# Knowledge Discovery in the Movie Database

Yuzhou Wang
School of Electrical and Computer Engineering
University of Waterloo
Waterloo, ON, Cananda
Email: ...@uwaterloo.ca

Sainan He
School of Electrical and Computer Engineering
University of Waterloo
Waterloo, ON, Cananda
Email: s66he@uwaterloo.ca

***Abstract*—The abundance of movie data in terms of review, rating or even detail information in the internet has encouraged many researches to formulate techniques to analyze the pattern in movie data invovlving in discovering factors which will influence the success of movies and developping recommendation systems of movie according to user reviews. With these techniques, poeple are able to analyse links between attributes in real data sets so that salesperson can boost their sales by puting associated or similar products together, or by recommending products the customers will most likely be interested in. In this paper, we apply different data mining techniques including association rules, clutering, classification and prediction on two Internet movie datasets.**

## I. Introduction

Nowadays, we live in a world with vast amounts of data collected everyday. Analyzing such data is an urgent need, thus, data mining has become a popular topic and a fast-growing field. Data mining techniques are already widely deployed in many essential area such as business, society, science, medicine, engineering and almost everywhere[1].

Data mining involves different knowledge discovery such as classification, clustering, association analysis[2]. Association rule learning is a method for discovering interesting relations between variables in large databases. Clustering is an unsupervised data mining technique for discovering interesting patterns from a given database. Classification is a supervised approach classifying records in a data set into predefined classes or even defining classes on the go.

It is very likely that a user will give a movie similar rating with another user who has the same taste. Such an approach that making predictions based on the interests of a user by collecting preference or tastes information from other users is Collaborative Filtering(CF) which is the most popular method in recommendation systems.

The objective of this paper is firstly, to provide a suitable approach along with necessary factors that are to be considered for association rules, clustering and classification using Internet Movie Database (IMDB) data. Performing different classification, a comparison is based on the evaluation the results. Lastly, apply collaborative filtering method to predict users' rating to movies using MovieLens dataset.

The organisation of the paper is as follows: Section 2 provides the literature review about the problem domain. Section 3 and 4 introduces the two dataset that we used in this paper and the data processing approaches. Section 5 gives an overview of the techniques we use to perform our analysis. Section 6 describes the actual analysis performed, and then presents the results and a discussion thereof. Section 7 gives the conclusions reached and a note about possible further work.

## II. Literature Review

### *A. Data mining for the internet of things: literature review and challenges[3]*

This is a review article surveying data mining through 3 perspectives. From the knowledge view, it illustrated classification, clustering, association analysis, time series analysis, outlier analysis and related realization such as SVM, Bayesian networks, SVD, etc from technique view. Then it introduced application of these theory and techniques in different fields. It also discussed challenges for data mining, like extraction useful data from large quantity of data with low quality. Finally, based on the former analysis on data mining algorithms and application, the author proposed a system architecture for big data mining system.

### *B. A Case Study in a Recommender System Based on Purchase Data[4]*

This paper present three kinds of collaborative filtering algorithms- memory based approach, matrix factorization and bigram matrix method- on a real-world dataset for recommending items to customer according customers purchase histories. The author established models for the three methods and applied them on different settings, comparing their results. They also proposed the multidimensional model for contextual analysis, but they didnt talk about it in detail due to the space limit. The research mainly drew the conclusion that: (1) the algorithm based on bigram association rules obtained the best performances; (2) the performance of these algorithms has slight difference compared to that of introducing contextual analysis.

### *C. Algorithms and Methods in Recommender Systems[5]*

In this report, the researcher have described traditional and modern recommender approaches with giving concrete examples and presenting their problems. The traditional ones which work with profiles of users, Content-based filtering measuring similarities and Collaborative filtering building neighbourhood, are usually combined to avoid some limitations

and problems for better results. Some modern methods are an extension of collaborative filtering, such as Context-aware, Semantic-based and Cross-domain based approaches, which outperform original one. Nevertheless, obtaining context information and creating new text mining techniques are some of the problems remaining to be solved. Others like Peer-to-Peer and Cross-lingual approaches are briefly introduced by the researcher.

