“Recommendation Engine for Movies Portal”

7th Semester Mini Project Report

Under The Guidance of
Prof. Dr. Ranjana Vyas
IIT-Allahabad

By-
Shweta Choudhary (RIT2012025)
Nakshatra Maheshwari(RIT2012074)
ShashankSharma (RIT2012075)

INDIANINSTITUTE OFINFORMATIONTECHNOLOGY ALLAHABAD
(AUNIVERSITY ESTABLISHED UNDERSEC.3OFUGCACT, 1956VIDENOTIFICATION NO.F.9-4/99-U.3DATED 04.08.2000 OF THEGOVT. OF INDIA)
ACENTREOFEXCELLENCEININFORMATION TECHNOLOGY ESTABLISHED BY GOVT. OF INDIA

November,2015

CANDIDATES'DECLARATION

We hereby declare that the work presented in this project report entitled "Recommendation Engine for Movie Portal", submitted towards completion of 7th Mid-Semester report of B.Tech. (IT) at Indian Institute of Information Technology, Allahabad, and is an authenticate record of our original work performed from July 2015 to Dec 2015 under the supervision of Dr. Ranjana Vyas. Due acknowledgements have been made in the text to all other material used. The project will be performed in full compliance to meet the stated objective with the stated Requirements and coping up with the constraints of the prescribed curriculum.

Date:
Place:Allahabad

Shweta Choudhary (RIT2012025)
Nakshatra Maheshwari (RIT2012074)
Shashank Sharma (RIT2012075)

CERTIFICATE

It is to state that the statement above made by the group is correct and appropriate to best of my knowledge .This work taken up as 7th semester mini-project by the above stated group is being performed under my supervision and guidance.

Date:
Place: Allahabad

Dr. Ranjana Vyas
IIIT-Allahabad

Contents

1. Abstract

Which movie I should watch tonight? Ever wonder how difficult this question could be if someone asks you to recommend the best movie? Best in terms of what? How this best is defined?

MOVIES fascinate every individual despite of his religion, age, zone, sex, and taste. Due to infinitely large number of movie options available for a user of specific taste, it is of utmost importance to develop a system that could recommend best movie (suited most to user preference). Many such systems have already been developed. But some of them recommend movies of a certain genre while some of them could not recommend to cold-start user. In this project, we aim to develop a system which is best in itself in terms of recommending as it is taking average value from available sources as well as user present needs and taste. Not only this but also, by taking average value, the cases of fraud ratings of already rated sites become lower in our recommendation, making it more reliable than present existing ones.

2. Introduction

In this dynamic world, where too much of information is available which keeps changing and increasing every minute, choosing domain of available options is indeed very large. Due to abundance of data floating everywhere, choosing one that is best suited to meet our demands and preferences is most times a very tedious job. Not only is this but it time consuming also. Therefore if we could rely on a system who can do this task for us and that too with accepted amount of accuracy, is definitely be a boon not only for problem solving in general, but also for specific individual in particular. Recommender system is the key.

Recommender systems are an important part of the information and e-commerce ecosystem. Whenever we talk of data which is exponentially large in comparison to a relatively small subset of options that are of interest, recommender system have always proved its existence. Recommender system takes input from user (say preferences, taste, necessity etc.), apply some apt algorithm on it and produce the recommender output for distinct user. Our recommendation engine will not only be based on available ratings but also on the distinct taste of individual user.

- Taking of various options available, the first question that arises is why movies only?
  The answer to this question lies on the fact that every individual despite of age, zone, sex, taste needs ENTERTAINMENT.

Why movie recommendation is the need of the hour?

- Preferably very large number of choices available.
- Drastically reduces the domain of choices.
- Movies guarantee entertainment (attracts user of all age groups).
- How our system is superior to the ones already existing?
- **“MAKE IN INDIA” approach.**
- More reliable dataset and in turn more accurate outcome.
- Can deal with fake ratings.

3. Problem Definition and Objectives

Why to depend on foreign movies dataset to produce movie Recommendation system algorithms?

- We will have complete Indian movies dataset where user can login and rate movies of their choices.
- We will also use ratings provided by different websites to produce collaborative rating for the recommendation-minimal false recommendation.
- Using the above information and applying algorithms for predicting rating, top movies will be recommended to the users.

Objectives:-

- Build dataset for Movie Recommender System.
- Choice of Algorithm for predicting ratings :-
  - Cosine similarity method to find similar users.
  - Matrix factorization techniques.

4. Literature Survey

Recommender systems emerged as an independent research area in the mid-1990s when researchers started focusing on recommendation problems that explicitly rely on the ratings structure.

Collaborative Filtering is the best technique for proposing individual recommendations to a user considering all parameters including his taste, zone, preferences, dislikes .Individual user should always feel as if the system is defined only for him giving him the most of what he desires.

4.1 Collaborative Filtering

The collaborative filtering algo works on the basis of user with similar interest. The main idea around which this theory is centered is to match people having similar preferences. It Recommends depending on neighbors. Neighbors –user with most similar taste to the user considered at that very moment.

4.2 Similarity Measures [2]

Similarity measure reflects the degree of closeness or separation of the target objects and should correspond to the characteristics that are believed to distinguish the object/group of object where the target object belongs.

