****

**Movie Recommendation System using Content Based Filtering**

**A Dissertation Proposal**

**Submitted by:**

Bishow Shrestha

Roll No. 13

Bandana Adhikari

Roll No. 8

Sabina Kumari Katwal

Roll No. 30

Indira Songmi

Roll No. 17

**Submitted to:**

Central Department of Computer Science and Information Technology

Tribhuvan University

Kirtipur, Nepal

**Under the supervision of**

Mr. Hiranya Bastakoti

1. **INTRODUCTION**

Recommender Systems (RS) keep growing in popularity in the world of machine learning and predictive modeling. It is ubiquitous and is used in virtually every industry. Due to its direct impact on different sectors and organization. In the business industry where movies are being produced in huge amount. Movies are available all over the world, it is challenging for a user to ﬁnd the appropriate movies suitable for his/her tastes. Different users like different movies or actors based on different genres and soon. It is important to ﬁnd a method of ﬁltering irrelevant movies and/or ﬁnd a set of relevant movies.

Currently, one major area where RS benefits the vast majority of Americans is in movie recommendations. Netflix, an entertainment company, employs over 300 people, spending a staggering $150 million annually, just to generate movie recommendations for viewers based on their previous preferences and movie-watching history. This shows that the money invested in the recommendation system is paying off.

The most well known recommendation systems are mainly based on Collaborative Filtering (CF) and Content-based Filtering. CF ﬁrst tries to ﬁnd out the groups of similar users automatically from a set of active users. The similarities between users are computed using correlation measure. It then recommends items to a user based on the opinions of the users groups. Whereas Content-based filtering, also referred to as cognitive filtering, recommends items based on a comparison between the content of the items and a user profile. The content of item is represented as a set of descriptors or terms, typically the words that occurs in a document. The user profile is represented as with the same terms and built up by analyzing the content of items which have been seen by the user.

1. **PROBLEM STATEMENT**

The learning methods applied to content-based filtering try to find the most relevant documents based on the user’s behavior in past. Such approach however restricts the user to documents similar to those already seen.

Its problem include limited content analysis, overspecialization and new user problem.

Due to which items are limited to their initial descriptions or features.

1. **Literature Review**

Recommendation system have become an important research field, since the emergence of the first paper on collaborative filtering in the mid 1990’s. The purpose of movie recommendation system is to help the users find the relevant movies. It focuses on the ability to numerically estimate the user preferences for unseen items or to provide users with item lists ranked in accordance to the estimated preferences. It has been a challenging task as when there are very large user/items dimensionally it can be very inefficient or hard to pre-compute all the recommendations. Predicting on unseen data means we cannot deal with unseen users or item and to deal with continuously changing features. Much advancement has been made in the literature and technology in last few years, most notably the application of machine learning to learn and extract feature representations from recommendation system.

In [1] “Matrix Factorization Techniques For Recommender Systems”, the authors Mr. Robert Bell and Chris Volinsky, Yehuda Koren, AT&T Labs-Research, presents the method which is one of the popular content filtering approach in the field of Recommendation system. This allows programs to associate users with matching products. Which analyze patterns of user interest in products to provide personalized recommendations that suit a user’s taste and learning algorithms to minimize equation, they are stochastic gradient descent and alternating least squares (ALS) [1].

In [2] “Recommendation system: Principles, methods and evaluation”, the authors F.O. Isinkaye , .O.Folajimi , B.A. Ojokoh, represents the success of any recommendation system. Depends largely on its ability to represent user’s current interests and mitigate some of the problems identified. Hybrid filtering, which combines two or more filtering techniques in different ways in order to increase the accuracy and performance of recommender systems t has been proposed. Using an implicit data as a check on explicit rating or allowing user to give explicit feedback only when he chooses to express explicit interest. Content based does not need the user profile, this technique need to have an in-depth knowledge and description of the features of the items in the profile. [2].

In [3] “Content-Based Recommendation Systems”, the authors Michael J. Pazzani, Daniel Billsus, systems that recommend an item to a user based upon a description of the item and a profile of the user’s interests and a means of comparing items to the user profile to determine what to recommend. The strengths and weaknesses of content-based recommendation systems. The way of representing items in table in row and columns which is a simple database with records [3].

1. **OBJECTIVES**

The objectives of Movie Recommendation system are as follows:

* To assist users in classifying similar movie interests.

