HYBRID RECOMMENDATION SYSTEM FOR MOVIES

Synopsis Report submitted in partial fulfillment
of the requirement for the degree of
B. E. (Information Technology)

Submitted By

Aniket Mirajkar

Rohan Nayak

Jeetesh Rokade

Under The Guidance Of

Prof.Girish Wadhwa

Department Of Information Technology

Vidyalankar Institute Of Technology
Wadala(East),Mumbai-400037

University Of Mumbai
2017-2018

CERTIFICATE OF APPROVAL
For
Project Synopsis

This is to Certify that

Aniket Mirajkar
Rohan Nayak
Jeetesh Rokade

Have successfully carried out Project Synopsis work entitled
HYBRID RECOMMENDATION SYSTEM FOR MOVIES

in partial fulfillment of degree course in
Information Technology
As laid down by University of Mumbai during the academic year
2017-18

Under the Guidance of
Prof. Girish Wadhwa

Signature Of Guide Head Of Department

Examiner 1 Examiner 2 Principal

Declaration

We declare that this written submission represents our ideas in our own words and where others' ideas or words have been included, we have adequately cited and referenced the original sources. We also declare that we have adhered to all principles of academic honesty and integrity and have not misrepresented or fabricated or falsified any idea/data/fact/source in our submission. We understand that any violation of the above will be cause for disciplinary action by the Institute and can also evoke penal action from the sources which have thus not been properly cited or from whom proper permission has not been taken when needed.

Sr.	Name Of Student	Roll No	Signature
1.	Aniket Mirajkar	14101B0039	
2.	Rohan Nayak	14101B0037	
3.	Jeetesh Rokade	14101B0032	

Date :

Acknowledgement

We are profoundly grateful to our guide Prof. Girish Wadhwa for his expert guidance and continuous encouragement throughout the course of the project to ensure that the project reaches its target, from its commencement to its completion

We would like to express deepest appreciation towards Prof. Ajitkumar Khachane, Head of Department of Information Technology and Prof. Deepali Nayak, Project Coordinator whose invaluable guidance supported us in completing this project.

At last, we must express our sincere heartfelt gratitude to all the staff members of Information Technology Department who helped us directly or indirectly during this course of work.

Contents

Abstract

Our project Hybrid Recommendation System for Movies uses the combination of collaborative and content based filtering in the context of web-based recommender systems. In particular, we will link the well-known MovieLens rating data with supplementary IMDB content information. The resulting network of user-item relations and associated content features will be converted into a unified mathematical model, which is applicable to our underlying neighbor-based prediction algorithm. By means of various experiments, we will demonstrate the influence of supplementary user as well as item features on the prediction accuracy of our proposed hybrid recommender. In order to decrease system runtime and to reveal latent user and item relations, we will factorize our hybrid model via singular value decomposition (SVD). Due to the enormous amount of information available online, the need for highly developed personalisation and filtering systems is growing permanently. Recommendation systems constitute a specific type of information filtering that attempt to present items according to the interests expressed by a user. Most web recommendation systems are employed for e-commerce applications or customer adapted websites, which assist users in decision making by providing personalized information.

Chapter 1

Introduction

Due to the enormous amount of information available online, the need for highly developed personalization and filtering systems is growing permanently. Recommender systems constitute a specific type of information filtering that attempt to present items according the interests expressed by a user. Most web recommenders are employed for e-commerce applications or customer adapted websites, which assist users in decision making by providing personalized information.Modern recommendation systems make use of two basic types of recommendation techniques, namely **content-based filtering** and **collaborative filtering**.

Collaborative filtering aims at predicting the user interest for a given item based on a collection of user profiles. Commonly, these profiles either result from asking users explicitly to rate items or are inferred from log-archives. Research started with memory-based approaches to collaborative filtering, that can be divided in user-based approaches like and item-based approaches. Given an unknown test rating (of a test item by a test user) to be estimated, memory-based collaborative filtering first measures similarities between test user and other users (user-based), or, between test item and other items (item-based). Then, the unknown rating is predicted by averaging the (weighted) known ratings of the test item by similar users (user-based), or the (weighted) known ratings of similar items by the test user (item-based). In both cases, only partial information from the data embedded in the user-item matrix is employed to predict unknown ratings (using either correlation between user data or correlation between item data). Because of the sparsity of user profile data however, many related ratings will not be available for the prediction.

