Information Technology, Allahabad, is an authentic record of my original work carried out under the guidance of Dr. Shirshu Varma . Due acknowledgements have been made in the text of the project to all other material used. This semester work was done in partial compliance with the requirements and constraints of the prescribed curriculum.

Place: Allahabad Sourabh Jain (IIT2010038)

Date:

Aman Kumar (IIT2010043)

Ishant (IIT2010024)

**CERTIFICATE FROM SUPERVISOR**

I do here by recommend that the mini project report prepared under my supervision by Dr. Shirshu Varma titled “Social Network Movie Recommendation System” be accepted in the partial fulfillment of the requirements of the completion of VII semester of Bachelor of Technology in Information Technology for Examination

Date: Prof. Shirshu Varma

Place: Allahabad Faculty designation, IIITA

Committee for Evaluation of the Thesis

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

**ACKNOWLEDGEMENTS**

We would like to express our greatest gratitude to the people who have helped & supported us throughout our project. We are grateful to our mentor Dr. Shirshu Varma for his continuous support for the project, from initial advice & contacts in the early stages of conceptual inception & through ongoing advice & encouragement to this day.   
  
A special thank of ours goes to our colleague who helped us in completing the project.

Date: Sourabh Jain

Place: Allahabad Aman Kumar

Ishant

B.Tech Final Year, IIITA

**ABSTRACT**:

Recommender systems are now popular both commercially and in the research community, where many approaches have been suggested for providing recommendations. In many cases a system designer that wishes to employ a recommendation system must choose between a set of candidate approaches. A first step towards selecting an appropriate algorithm is to decide which properties of the application to focus upon when making this choice. Indeed, recommendation systems have a variety of properties that may affect user experience, such as accuracy, robustness, scalability, and so forth. In this paper we discuss how to compare recommenders based on a set of properties that are relevant for the application.

One of the potent personalization technologies powering the adaptive web is collaborative filtering. Collaborative filtering (CF) is the process of filtering or evaluating items through the opinions of other people. CF technology brings together the opinions of large interconnected communities on the web, supporting filtering of substantial quantities of data. In this paper we introduce the core concepts of collaborative filtering, its primary uses for users of the adaptive web, the theory and practice of CF algorithms.

Table of Contents

1. Introduction………………………………………………………………………………7
   1. Recommender system and social media…………………….…………………..7
   2. Approaches……………………………..………………………………………...8
   3. Problem definition and scope.……………………………………………………8
   4. Core concepts……………………………………………………………………9
2. Collaborative Filtering ………………….………………………………………….…...11

2.1 Beginning of collaborative filtering ……………………………………………..11

2.2 Use of Collaborative filtering...............................................................................12

2.3 Properties for domain suitable for collaborative filtering…………………….13

3. Comparison of collaborative filtering and content based approach…………………..15

4. Development of Algorithm ………………………………………..................................16

* 1. Collaborative filtering algorithm………….………………………………..…16
  2. Practical challenges……………………………………………………………..18

5. Conclusions………………………………………………………………………………19

6. References……………………………………………………………………………….20

**1. INTRODUCTION**

Recommender Systems that target the social media domain, aim at coping with the challenge of social overload by presenting the most attractive and relevant content, aim at increasing adoption and engagement. Often apply personalization techniques.

**1.1 Recommender systems and social media**:

Recommender systems are an augmentation of the social process, in which we rely on advices and suggestions from other people. Any CF system has social characteristics. However, here we focus on the social media domain. Social media and recommender systems mutually benefit from each other. RS can significantly impact the success of social media, ensuring each user is present with her most relevant items that suit her personal needs.

Collaborative Filtering is the process of filtering or evaluating items using the opinions of other people. While the term collaborative filtering (CF) has only been around for a little more than a decade, CF takes its roots from something humans have been doing for centuries - sharing opinions with others.

**1.2 Approaches**

There are typically two types of algorithms for recommender systems.

1. Content-based methods

2. Collaborative filtering

Content-based methods measure the similarity of the recommended item (target item) to the ones that a target user (i.e., user who receives recommendations) likes or dislikes based on item attributes.

On the other hand, collaborative filtering finds users with tastes that are similar to the target user’s based on their past ratings. Collaborative filtering will then make recommendations to the target user based on the opinions of those similar users.

**1.3 Problem Definition and scope**

Given a table with each row corresponding to every user and each column corresponding to each movie. All entries in the table are not available. We predict those undefined ratings using our algorithm.