1. **SCOPE AND LIMITATIONS**

The scopes of movie recommendation system are as follows:

* This makes use of the information provided by users, analyzes them and then recommends the movie that is best suited to the user at that time.
* Suggesting movies the user will like.

As every invention and technology has the pros and cons. This application, too has some limitations along with many of its own limitations. Some of its limitations are mentioned below:

* User gets recommended what is similar to previous recommendation.
* Whenever new user comes there is lack of information to build solid profile for a user to recommend.

1. **METHODOLOGY**

The research method for this module is discussed in this chapter. The different sections are:

* 1. **Data Collection**

The movies datasets are being collected from the different websites and again at different time at different circumstances. Those collected data are kept at different databases to keep safely as the data provides good environment for future.

* 1. **Input text to Vector**

The input or query input by the user are converted into vector form. It is able to generate corresponding vector according to its semantics in text. And each value in the vector represents a dimension of feature. So for different text, it can calculate their similarity by scores in vectors.

* 1. **Proposed Implementation Model**

In implementation model, we will be implementing the Matrix Factorization Algorithm. The algorithm is used to find the dependences among users and movies. The implementation model is given below:

Input/Query

Vector

Matrix Factorization

Perform Recommendation

Result Display

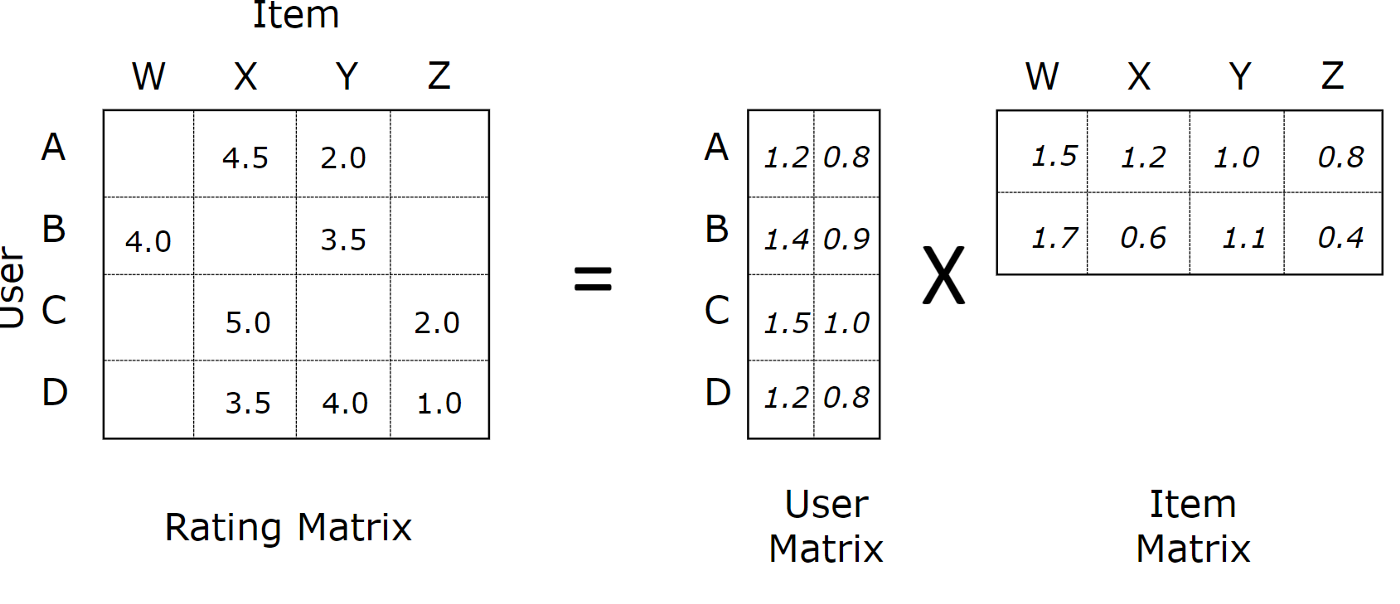
* 1. **Learning and Training**

After the vector conversion and extracting the features the process of learning and training begins. In the training phase recommendation system learns patterns of different domains from input domains. The supervised learning method is applied as learning and training. System is then tested against testing domains and accuracy and efficiency of the system is calculated as per the correctness of the model. We tend to use Matrix Factorization algorithm for all the process.