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 each item is represented as a set of descriptors or terms, typically the words that occur in a document. The user profile is represented with the same terms and built up by analysing the content of items which have been seen by the user. Several issues have to be considered when implementing a content-based filtering system. First, terms can either be assigned automatically or manually. When terms are assigned automatically a method has to be chosen that can extract these terms from items. Second, the terms have to be represented such that both the user profile and the items can be compared in a meaningful way.The information source that content-based filtering systems are mostly used with are text documents. A standard approach for term parsing selects single words from documents. The vector space model and latent semantic indexing are two methods that use these terms to represent documents as vectors in a multi-dimensional space. Relevance feedback, genetic algorithms, neural networks, and the Bayesian classifier are among the learning techniques for learning a user profile. The vector space model and latent semantic indexing can both be used by these learning methods to represent documents. Some of the learning methods also represent the user profile as one or more vectors in the same multi-dimensional space which makes it easy to compare documents and profiles. Other learning methods such as the Bayesian classifier and neural networks do not use this space but represent the user profile in their own way.

In our project we are going to combine both strategies into one hybrid approach, which utilizes supplementary content features in order to improve the prediction accuracy of traditional collaborative filtering. We are going to make use of dataset which is derived by joining the well-known MovieLens ratings database with the IMDB movie database. Both the databases will be joined in a unique mathematical model which will give us a set of interdependencies. This model enables us to extract latent user/item relations by means of matrix factorization. This factorization will be achieved by means of a technique called as singular value decomposition. The dimensionally reduced data can be employed to directly estimate unknown ratings (pure SVD approach) or rather to accelerate collaborative filtering.

Chapter 2

Aim and Objective

1. The main aim would be to develop a hybrid recommender system which incorporates and enhances properties of existing recommendation systems along with a new approach in order to decrease system runtime and to reveal latent user and item relations with great accuracy.

2. Developing a popularity score which will help users judge the movie in a better way and success prediction for movies before release will provide better feedback to movie makers.

Chapter 3

Literature Surveyed

1.Unifying User-based and Item-based Collaborative Filtering Approaches by Similarity Fusion

- Traditional memory-based collaborative filtering methods assign ratings for recommendations by averaging ratings between similar users or items. In both cases, only partial information from the user-item matrix is used for predicting ratings because of data sparsity problem (lack of ratings by similar user or similar items).

- The paper suggests fusion of predictions from 3 sources: prediction using same user on other items, prediction using other users on same item and prediction based on other similar users on other similar items i.e. SIR, SUR and SUIR. This method of similar fusion uses SIR and SUR for base predictions while SUIR is used for smoothing the predictions.

- The method makes use of 2 parameters in the probability estimation- and to control the weightage of SIR, SUR and SUIR. When equals one, the algorithm corresponds to a user-based approach, while equal to zero results in an item-based approach. Tuning parameter controls the impact of smoothing from the background model (i.e. SUIR).

2.Improved Neighborhood-Based Algorithms for Large-Scale Recommender Systems

The paper suggests an improved approach for evaluating neighbourhood relationships to calculate similarities between users and items. To improve efficiency and reduce memory requirement, it suggests matrix factorization using RMF (regularized matrix factorization) to produce an approximation of the rating matrix.

Data is pre-processed to remove global effects to improve efficiency of the neighbourhood-based approach. It describes an alternate approach where the matrix of similarities between items is learned by the machine itself.

Pearson correlation is used for measuring similarity of two users or items. This scaling of neighbourhood-based methods makes the algorithms scalable to large-scale problems.

3.Hydra- A hybrid recommendation system

- This paper proposes a novel approach to rating prediction, in which collaborative filtering and content-based filtering techniques are combined. This hybrid approach is special in that rating data as well as content information are joined in a unified model, which leads to less parameters and more reasonable prediction results.
- In particular, they describe the linkage of the well-known MovieLens rating data and the IMDB movie information. By means of various experiments, they demonstrate that the retrieved user and movie content features are beneficial to the prediction accuracy of traditional collaborative filtering algorithms.
- For the purpose of minimizing the runtime of our designed hybrid recommender system as well as to extract latent user and movie relations, they factorize our unified model by means of singular value decomposition. The dimensionally reduced data can be employed to directly estimate unknown ratings (pure SVD approach) or rather to accelerate collaborative filtering (SVD-KNN as well as HYBSVD-KNN algorithm)

.

4.Improved Neighborhood-Based Algorithms for Large-Scale Recommender Systems

- This article proposed several neighborhood-based algorithms for large-scale recommender systems. An important property of these algorithms is that their memory usage scales linearly with the number of users or items as compared to a quadratic scaling of most other neighborhood based approaches. This makes the algorithms scalable to large-scale problems. To date it seems that powerful solutions for collaborative filtering problems need to combine the predictions of a diverse set of single algorithms.
- To date it seems that powerful solutions for collaborative filtering problems need to combine the predictions of a diverse set of single algorithms.