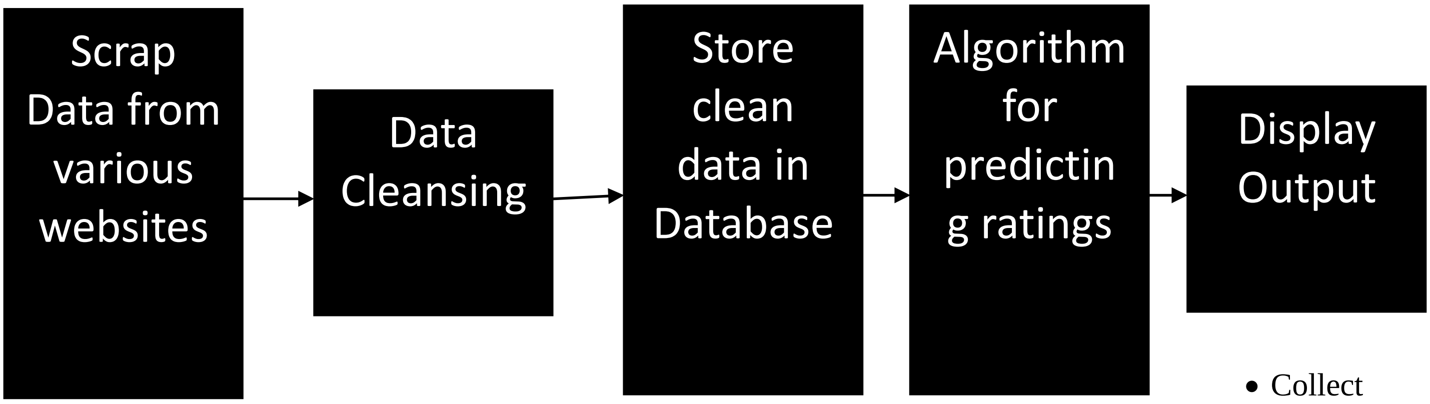
4.2.1Euclidean Distance:-Euclidean distance is a standard metric for geometrical problems.

4.2.3 Cosine Similarity:-Cosine similarity is a measurement of similarity between the two vectors of an inner product space that measure the cos of the angle between them.

$$SIM_C(\overrightarrow{t_a}, \overrightarrow{t_b}) = \frac{\overrightarrow{t_a} \cdot \overrightarrow{t_b}}{|\overrightarrow{t_a}| \times |\overrightarrow{t_b}|},$$

5. Proposed Methodology And Complete Overview

Overview →


Scrap Data from various websites
Data Cleansing
Store clean data in Database
Algorithm for predictin g ratings
Display Output

- Collect data about movies (Year, Genre, Artist, etc.). Wikipedia is the best source to collect Indian movies data.
- Scrap the movies rating from various rating websites, like, **Timesofindia.com, Imdb.com, Bollywoodhungama.com, Ndtv.com, Glamsham.com, Rottentomatoes.com, etc.**
- Apply Data Cleansing to remove noise and clean Data.(Using Heuristic Levenshtein distance)
- Normalize the Schema for storing the above clean data in database.
- Algorithm for predicting ratings.
- Output->Display the recommended movies to the users.

5.2APPROACH: Data Cleaning

- We use different approaches for this (and for most efficiency we mix them all):
- Ignore everything that is in parenthesis.
- Breaking down into cases based on Artist, Genre, etc.
- Define words automatically drop like “movie”, “new”, etc.
- Make a dictionary of feature set (Genre, Year of release, etc.) for different catalog which should be same while declaring any movie as common movie.
- Compare the names via their Levenshtein distance [5] and use this distance for clustering.
- Heuristic Levenshtein distance [5].
- If there are n movies, then overall complexity= ($O(n^2l^2$), where l=product title length

5.3Heuristic Levenshtein distance

By considering Levenshtein distance between dictionary movie title and the movie title we scraped, all 3 operations are considered:

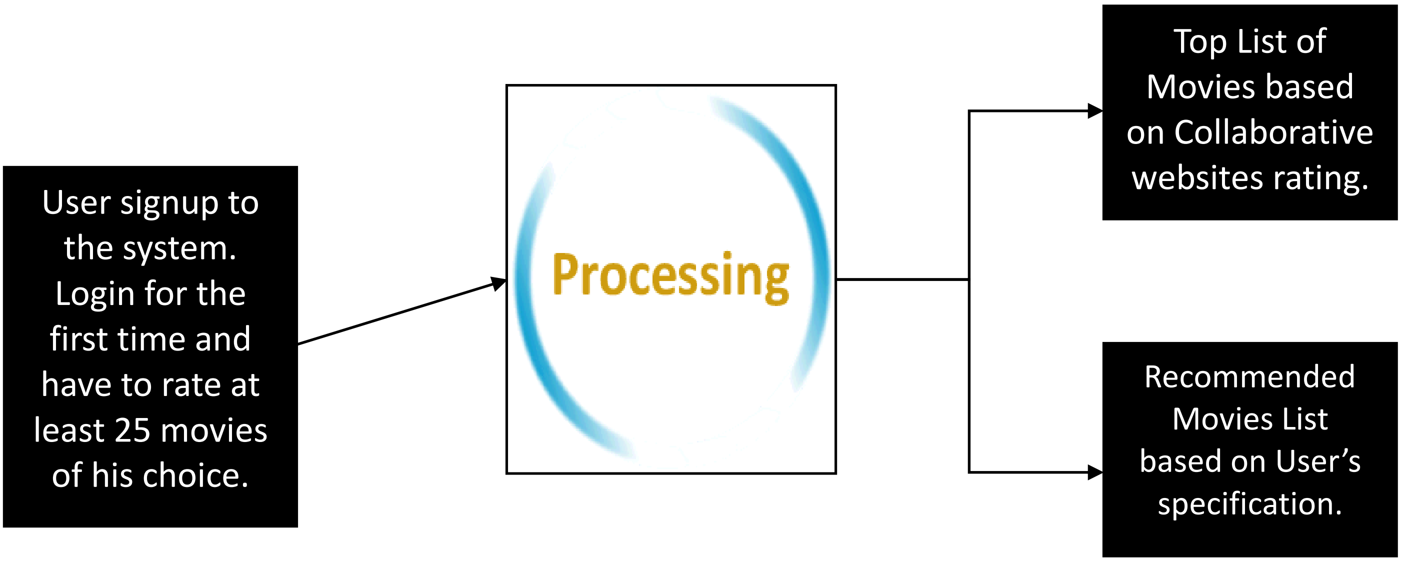
- DELETE OPERATION
- SUBSTITUTION OPERATION
- INSERT OPERATION

The optimal solution count consists of all the counts of respective operations.
If we consider only SUBSTITUTION OPERATION count from the optimal solution, then count comes nearly equal to 0.
Then we predict them similar.

5.4Algorithms

We will use either of the two methods:

- **5.4.1 Collaborative Filtering using Similarity Measures like Cosine Similarity:**
    - In this method, movies are recommended by finding out the similarity between the users preferences using cosine distance similarity [6] and demographic information [6] of users.
    - After that we will apply several heuristics to produce final result.
    - Such methods have justification of the results, but its performance decreases when matrix data gets scattered.
- **5.4.2 Matrix Factorization for Collaborative Filtering:**
    - In this method, we will use matrix-factorization method of collaborative Filtering for the rate prediction and ranking using SVDFeature [7].
    - SVDFeature [7] is a machine learning toolkit for feature-based collaborative filtering.
    - The feature-based setting allows us to build factorization models.
    - SVDFeature [7] will learn a feature-based matrix factorization model with the given training data and make predictions on supplied test feature files.


User signup to the system. Login for the first time and have to rate at least 25 movies of his choice.
Processing
Top List of Movies based on Collaborative websites rating.
Recommended Movies List based on User's specification.

6. HARDWARE &SOFTWARE REQUIREMENTS

Hardware tools –

- Server to handle multiple User requests.

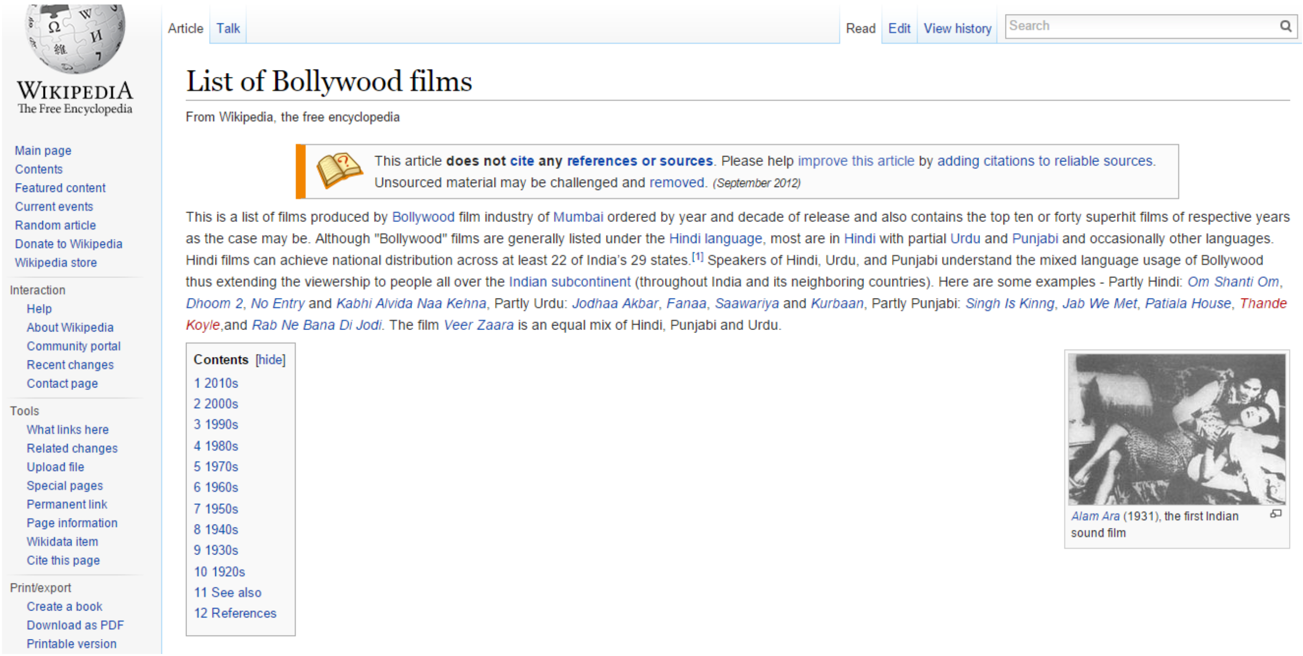
Software tools –

- Python based IDE (Canopy).
- C++ Based IDE (Code blocks, DevC++).
- JAVA Based IDE (Netbeans).
- LAMP (Linux, Apache, MYSQL, PHP) server.
- SVDFeature toolkit.

7 .Work done till mid semester

MONTH	SUMMARY
JULY	Study on various recommendation system available.
AUGUST	Analyze different recommendation algorithm. Learnt Scrappy tools to extract data. Find places where we can find the dataset.