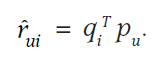
* 1. **Matrix Factorization**

**Matrix factorization** is a class of [collaborative filtering](https://en.wikipedia.org/wiki/Collaborative_filtering) algorithms used in [recommender systems](https://en.wikipedia.org/wiki/Recommender_systems). Matrix factorization algorithms work by decomposing the user-item interaction [matrix](https://en.wikipedia.org/wiki/Matrix_(mathematics)) into the product of two lower dimensionality rectangular matrices.

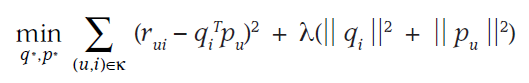
One strength of matrix factorization is the fact that it can incorporate implicit feedback, information that are not directly given but can be derived by analyzing user behavior. Using this strength we can estimate if a user is going to like a movie that (he/she) never saw. And if that estimated rating is high, we can recommend that movie to the user.

****

The concept of matrix factorization can be written mathematically to look something like below.

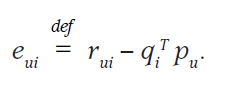
****

Then we can create an objective function (that we want to minimize) with respect to q and p, which are (m, k) and (k, n) matrices.

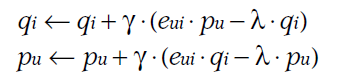


The term on the right is the regularization term, this is added since we do not want our decomposed matrix q and p to over-fit to the original matrix. Since our goal is to generalize the previous ratings in a way that predicts future, unknown ratings, we should not over-fit our model.

**Learning Methods**

****

One obvious method to find matrix q and p is the gradient descent method. Since we have the loss function defined, take the partial derivative respect to q and p to optimize those values.

****

By taking partial derivatives, the update rule would look something like above. But the error surface is not convex, we can also take the alternative approach in which we fix q and p alternatively while optimizing for another.

1. **Project Schedule**

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Year | | 2019 | | | | | | | |
| Activity | 1 Week | 2 Week | 3 Week | 1  Week | 3 Week | 2 Week | 2 Week | 2  Week | 1 Week |
| Collection of Literature |  |  |  |  |  |  |  |  |  |
| Study of Literature |  |  |  |  |  |  |  |  |  |
| Analysis of proposed Scheme |  |  |  |  |  |  |  |  |  |
| Data  Collection |  |  |  |  |  |  |  |  |  |
| Preparation Of Model |  |  |  |  |  |  |  |  |  |
| Implementation and debugging |  |  |  |  |  |  |  |  |  |
| Analysis and Simulation |  |  |  |  |  |  |  |  |  |
| Report Writing and Documentation |  |  |  |  |  |  |  |  |  |
| Presentation& Submission |  |  |  |  |  |  |  |  |  |

# **EXPECTED OUTCOME**

At the end of this project, the proposed model is to be expected to recommend the movie based on the Matrix Factorization. More over the accuracy of the proposed model is based on the calculation of the error between the given output and calculated input, then partial derivative is performed to optimize the value.

1. **REFERENCES**

[1] F.O. Isinkaye , .O.Folajimi , B.A. Ojokoh , *Recommendation system : Principles ,methods and evaluation*, Received 13 March 2015; Revised 8 June 2015; Accepted 30 June 2015; Available online 20 August 2015 .

[2] Rahul Katarya , Om Prakash Verma , *Egyptian Informatics Journal*

Volume 16, Issue 3 , Pages 261-273, November 2015.

[3] Jieun Son , Seoung Bum Kim , *Content-based filtering for recommendation systems using multiattribute networks* , Received 21 April 2017; Revised 1 August 2017; Accepted 2 August 2017; Available online 3 August 2017.

[4] Robert Bell , Yehuda Koren , Chris Volinsky , *Matrix Factorization Techniques for Recommender Systems*, pp. 30-37, vol. 42, August 2009.

[5] Himang Sharatun, *Matrix Factorization Intuition for Movie Recommender System*, April 1, 2018.

[6] Simple Matrix Factorization example on the Movielens dataset using Pyspark. (2018). Medium. Retrieved 22 November 2018, from <https://medium.com/@connectwithghosh/simple-matrix-factorization-example-on-the-movielens-dataset-using-pyspark-9b7e3f567536>