5.Application of dimensionality reduction in recommender systems

- This study shows that Singular Value Decomposition (SVD) may be such a technology in some cases. They tried several different approaches to using SVD for generating recommendations and predictions, and discovered one that can dramatically reduce the dimension of the ratings matrix from a collaborative filtering system.
- The SVD-based approach was consistently worse than traditional collaborative filtering in se of an extremely sparse e-commerce dataset. However, the SVD-based approach produced results that were better than a traditional collaborative filtering algorithm some of the time in the denser MovieLens data set.
- This technique leads to very fast online performance, requiring just a few simple arithmetic operations for each recommendation. Computing the SVD is expensive, but can be done offline. Also, there are many other ways in which SVD could be applied to recommender systems problems, including using SVD for neighborhood selection, or using SVD to create low-dimensional visualizations of the ratings space

6.Burke. Hybrid web recommender systems

- This paper surveys the space of two-part hybrid recommender systems, comparing four different recommendation techniques and seven different hybridization strategies.
- Such a component can be combined in numerous ways to build hybrids and in fact, some of the best performing recommenders seen in these experiments were created by using the knowledge-based component as a secondary or contributing component rather than as the main retrieval component.
- In search of better performance, researchers have combined recommendation techniques to build hybrid recommender systems.
- The study finds that cascade and augmented hybrids work well, especially when combining two components of different strengths Three general results, however, can be seen.
- Second, cascade recommendation, although rare in the hybrid recommendation literature, turns out to be a very effective means of combining recommenders of differing strengths.
- Adaptive web sites may offer automated recommendations generated through any number of well-studied techniques including collaborative, content-based and knowledge-based recommendation.
- None of the weak/strong cascade hybrids that were explored ranked less than eighth in any condition, and in the average results, they rank in four of the top seven positions.
- Adopting this approach requires treating the scores from a primary recommender as rough approximations, and allowing a secondary recommender to fine-tune the results.
- First, the utility of a knowledge-based recommendation engine is not limited strictly to its ability to retrieve appropriate products in response to user queries.
- Finally, the six hybridization techniques examined have very different performance characteristics.

Chapter 4

Problem Statement

Formally speaking we aim to develop a recommendation system that enhances the properties of existing system with a newer and a more efficient approach that reduces the system run time and determine item relations with a greater accuracy.