Complete List of Bollywood movies available in Wikipedia (Year wise):



List of Bollywood films

From Wikipedia, the free encyclopedia

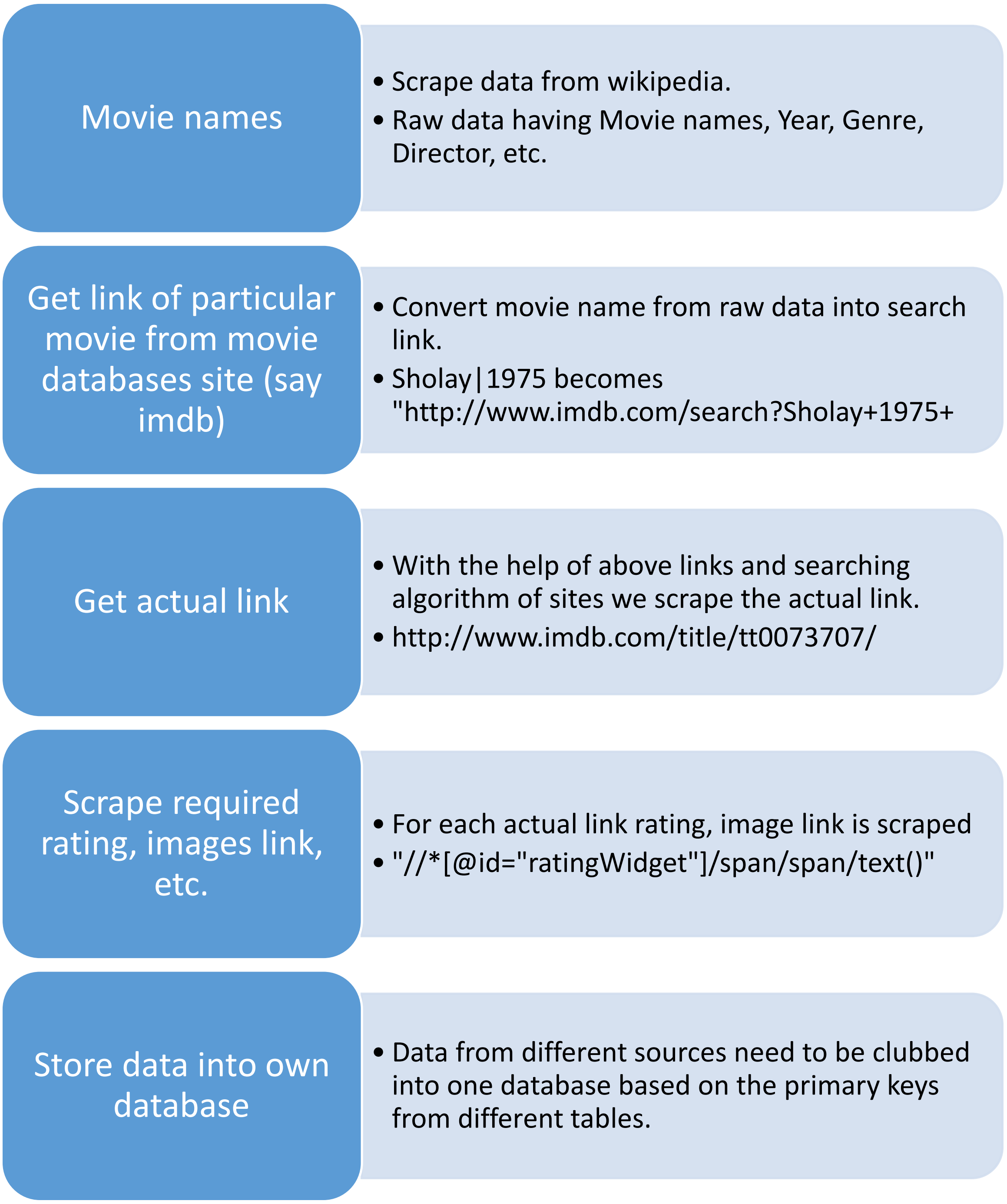
This article **does not cite any references or sources**. Please help improve this article by adding citations to reliable sources. Unsourced material may be challenged and removed. *(September 2012)*

This is a list of films produced by Bollywood film industry of Mumbai ordered by year and decade of release and also contains the top ten or forty superhit films of respective years as the case may be. Although "Bollywood" films are generally listed under the Hindi language, most are in Hindi with partial Urdu and Punjabi and occasionally other languages. Hindi films can achieve national distribution across at least 22 of India's 29 states.[1] Speakers of Hindi, Urdu, and Punjabi understand the mixed language usage of Bollywood thus extending the viewership to people all over the Indian subcontinent (throughout India and its neighboring countries). Here are some examples - Partly Hindi: *Om Shanti Om*, *Dhoom 2*, *No Entry* and *Kabhi Alvida Naa Kehna*, Partly Urdu: *Jodhaa Akbar*, *Fanaa*, *Saawariya* and *Kurbaan*, Partly Punjabi: *Singh Is Kinng*, *Jab We Met*, *Patiala House*, *Thande Koyle*,and *Rab Ne Bana Di Jodi*. The film *Veer Zaara* is an equal mix of Hindi, Punjabi and Urdu.

Contents [hide]

Alam Ara (1931), the first Indian sound film

8.AFTER MID-SEMESTER



Movie names

- Scrape data from wikipedia.
- Raw data having Movie names, Year, Genre, Director, etc.

Get link of particular movie from movie databases site (say imdb)

- Convert movie name from raw data into search link.
- Sholay|1975 becomes "http://www.imdb.com/search?Sholay+1975+

Get actual link

- With the help of above links and searching algorithm of sites we scrape the actual link.
- http://www.imdb.com/title/tt0073707/

Scrape required rating, images link, etc.