## III. Data Collection

In this paper, we will implement association rules, clustering, classification and rating prediction. We choose two widely used datasets about moives- IMDB and Movielens. IMDB dataset is used for the former three data mining methods and MovieLens is for rating predicting.

### A. IMDB

The Internet Movie Database (IMDb) is a comprehensive online database having information about movies, actors, television shows, production, etc. The IMDb web site[6] provides more than 50 text files in ad-hoc format (called lists) containing different characteristics about movies (e.g. actors.list or running-times.list). Given the large scale of the data and the degree of interactions between the people, IMDb is a fertile source of data mining problems.

### B. MovieLens

MovieLens is a movie recommender project, developed by the Department of Computer Science and Engineering at the University of Minnesota. MovieLens is a typical collaborative filtering system that collects movie preferences from users and then groups users with similar tastes. Based on the movie ratings expressed by all the users in a group it attempts to predict for each individual their opinion on movies they have not yet seen. Two data sets are available at the MovieLens web site[7]. In this paper, we will use the MovieLens 100K dataset which consists of 100,000 ratings for 1682 movies by 943 users.

## IV. Data Preparation

### A. Reconstruct Data in MySQL Database

At first we tried to store all the data in IMDB into MySQL. As mentioned above, the original dataset in IMDB is in text files which is a list format within natural language.This provided the insight that raw IMDB data are unsuitable for data mining unless they are processed through some natural language processing tool. To minimize the effort spent on parsing all the text files and then converting each one to a table in a database, a third party tool name IMDBPY as obtained from[8] is used. It is an alternative way to navigate through movie information. IMDBPY automatically imports the list files and creates a MySQL database with tables and populates the tables with required data.

The dataset contains abundant information about movies which makes it a very large amount of records. It takes more than 3 hours to convert all the files and import the data into

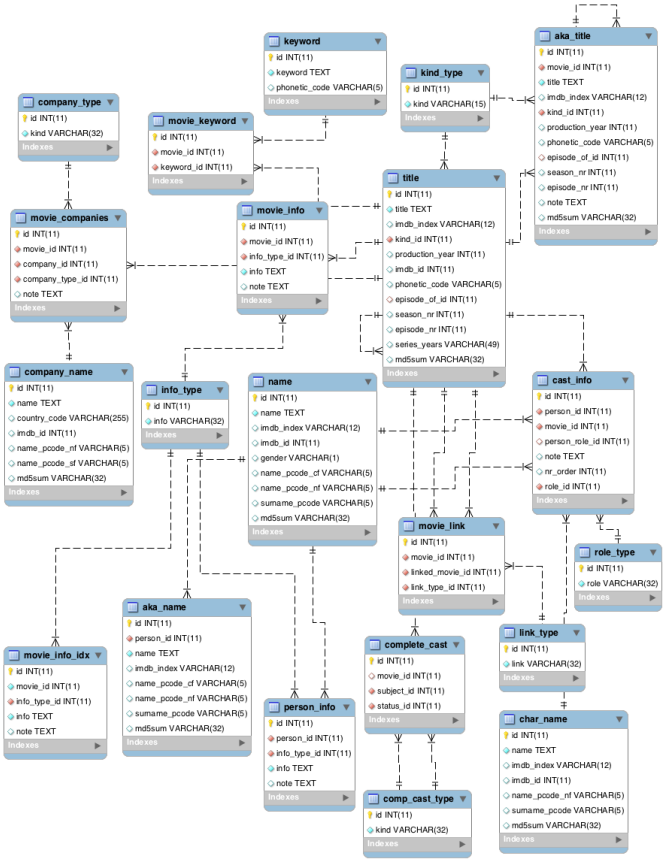


Fig. 1. IMDB Database

MySQL with IMDBPY. Then we add relevant indexes and foreign keys in the database. Figure 1 shows the relational schema of the IMDB database. We can see that this structure makes query in the database complicated because many additional joins are needed.