Chapter 5

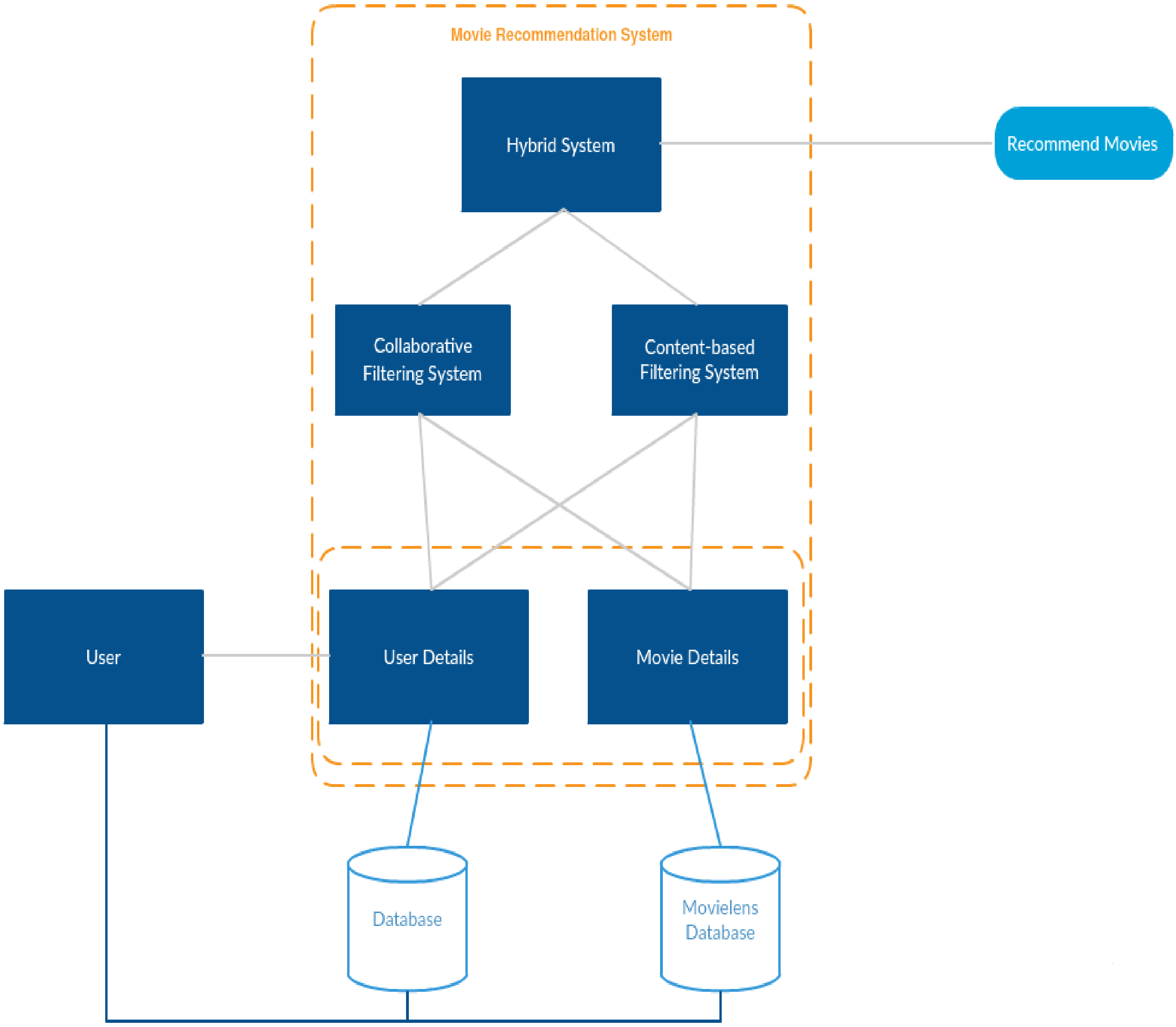
Scope

The project scope encompasses a hybrid recommendation system which will make use of item-based and user-based filtering to provide personalized recommendations. The project will incorporate sentiment analysis based on movie reviews and will also incorporate a success predictor to estimate the success rate of upcoming movies based on various parameters

Chapter 6

Proposed System

New users will have to sign up using the user interface provided on the website. The users will be asked to provide feedback on certain movies and movie genre. Based on the feedback provided, the user will be segregated, and a set of recommendations will be provided. Real time analysis ensures that the system will adopt dynamically based on user behavior. Registered users will be able to access various features such as viewing movie details, add movies to watchlist.


Movie Recommendation System
Hybrid System
Recommend Movies
Collaborative Filtering System
Content-based Filtering System
User
User Details
Movie Details
Database
Movielens Database

Chapter 7

Methodology

We have planned on using the below approaches to solve the given problem:

- Datamining tool (WEKA)
- R-programming language
- Hadoop
- JavaScript
- Apache Spark

Data mining tools like WEKA will be used for performing database operations such as classification, clustering and outlier analysis. Hadoop, R-programming and Apache Spark will be used for developing the back-end and performing real time analysis. The front-end will be developed using JavaScript and HTML/CSS.

Chapter 8

Analysis

8.1 Process model used for the project

Process model used will be the Prototype Model.

A prototype (an early approximation of a final system or product) will be built, tested, and then reworked as necessary based on feedback until an acceptable prototype is finally achieved from which the complete system or product can now be developed.

8.2 Feasibility Study

Technology Considerations

Movie recommendation systems available in the market are dependent on the dataset to contain large of clusters of similar users and items. They also do not provide services such as effective remote access via cloud, customer interaction modules, etc. to be solved with the proposed system.

Product/Service Marketplace

The Movie recommendation system will impact client institutions in several ways. The following provides a high-level explanation of how the organization, tools, processes, and roles and responsibilities will be affected as a result of the movie recommendation system implementation:-

Tools: The existing requirement for on site management systems will be eliminated completely with the availability of a cloud based system.

Processes: With the Movie recommendation system comes more efficient and streamlined administrative and customer relations processes.

Hardware/Software: Clients will need to handle no extra software or hardware apart from a stable high speed Internet connection and a computer device.

Operational Feasibility

The project will be implemented in a way that it will allow the functioning of recommendations smoothly. It will provide a user-friendly user interface in a modular fashion.

Schedule

The recommendation system campaign is expected to take seven months from project approval to launch of the system. following is a high level schedule of some significant milestones for this initiative:

1. July 15, 2017: Initiate Project
2. August 1, 2017: Project kickoff meeting
3. October 31, 2017: Complete recommender system design
4. January 1, 2018: Complete testing of recommender system
5. January 15, 2018: Beta testing trials of recommender system
6. February 1, 2018: Go live with system launch