- For each actual link rating, image link is scraped
- "//*[@id="ratingWidget"]/span/span/text()"

Store data into own database

- Data from different sources need to be clubbed into one database based on the primary keys from different tables.

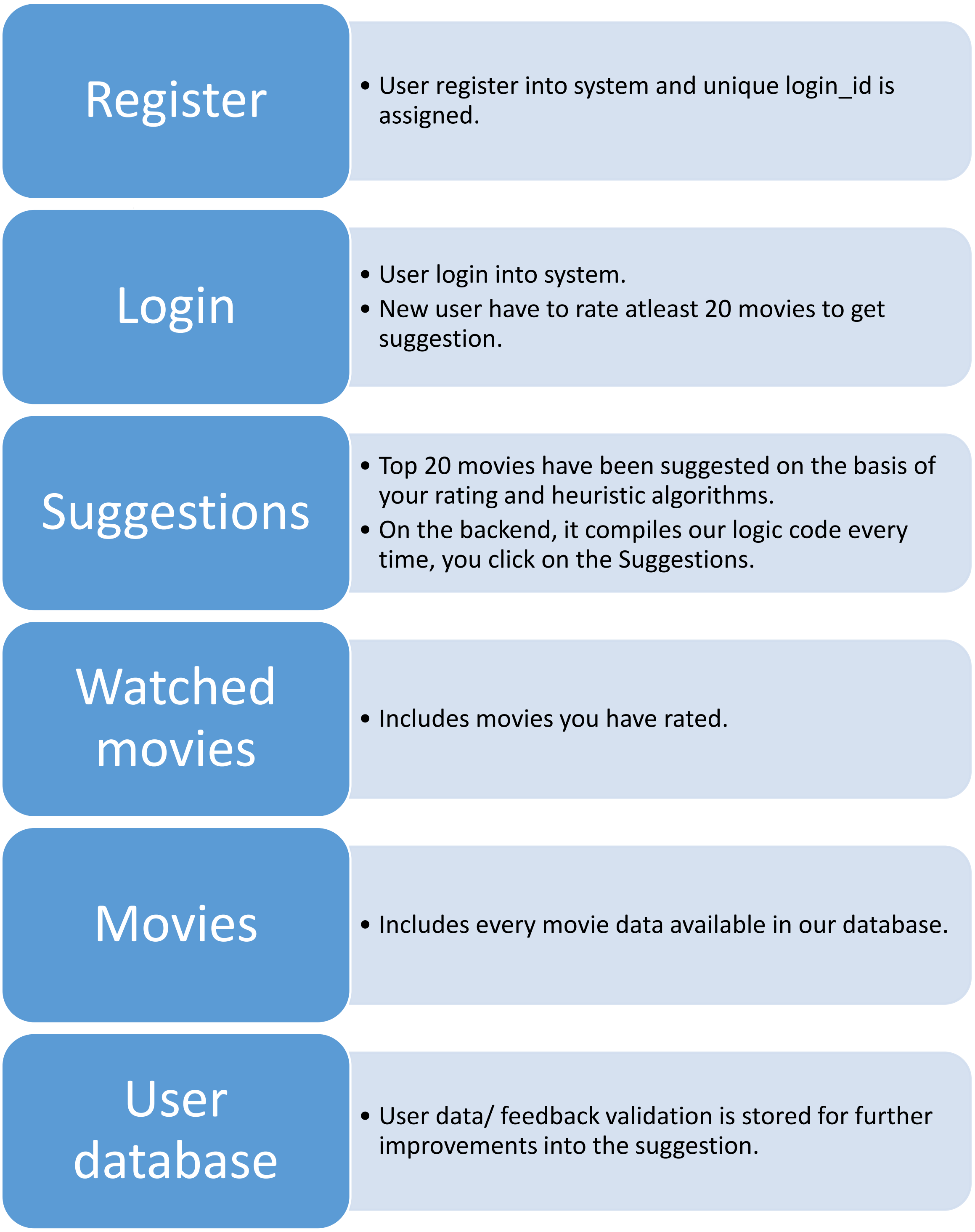
movie_id	movie_name	movie_director	Movie_genre	Movie_imdb_rawlink

Movie_imdb_rawlink	Movie_imdb_actuallink

Movie_imdb_actuallink	Movie_rating	Movie_image_link

Different databases are joined using MySQL queries:

SELECT list.*, imdbrating.*, rotten.rottenlink **FROM** imdbrating, list, imdb, rotten **WHERE LTRIM**(**RTRIM**(list. Movie_imdb_rawlink))=**LTRIM**(**RTRIM**(imdb. Movie_imdb_rawlink)) && LTRIM(RTRIM(imdb.Movie_imdb_actuallink))=LTRIM(RTRIM(imdbrating. Movie_imdb_actuallink)) && **LTRIM**(RTRIM(list.id))=**LTRIM**(RTRIM(rotten.id));



Register

- User register into system and unique login_id is assigned.

Login

- User login into system.
- New user have to rate atleast 20 movies to get suggestion.

Suggestions

- Top 20 movies have been suggested on the basis of your rating and heuristic algorithms.
- On the backend, it compiles our logic code every time, you click on the Suggestions.

