

Knowledge Discovery in the Movie Database

Yuzhou Wang

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

Sainan He

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

Abstract—Abundance of movie data across the internet makes it an obvious candidate for machine learning and knowledge discovery.

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.

Data mining involves different knowledge discovery such as classification, clustering, association analysis. 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 [3]. 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

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

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

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

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 (<http://www.imdb.com/> [4]) 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 (<http://movielens.umn.edu/> [2]). The first one consists of 100,000 ratings for 1682 movies by 943 users. The second one consists of approximately 1 million ratings for 3883 movies by 6040 users.

C. Subsection Heading Here

Subsection text here.

1) Subsubsection Heading Here: Subsubsection text here.

IV. DATA PREPARATION

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:

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

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 .

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

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.

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

$$MAE = \sqrt{\frac{\sum |r'_{u,i} - r_{u,i}|}{N}} \quad (4)$$

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.

VI. EXPERIMENTAL EVALUATION

VII. CONCLUSION

The conclusion goes here.

ACKNOWLEDGMENT

The authors would like to thank...

REFERENCES

- [1] H. Kopka and P. W. Daly, *A Guide to L^AT_EX*, 3rd ed. Harlow, England: Addison-Wesley, 1999.