8.3 Cost analysis

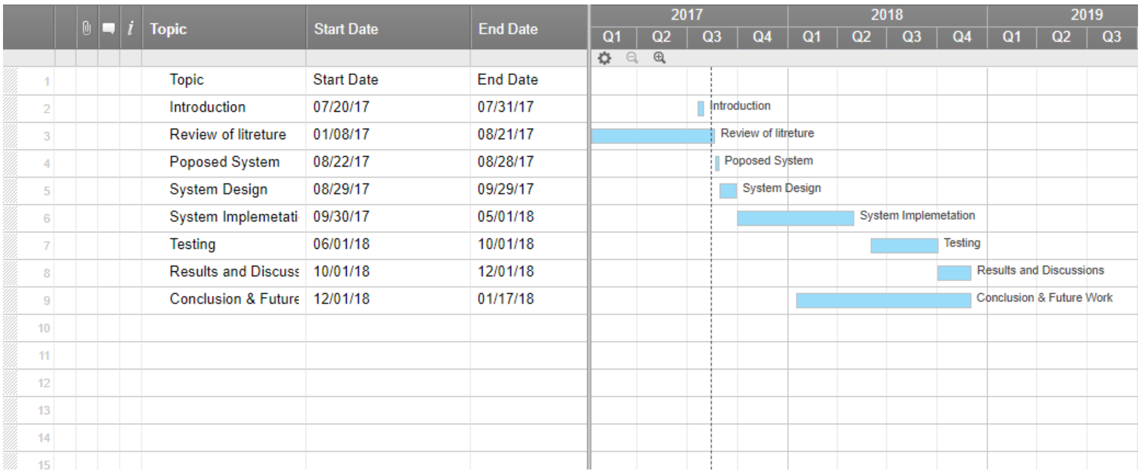
As resources needed are open source, there is no need of resource cost but if this project is used as a business idea then cost associated for developing this project will be as follows:

Assumptions made under COCOCO Model
(According to Semi-Detached Model)

No. of project members: 3
Salary per person assumed as: Rs. 1000/ month
Size of code (In KLOC): 10
Estimation of Development Effort: 26.9 per month
Estimated Development Time: 9 months

Cost required to develop the product: 9 months * 10kLOC * 100rs/hr * 3 project members= 27000

8.4 Timeline Chart



	Topic	Start Date	End Date
1	Topic	Start Date	End Date
2	Introduction	07/20/17	07/31/17
3	Review of litreture	01/08/17	08/21/17
4	Poposed System	08/22/17	08/28/17
5	System Design	08/29/17	09/29/17
6	System Implemetati	09/30/17	05/01/18
7	Testing	06/01/18	10/01/18
8	Results and Discuss	10/01/18	12/01/18
9	Conclusion & Future	12/01/18	01/17/18
10			
11			
12			
13			
14			
15			

Chapter 9

Design

9.1 Data Flow Diagrams

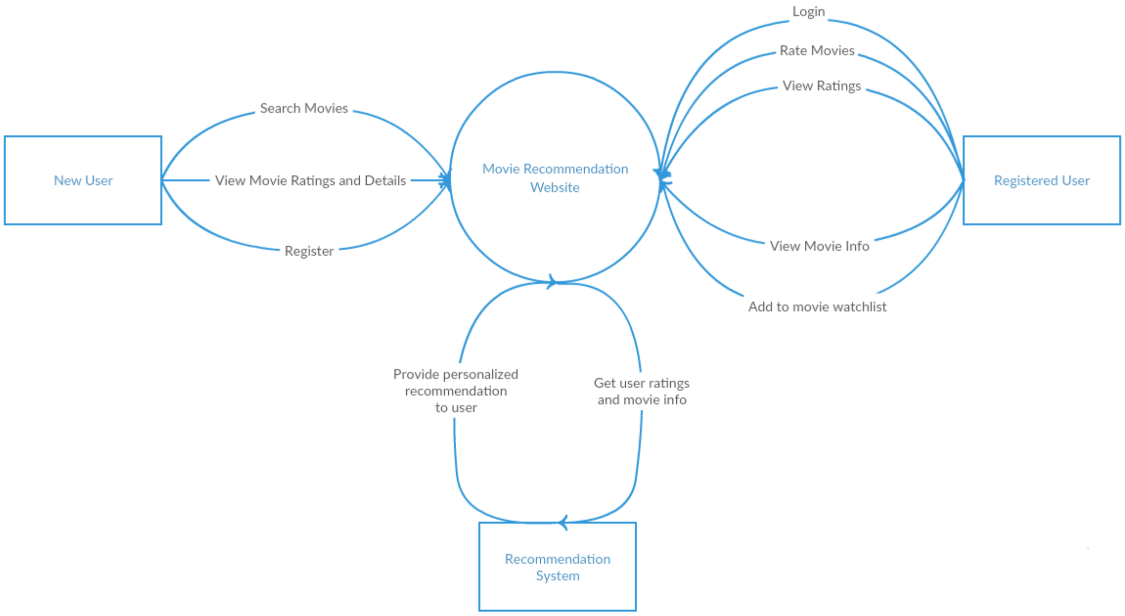

Login
Rate Movies
View Ratings
Search Movies
New User
View Movie Ratings and Details
Movie Recommendation Website
Registered User
Register
View Movie Info
Add to movie watchlist
Provide personalized recommendation to user
Get user ratings and movie info
Recommendation System