Watched movies

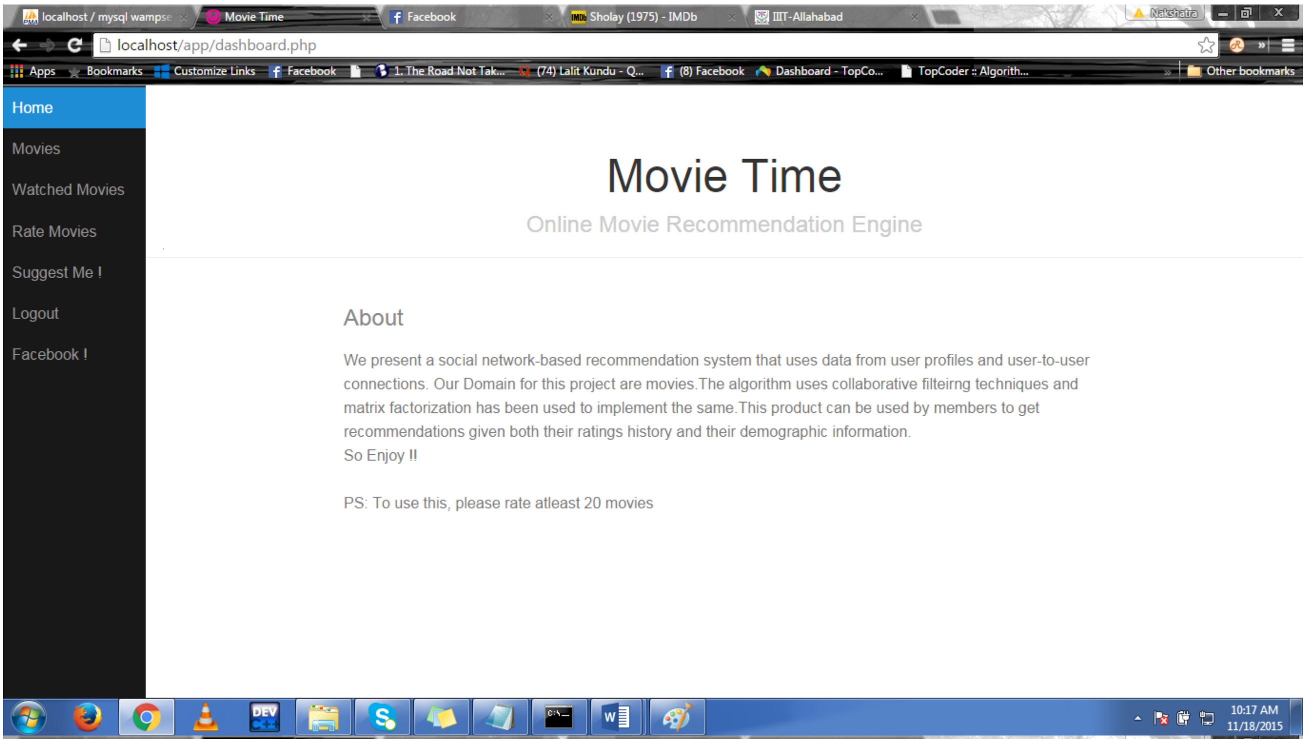
- Includes movies you have rated.

Movies

- Includes every movie data available in our database.

User database

- User data/ feedback validation is stored for further improvements into the suggestion.