### B. Parsing Data in R

## V. Methodologys

### A. Association Rules

### B. Clustering

### C. Classification

### D. Collaborative Filtering

*User-based collaborative filtering*, also know as *k-NN collaborative*,was the first of the automated CF methods. It find other users whose past rating behavior is similar to that of current user and use their ratings on that item to predict what the current user will like.

*Item-based Collaborative Filtering* is similar to User-based CF, it uses similarity between the rating patterns of items. If two items tend to have the same users like and dislike them,then they are similar and users are expected to have similar preferences for similar items.

*1) Computing Predictions:* To compute predictions or recommendations for a user u, user-user CF firstly needs to determine the number N of neighbors will be used to generate the result. Then computing the weighted average of the chosen neighboring users' rating i by using similarity as weights. The formula is given as below:

$$p_{u,i} = \bar{r}_u + \frac{\sum_{u' \in N} s(u,u')(r_{u',i} - \bar{r}_{u'})}{\sum_{u' \in N} |s(u,u')|} \tag{1}$$

In order to eliminate the differences in users's use of the rating scale, subtracting the user's mean rating $\bar{r}_{u'}$ to compensate is necessary. The parameter $p_{u,i}$ is predicated rating on item i for user $u$. $\bar{r}_{u'}$ is average rating on all items rated by user $u$.The parameter $r_{u',i}$ indicates the rating of user $u'$ on item i.$s(u,u')$ is similarity between user u and $u'$. N is the number of neighbors chosen for user $u$.

*2) Measure of Similarity:* An critical parameter used to calculate predications is similarity function. One of the most common and typically measurements is the cosine similarity.

In Cosine Similarity model, users are represented as $|I|$-dimensional vectors of rating on $|I|$ items. Similarity is measured by the cosine distance between two rating vectors. The formula is given below indicating how to calculate the Cosine Similarity between user $u$ and $v$.

$$s(u,v) = \frac{\sum_i r_{u,i} r_{v,i}}{\sqrt{\sum_i r_{u,i}^2}\sqrt{\sum_i r_{v,i}^2}} \tag{2}$$

$r_u$ is rating vector of user $u$.

*3) Evaluation Metrics:* Our goal is to predict the rating a user would give to a restaurant. We predict the rating that user has not rated in the training dataset, but the true rating is stored in the test dataset. We use the root-mean-square error and mean-absolute error for evaluation.

$$RMSE = \sqrt{\frac{\sum \left(r'_{u,i} - r_{u,i}\right)^2}{N}} \tag{3}$$

$$MAE = \sqrt{\frac{\sum \left|r'_{u,i} - r_{u,i}\right|}{N}} \tag{4}$$

Here $r'_{u,i}$ is the predicted rating from user $u$ on item $i$ and $r_{u,i}$ is the true rating; $N$ is the size of test dataset.

Another evaluation method is precision-recall: precision tells us how good the predictions are. In other words, how many were a hit.; recall tells us how many of the hits were accounted for, or the coverage of the desirable outcome.

$$precision = \frac{|\{relevantdocuments\} \bigcap \{retrieveddocuments\}|}{|\{retrieveddocuments\}|} \tag{5}$$

$$recall = \frac{|\{relevantdocuments\} \bigcap \{retrieveddocuments\}|}{|\{retrieveddocuments\}|} \tag{6}$$