Figure: Level 0 Data Flow Diagram

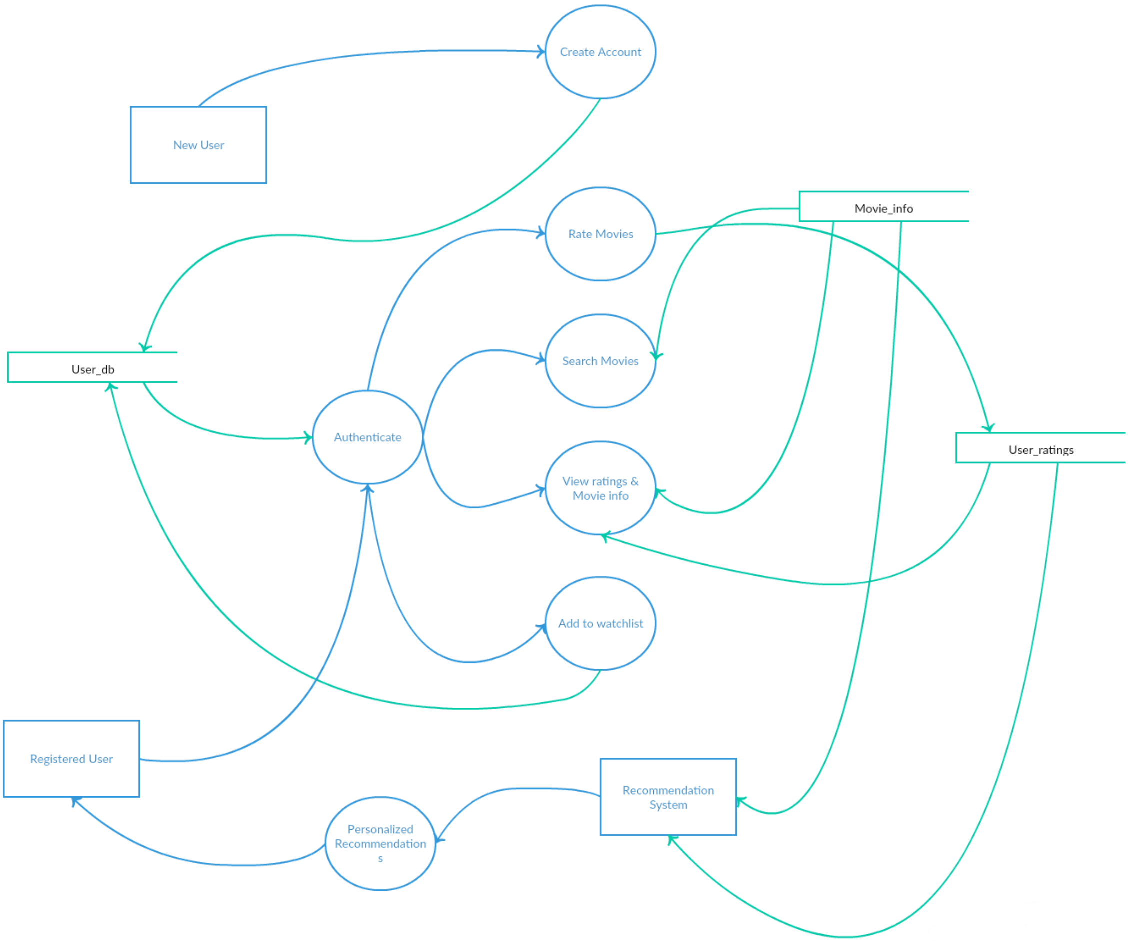


Figure: Level 1 Data Flow Diagram

9.2 UML Diagrams