**1.4. Core Concepts**

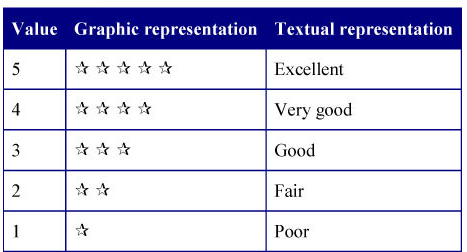
A ratingconsists of the association of two things – user and item – often by means of some value. One way to visualize ratings is as the following matrix:



The term user refers to any individual who provides ratings to a system. Most often, we use this term to refer to the people using a system to receive information (e.g., recommendations) although it also refers to those who provided the data (ratings) used in generating this information.

Collaborative filtering systems produce predictions or recommendations for a given user and one or more items. Items can consist of anything for which a human can provide a rating, such as art, books, CDs, journal articles, or vacation destinations.

1. Scalar ratings can consist of either numerical ratings, such as the 1-5 stars provided in Movie Lens or ordinal ratings such as strongly agree, agree, neutral, disagree, strongly disagree.
2. Binary ratings model choices between agree/disagree or good/bad.
3. Unary ratings can indicate that a user has observed or purchased an item, or otherwise rated the item positively. The absence of a rating indicates that we have no information relating the user to the item (perhaps they purchased the item somewhere else).



Liker Scale

Ratings may be gathered through explicit means, implicit means, or both. Explicitratingsare those where a user is asked to provide an opinion on an item. Implicitratingsare those inferred from a user’s actions.

Ratingsin a collaborative filtering system can take on a variety of forms.

**2. Collaborative filtering**

**2.1 The Beginning of Collaborative Filtering**

As content bases grew from mostly "official" content, such as libraries and corporate document sets, to "informal" content such as discussion lists and e-mail archives, the challenge of finding quality items shifted as well.

Pure content-based techniques were often inadequate at helping users find the documents they wanted. Keyword-based representations could do an adequate job of describing the content of documents, but could do little to help users understand the application of the keywords or the quality of those documents.

Two solutions proposed

1. Wait for improvements in artificial intelligence that would allow better automated classification of documents
2. Bring human judgment into the loop

**The Tapestry system**: Incorporating user actions and opinions into a message database and search system. Tapestry stored the contents of messages, along with metadata about authors, readers, and responders.

Tapestry users could form queries that combined basic textual information (e.g. contains the phrase "recommender systems") with semantic metadata queries (e.g. written by John OR replied to by Joe) and annotation queries (e.g. marked as "excellent" by Chris), Pull active collaborative filtering.

**2.2 Uses of Collaborative Filtering**

In this section we consider this question by exploring user tasks that CF supports, then the services that CF systems provide, and finally, contrasting CF with content filtering.

**User Tasks:**

Tasks for which people use collaborative filtering that have been studied include:

1. **Help me find new items I might like**. In a world of information overload, I cannot evaluate all things. Present a few for me to choose from. This has been applied most commonly to consumer items (music, books, movies), but may also be applied to research papers, web pages, or other ratable items.
2. **Advise me on a particular item**. I have a particular item in mind; does the community know whether it is good or bad?
3. **Help me find a user (or some users) I might like**. Sometimes, knowing who to focus on is as important as knowing what to focus on. This might help with forming discussion groups, matchmaking, or connecting users so that they can exchange recommendations socially.
4. **Help our group find something new that we might like**. CF can help groups of people find items that maximize value to group as a whole. For example, a couple that wishes to see a movie together or a research group that wishes to read an appropriate paper.
5. **Help me find a mixture of “new” and “old” items.** I might wish a “balanced diet” of restaurants, including ones I have eaten in previously; or, I might wish to go to a restaurant with a group of people, even if some have already been there; or, I might wish to purchase some groceries that are appropriate for my shopping cart, even if I have already bought them before.
6. **Help me with tasks that are specific to this domain**. For example, a research paper recommender might also wish to support tasks such as “recommend papers that my paper should cite” and “recommend papers that should cite my paper.” Similarly, a recommender for a movie and a restaurant might be designed to distinguish between recommendations for a first date versus a guys’ night out. Recommenders for some domain-specific tasks have been explored, many have not. To date, much research has focused on more abstract tasks (like “find new items”) while not probing deeply into the underlying user goals (like “find a movie for a first date”).

**2.3 Properties of Domains Suitable for Collaborative Filtering:**

We group these properties below into data distribution, underlying meaning, and data persistence.

**Data Distribution:**

These properties are about the numbers and shape of the data:

1. **There are many items.** If there are few items to choose from, the user can learn about them all without need for computer support.
2. **There are many ratings per item.** If there are few ratings per item, there may not be enough information to provide useful predictions or recommendations.
3. **There are more users rating than items to be recommended.** If there are few ratings per user, you will need many users for efficient recommendations.
4. **Users rate multiple items**. If a user rates only a single item, this provides some information for summary statistics, but no information for relating the items to each other.