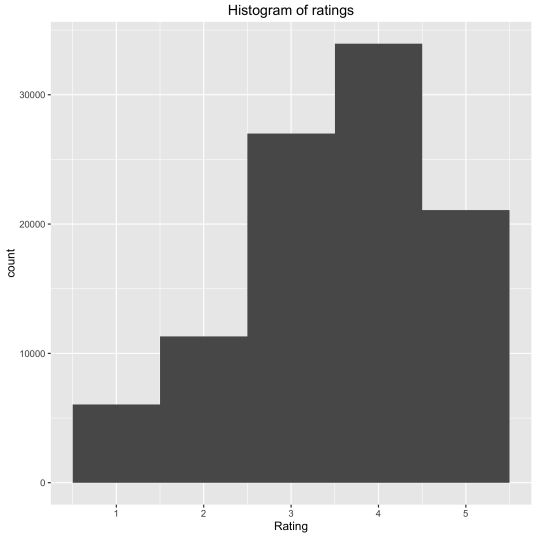


Fig. 2. Rating Distribution

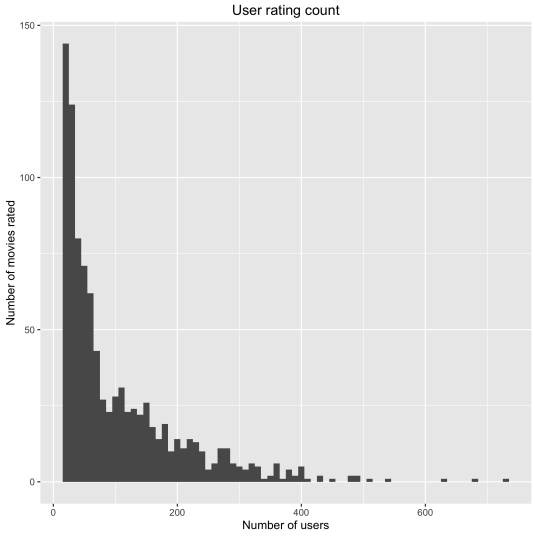


Fig. 3. User Rating Count Distribution

## VI. Experimental Evaluation

### A. title

### B. title

### C. Rating Prediction

*1) Visualizing Data:* Data for this part is from the MovieLens dataset which is a rich resource for recommendation. We firstly visualize the data by ploting several histograms. Figure 1 is the rating distribution of the raw data. The rating range is from 1 star to 5 stars. Then we use a z-score normalization to normalize the data for further analysis. Figure 2 shows the distribution of user's rating count. We can see that many people just reviews quite a few movies. Figure 3 is the distribution of each movie's average rating. It indicates that most movies received 3 to 4 stars.