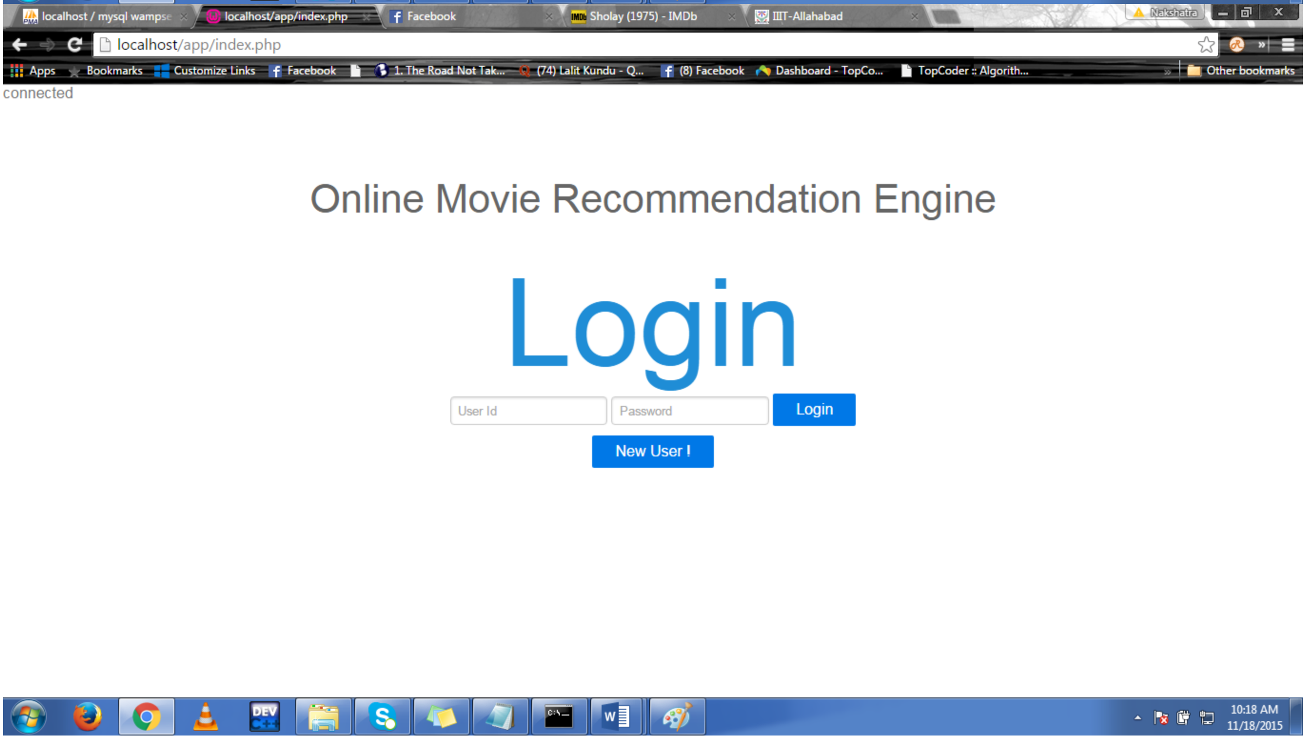
Movie Time

Online Movie Recommendation Engine

About

We present a social network-based recommendation system that uses data from user profiles and user-to-user connections. Our Domain for this project are movies.The algorithm uses collaborative filteirng techniques and matrix factorization has been used to implement the same.This product can be used by members to get recommendations given both their ratings history and their demographic information.
So Enjoy !!

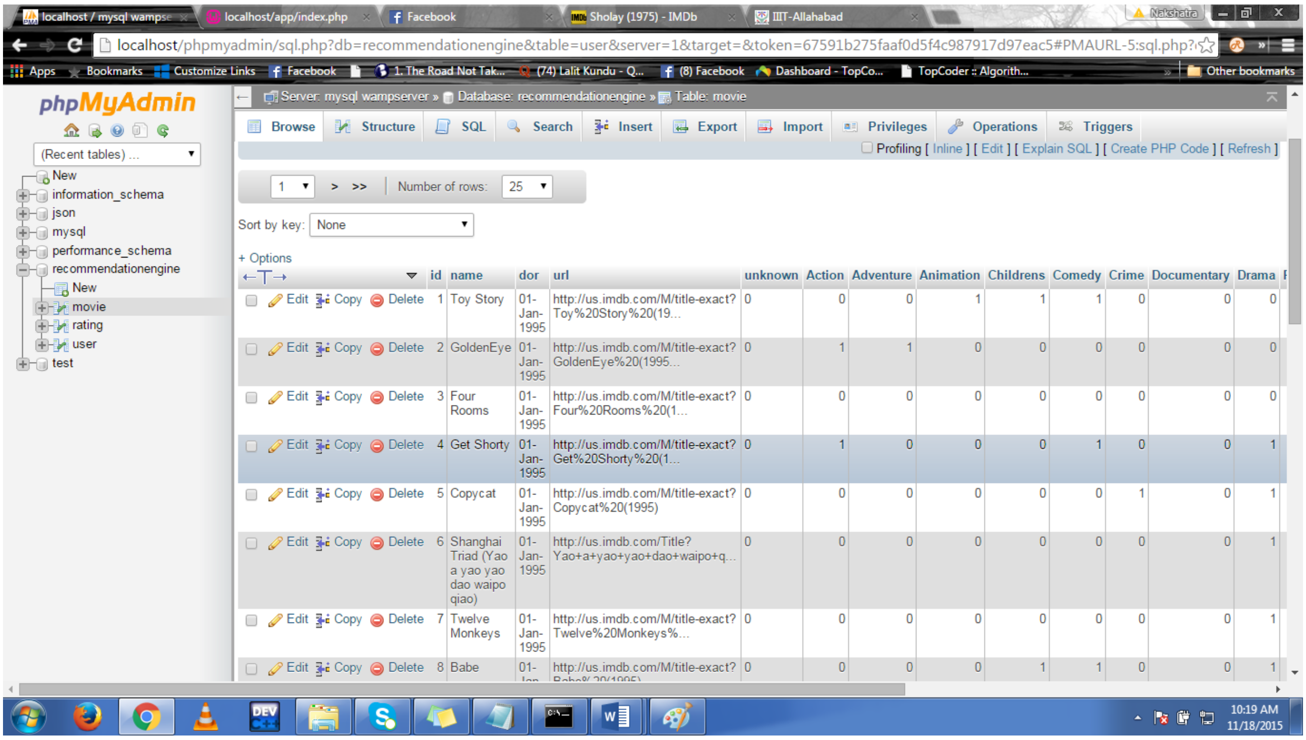
PS: To use this, please rate atleast 20 movies