**Underlying Meaning:**

These properties are of the underlying meaning of the data:

1. For each user of the community, there are other users with common needs or tastes.
2. Item evaluation requires personal taste.
3. Items are homogenous

**Data Persistence:**

These are properties of how long the data is relevant:

1. **Items persist**, In order for a CF system to generate a prediction for me regarding a recently appeared news story, a typical CF algorithm requires that a) one or more users have rated the story and b) these users have also rated some other stories that I have also rated.
2. **Taste persists**, CF has been most successful in domains where users’ tastes don’t change rapidly: e.g., movies, books, and consumer electronics.

**3. Comparing Collaborative Filtering to Content-Based Filtering:**

1. **Collaborative filtering** uses the assumption that people with similar tastes will rate things similarly**. Content-based filteringuses** the assumption that items with similar objective features will be rated similarly. For **example**, if you liked a web page with the words “tomato sauce,” you will like another web page with the words “tomato sauce.”
2. **Collaborative filtering** needs ratings for an item in order to predict for it. On the other hand, **content-based filtering** needs content to analyze. Collaborative filtering does not require content. A content filtering model can only be as complex as the content to which it has access.
3. **Content-based filtering** may over-specialize. Items are recommended that match the content features in the user's interest profile or query. Items that do not contain the exact features specified in the interest profile may not get recommended even if they are similar (e.g., due to synonymy in keyword terms). Researchers generally believe collaborative filtering leads to more unexpected or different items that are equally valuable.

**4. Development of algorithm**

**4.1 Collaborative Filtering Algorithms**

CF Algorithms are classified into two classes:

1. Memory based algorithms that require all ratings, items, and users are stored in memory.Model-basedalgorithms that periodically create a summary of ratings patterns offline.
2. Non-probabilistic algorithms and probabilistic algorithms.

**Non-probabilistic Algorithms**

The most well-known CF algorithms are nearest neighbor algorithms. We introduce the two different classes of nearest neighbor CF algorithms.

1. User-based nearest neighbor
2. Item-based nearest neighbor

**User-Based Nearest Neighbor Algorithms:**

If a user *n* is similar to a user *u,* we say that *n* is a *neighbor* of *u*. User-based algorithms generate a prediction for an item iby analyzing ratings for ifrom users in u’s neighborhood. Naively, we could average all neighbors’ ratings for item i. Equation 1 gives this average-user formulation, where rniis neighbor n’s rating for item i.

More accurate predictions by weighting ratings from users who are similar to u more heavily. Thus, if user Sim(u,n) is a measure of the similarity between a target user u and a neighbor n, a prediction can be given by equation 2.

Equation 3 normalizes the prediction by dividing by the sum of the neighbors’ similarities.

Users vary in their use of rating scales. That is, one optimistic happy user may consistently rate things 4 of 5 stars that a pessimistic sad user rates 3 of 5 stars. They mean the same thing (“one of my favorite moves”), but use the numbers differently.

**4.2 Practical Challenges of User-Based Algorithms**

Ratings data is often sparse, and pairs of users with few co-ratings are prone to skewed correlations. For example, if users share only three co-rated items, it is not uncommon for the ratings to match almost exactly (a similarity score of 1). If such Similarities are not adjusted, these skewed neighbors can dominate a user’s neighborhood.

Calculating a user’s perfect neighborhood is expensive – requiring comparison against all other users. Thus, in a naïve implementation, the time and memory requirements of user-based algorithms scale linearly with the number of users and ratings.

Techniques to reduce processing time and memory consumption

1. **Sub-sampling:**In sampling, a subset of users is selected prior to prediction computation. Neighborhood computation time remains fixed, and schemes have been proposed to intelligently choose neighbors in order to achieve virtually identical accuracy.
2. **Clustering**:Clustering algorithms have been used to quickly locate a user's neighbors. In these schemes, a user is compared to groups of users, rather than individual users. Clusters of users similar to the target are quickly discovered, and nearest neighbors can be selected from the most similar clusters.

**Conclusions:**

Collaborative filtering is one of the core technologies that will power the adaptive web. Content-based personalization can be effective in limited circumstances.

Filter information based on such complex dimensions, we need to include people in the loop, who analyze the information and condense their opinions into data that can be easily processed by software – ratings.

**References**:

[1]. Adomavicius, G., Tuzhilin A, 2005, “Toward the Next Generation of Recommender Systems: A Survey of the State-of-the-Art and Possible Extensions. IEEE Transactions on Knowledge and Data Engineering”.

[2]. Avery, C., Resnick P., Zeckhauser, R, 1999 “The Market for Evaluations. American Economic Review”.

[3]. Balabanovíc, M., Shoham, Y, 1997, “Content-Based, Collaborative Recommendation*.* Communications of the ACM”.

[4]. Springer-Verlag Berlin Heidelberg, 2007, “Collaborative Filtering Recommender Systems**”**.

Suggestions of the Board members