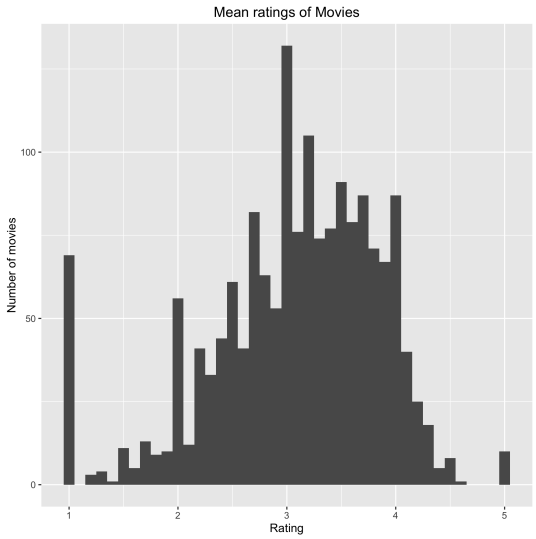


Fig. 4. Movie Average Rating Distribution

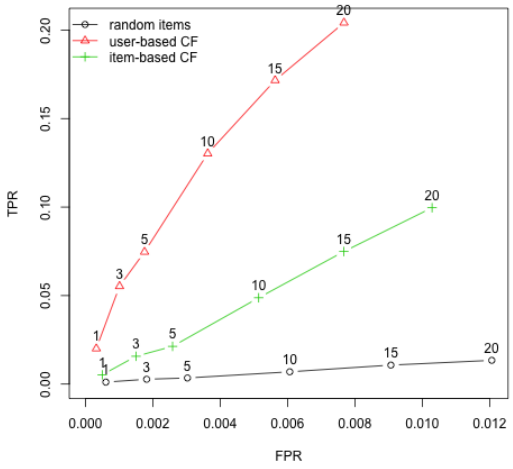


Fig. 6. ROC Curves

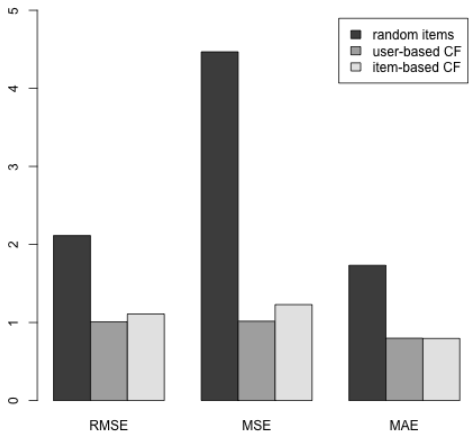


Fig. 5. RMSE

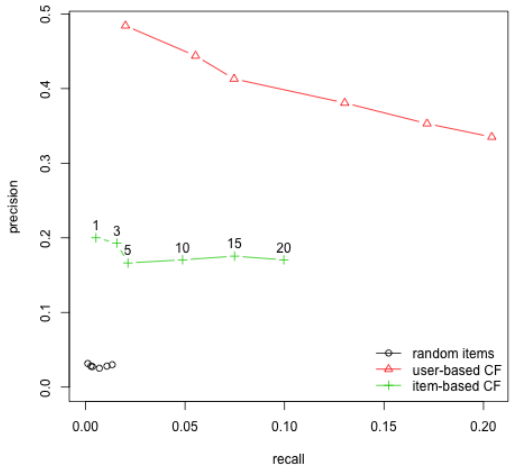


Fig. 7. Precision and Recall Curves

*2) Collaborative Filtering Results:* We apply both user-based CF and item-based CF for rating predicting. Also, we use a random prediction for the baseline. Figure 5 compares the prediction accuracy of the three methods.

Table 1 shows the model time and prediction time of each applied algorithms.

TABLE I
RUNNING TIME OF THREE ALGORITHMS

| Method | model time | prediction time |
|---|---|---|
| Random | 0.025sec | 0.062sec |
| User-based | 0.054sec | 0.859sec |
| Item-based | 37.911sec | 0.591sec |

## VII. Conclusion

The conclusion goes here.


## Acknowledgment

The authors would like to thank...


## References


[1] Jiawei Han, Micheline Kamber, Jian Pei. *Data Mining Concepts and Techniques*, 3rd Edition, 2012.

[2] Usama Fayyad, Gregory Piatetsky-Shapiro, and Padhraic Smyth. *From Data Mining to Knowledge Discovery in Databases*. AI Magazine Volume 17 Number 3, 1996.

[3] Feng Chen, Pan Deng, Jianfu Wan and etc. *Data Mning for the Internet of Things: Literature Review and cChallenges*. International Journal of Distributed Sensor Networks, 2015.

[4] Bruno Pradel, Savaneary Sean, Julien Delporte. *A Case Study in a Recommender System Based on Purchase Data*. Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining, Pages 377-385, 2011.

[5] Daniar Asanov. *Algorithms and Methods in Recommender Systems.* Berlin Institute of Technology, Berlin, Germany, 2011.
[6] http://www.imdb.com/
[7] http://movielens.umn.edu/
[8] http://imdbpy.sourceforge.net/
[9] H. Kopka and P. W. Daly, *A Guide to LaTeX*, 3rd ed. Harlow, England: Addison-Wesley, 1999.