connected

Online Movie Recommendation Engine

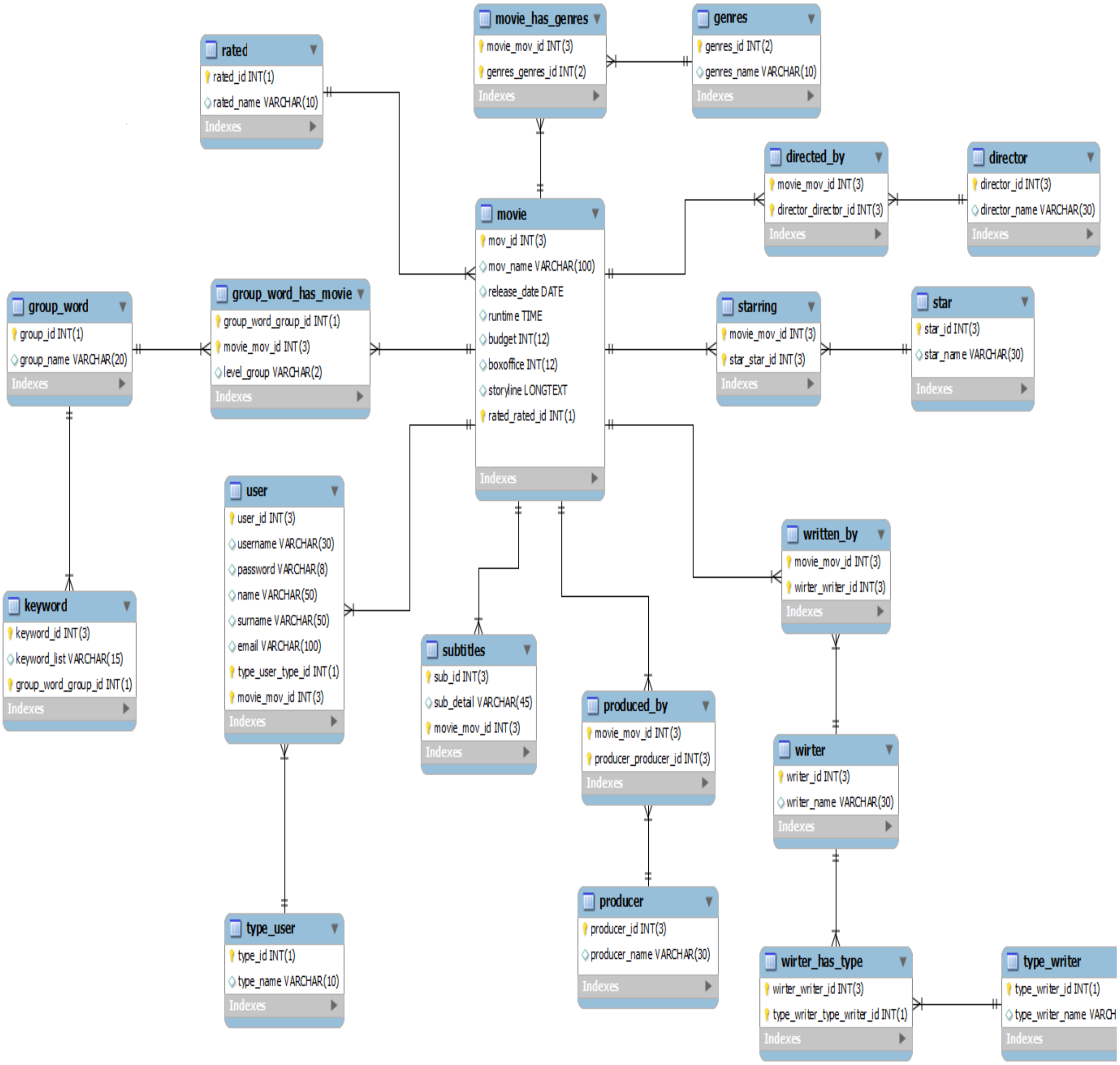
Login



Server: mysql wampserver » Database: recommendationengine » Table: movie

	id	name	dor	url	unknown	Action	Adventure	Animation	Childrens	Comedy	Crime	Documentary	Drama
□ Edit Copy Delete	1	Toy Story	01-Jan-1995	http://us.imdb.com/M/title-exact?Toy%20Story%20(19...	0	0	0	1	1	1	0	0	0
□ Edit Copy Delete	2	GoldenEye	01-Jan-1995	http://us.imdb.com/M/title-exact?GoldenEye%20(1995...	0	1	1	0	0	0	0	0	0
□ Edit Copy Delete	3	Four Rooms	01-Jan-1995	http://us.imdb.com/M/title-exact?Four%20Rooms%20(1...	0	0	0	0	0	0	0	0	0
□ Edit Copy Delete	4	Get Shorty	01-Jan-1995	http://us.imdb.com/M/title-exact?Get%20Shorty%20(1...	0	1	0	0	0	1	0	0	1
□ Edit Copy Delete	5	Copycat	01-Jan-1995	http://us.imdb.com/M/title-exact?Copycat%20(1995)	0	0	0	0	0	0	1	0	1
□ Edit Copy Delete	6	Shanghai Triad (Yao a yao yao dao waipo qiao)	01-Jan-1995	http://us.imdb.com/Title?Yao+a+yao+yao+dao+waipo+q...	0	0	0	0	0	0	0	0	1
□ Edit Copy Delete	7	Twelve Monkeys	01-Jan-1995	http://us.imdb.com/M/title-exact?Twelve%20Monkeys%...	0	0	0	0	0	0	0	0	1
□ Edit Copy Delete	8	Babe	01-	http://us.imdb.com/M/title-exact?	0	0	0	0	1	1	0	0	1

DATABASE CONTEXTS



9. RESULT

We present a social network-based recommendation system that uses data from user profiles and user-to-user connections. The algorithm uses collaborative filtering techniques and matrix factorization has been used to implement the same. This product can be used by members to get recommendations given both their ratings history and their demographic information.

10.REFERNCES

[1] (http://en.wikipedia.org/wiki/Levenshtein_distance)
[2] Anna Huang (2008) ,"Similarity Measures for Text Document Clustering" , Department of Computer Science The University of Waikato, Hamilton, New Zealand.
[3] Tianqui Chen. " SVDFeature : A Toolkit for Feature-based Collaborative Filtering," Shanghai Jiao Tong University.
[4] http://www.cnet.com/news/top-10-movie-recommendation-engines/
[5] http://maheshakya.github.io/gsoc/2014/05/18/preparing-a-bench-marking-data-set-using-singula-value-decomposition-on-movielens-data.html
[6] Gilda MoradiDakhel,"A New Collaborative Filtering Algorithm Using K-means Clustering and Neighbors' Voting" Faculty of Electrical, Computer & IT Engineering, Azad University, Qazvin Branch, Qazvin, Iran

REMARKS