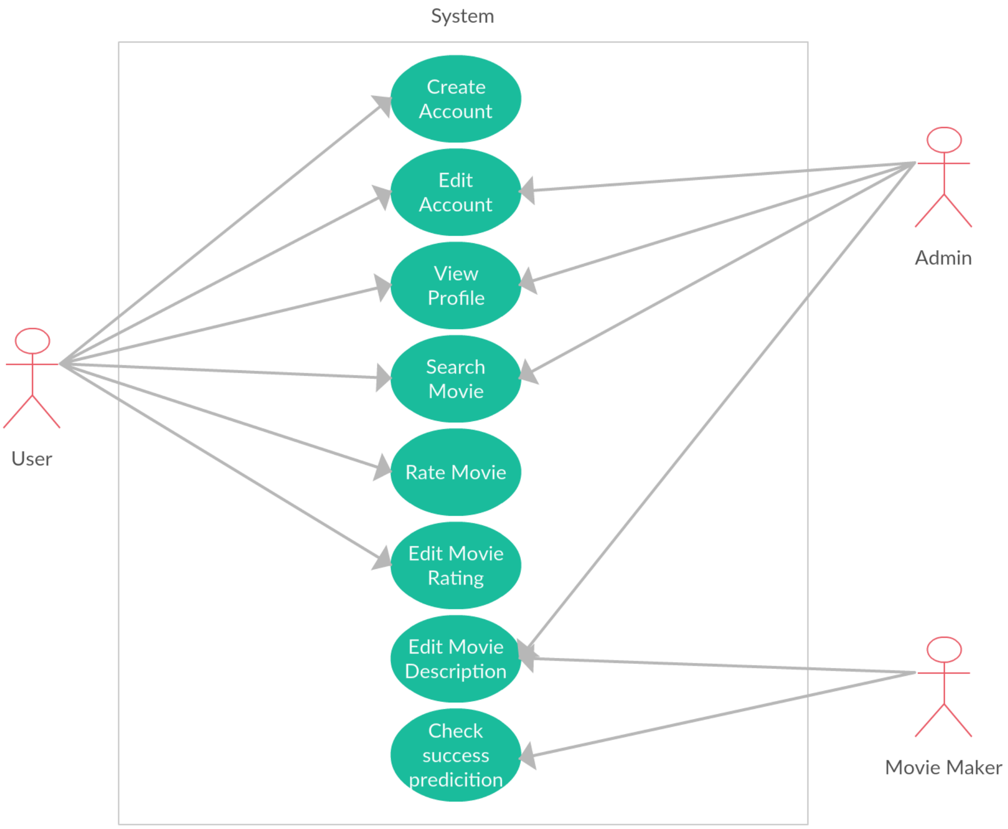


Figure: Use Case Diagram

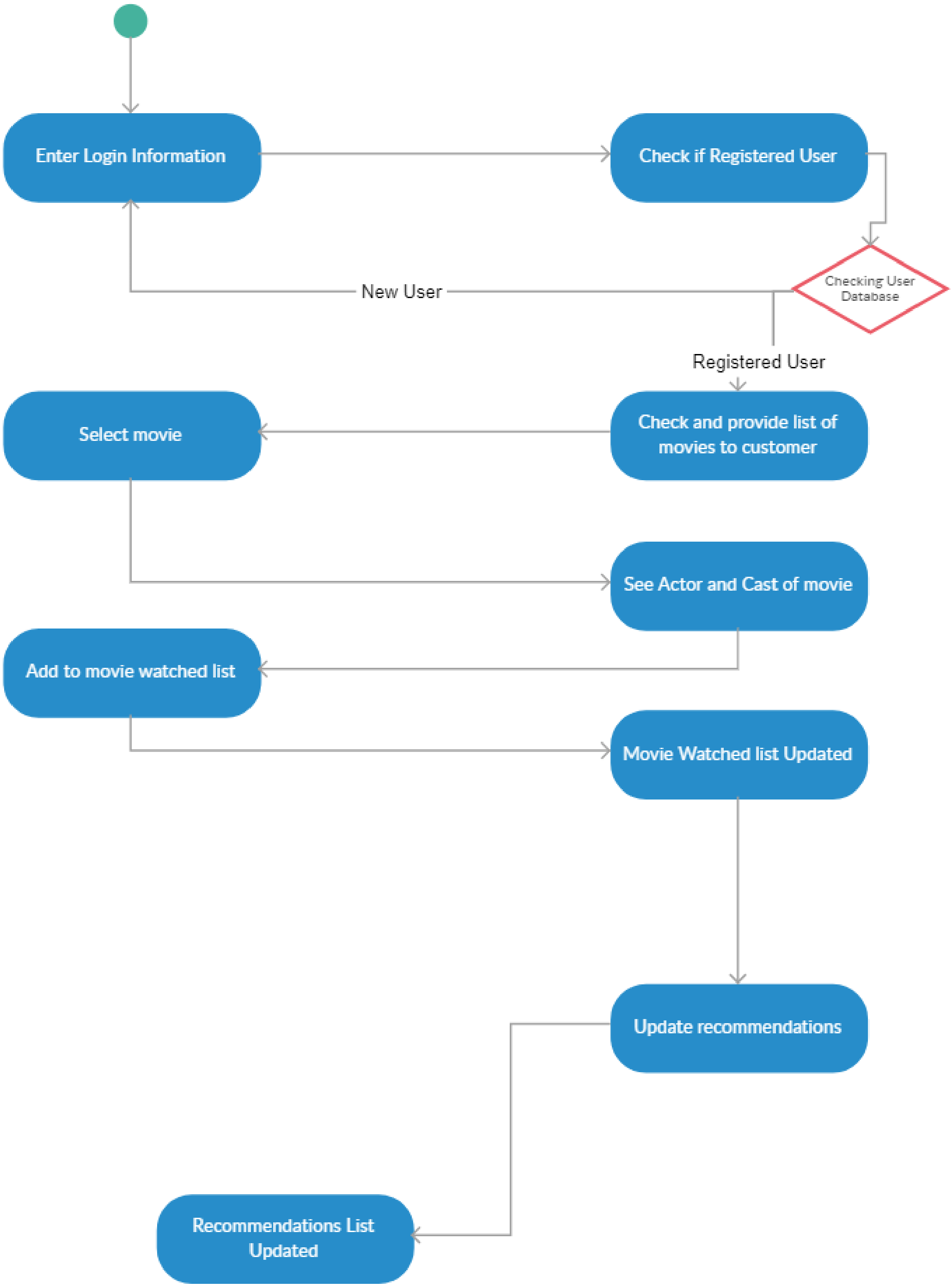
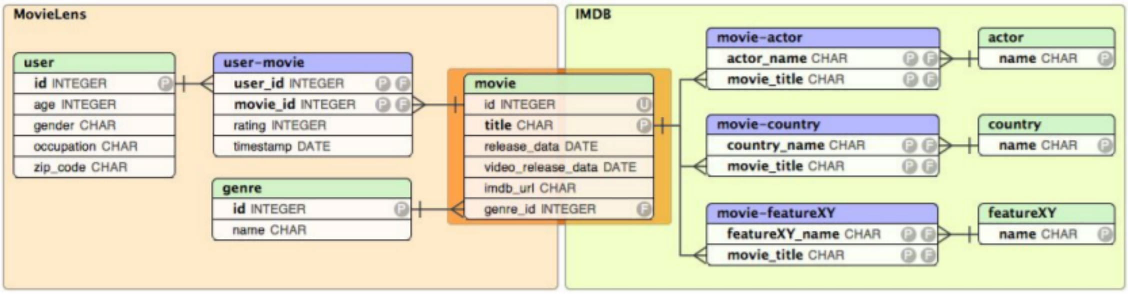

Enter Login Information
Check if Registered User
Checking User Database
New User
Registered User
Select movie
Check and provide list of movies to customer
See Actor and Cast of movie
Add to movie watched list
Movie Watched list Updated
Update recommendations
Recommendations List Updated

Figure: Activity Diagram

9.3 EER Diagram



Chapter 10

Hardware and Software Requirements

1. Hardware Requirements
   - Computer/laptop/smartphone
   - High speed internet connection
   - Server-side functionality (provided by AWS)
2. Software Requirements
   - Apache Spark
   - Hadoop
   - R-programming
   - Web programming languages such as HTML, CSS, JavaScript, etc.

Chapter 11

References

1	**Hydra: A hybrid recommendation system** Stephan Spiegel, Jrme Kunegis, Fang Li, Proc. Workshop on Complex Networks in Information and Knowledge Management, 2009
2	**Improved Neighborhood-Based Algorithms for Large-Scale Recommender Systems** A. Toescher, M. Jahrer, and R. Legenstein, KDD 08: Proceeding of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining, 2008
3	**Unifying User-based and Item-based Collaborative Filtering Approaches by Similarity Fusion** J. Wang, A. P. de Vries, and M. J. T. Reinders, SIGIR 06: Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval, 2006
4	**Matchbox: Large scale online bayesian recommendations** D. H. Stern, R. Herbrich, and T. Graepel, WWW 09: Proceedings of the 18th international conference on World wide web, 2009

5	**Factorization meets neighborhood- a multifaceted collaborative filtering model** Y. Koren, n KDD 08: Proceeding of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining, 2008
6	**Item-based collaborative filtering recommendation algorithms** B. Sarwar, G. Karypis, J. Konstan, and J. Reidl., WWW01: Proceedings of the 10th international conference on World Wide Web, 2001
7	**Toward the next generation of recommender systems: A survey of the state-of-the-art and possible extensions** Adomavicius and Tuzhilin, IEEETKDE, 2005
8	**Amazon.com recommendations: Item-to-item collaborative filtering** G. Linden, B. Smith, and J. York, IEEE Internet Computing, 2003
9	**Application of dimensionality reduction in recommender systems** B. M. Sarwar, G. Karypis, J. A. Konstan, and J. T. Riedl, ACM WebKDD Workshop, 2000
10	**Hybrid web recommender systems** R.D. Burke, The Adaptive Web, Methods and Strategies of Web Personalization, 2007