Movie Recommender System using Supervised Learning Techniques

Ng Shin Cheng[a], Nabilah Filzah Mohd Radzuan*[a], Mohd Norshahriel Abd Rani[a]

[a]*Centre for Emerging Technologies in Computing (CETC), Faculty of Information Technology, INTI International University*

Abstract

Nowadays, with the rapid development of the internet, the huge volume of information and data can present in front of us. Amazon contains up to millions of books, Netflix with up to 10 thousand of movies. With such a large amount of information, it is very difficult for users to find the part that they interested, this is the main reason why the recommendation system is used. The recommendation system can suggest user the products or movies that user interested based on the user's recorded information or history. It is called the as user's behavior. Along with the expansion of domain, comes information overload and difficulty in extraction of data. The number of movies has increased rapidly, therefore the movie recommendation system plays a very significant role. It is used to enhance the user experience by giving fast and coherent suggestions. Users can save a lot of time when searching for the movies they are interested and get new suggestions for movies which are new to them. The recommender system is widely used nowadays not only for movies but other domains also. Example, the recommendersystem is also used in big companies such as Lazada, Shopee, Spotify, Amazon, etc. Infact, the recommender system helps the companies to increase revenue in a large margin as it satisfies the user's need. Technically, there are two main methods used in the recommender system, collaborative filtering method and content-based filtering method. Collaborative filtering method is based on the interaction between items and users so-called "user-item interactions matrix" while content-based filtering method is based on the characteristics of the items and link similar ones where provides results based on the activities of the specific user. A suitable algorithm can enhance the performance and efficiency of the system. It would benefit the users by satisfying their user's experience and improve the accuracy of the system when recommending users movies that suit their preferences.

Keywords: movie, recommendation system, supervised learning, big data

*Corresponding author. Tel.: +6-017-5999732; Fax: +6-06-798 2000E-mail: nabilah.radzuan@newinti.edu.my

1.0 Introduction

Nowadays, information technology is moving with a very fast pace. There are tons of data and information existed, data overload might be occurred such as movies data. It is quite difficult for a user to find the specific movie that the user is looking for within this large volume of data and information. To overcome this issue, recommendation system is introduced, the recommendation system is a solution for users to recommend user related products and items he/she might be interested in. A lot of companies preferrecommendation system because it is effective, and it helped the company to earn more profits. Recommendation system is a type of filtering technique. It is an integrated system data mining algorithm with various functions and user-related data to obtain user preferences. The recommendation system is used to analyze user's behavior based on the passed historical data of the user such as personal preferences of movies and recommend the similar movies for user that he/she might be interested. There are a lot of big companies that are using the recommendation engine such as Amazon and Netflix. 35% of Amazon's revenue is generated by its recommendation engine. 80% ofstream time is achieved through Netflix's recommender system. There are two significant methods in developing a recommendation system. Collaborative Filtering technique and Content-Based Filtering technique. Today, most of the researchers used the collaborative filtering method and content-based filtering method to predicts the user's requirements. However, author will use several classification techniques to improve the accuracy instead of using the collaborative filtering method and content- based filtering method. Supervised machine learning, classification techniques will be applied to automate analytical model building. The algorithms are KNN, SVM and Linear regression. Author will compare these three algorithms and find out which is the most accurate and efficient for the movie recommendation system.

As the most classic type of recommendation algorithm, collaborative filtering includestwo parts: online collaboration and offline filtering. The so-called online collaborationis to find items that users may like through online data, while offline filtering is to filterout some data that is not worth recommending, such as data with a low recommendationscore, or data that the user has purchased although the recommendation value is high. Linear Regressions Clustering and K-Nearest Neighbor techniques are under Collaborative Filtering method. The advantages are no domain knowledge required andthe quality is directly proportional to time. The disadvantage is new user boost-up problem.

Content-based recommendation algorithm is to build a recommendation algorithm model based on the subject-related information, user-related information, and the user's operating behavior on the subject, to provide users with recommendation services. Thesubject-related information here may be metadata information, tags, user comments, and manually labeled information that describe the subject matter in text. User-related information refers to demographic information (such as age, gender, preferences, region, income, etc.). Clustering Decision Trees is the technique used under Content- Based Filtering method. The advantage is no domain knowledge required and the disadvantages are new user-boost up problem and the quality is depending on big data. Other than that, the Hybrid recommendation method is the combination of both Collaborative Filtering method and Content-Based recommendation method. It uses different voting methods and creating single unified model such as linear regression and support vector machine (SVM). The advantage is it can avoid issues like content description. The disadvantage is user boost-up problem.

Recently, most of the researchers attempt several approaches and algorithms to improve the efficiency of the recommendation system. This paper will be comparing 3 techniques that are under classification which is K-nearest neighbors (KNN), linear regression and support vector machine (SVM) and find out the most effective algorithm.

The K-Nearest Neighbors (KNN) algorithm is a classification algorithm, and it is also the simplest and most understandable machine learning algorithm. The application scenarios include character recognition, text classification, image recognition and other fields. The idea of the algorithm is a sample is most similar to the k samples in the dataset, and if most of the k samples belong to a certain category, the sample also belongs to this category. It is suitable to predict what user would rate the movies before they had rated them.

Linear regression is a statistical analysis method that uses mathematical statistics and regression analysis to determine the quantitative relationship between two or more variables. It is used to predict the value of a variable based on the value of another variable. It is helpful in prediction to find the pattern, or the inner relationship among users' rating habits if there exists a pattern in rating records between any given two users.

Support Vector Machine (SVM) is a generalized linear classifier that performs binary classification of data according to supervised learning, and its decision boundary is the maximum-margin hyperplane. SVM uses the hinge loss function to calculate empirical risk and adds a regularization term to the solution system to optimize structural risk. It is a classifier with sparsity and robustness. SVM can perform non-linear classification through the kernel method, which is one of the common kernel learning methods.

2.0 Research Objective

- To execute descriptive analysis and predictive analysis. Descriptive analysis based on the current existing information to find out the top trending movies. Predictive analysis is to find out the similarities of movies genre and ratings.
- To develop recommendations movie system based on existing history information such as movie ratings and taglines on the user's preferences.
- To evaluate movie recommendation system to find out the most effective and accurate algorithm through Supervised Learning Techniques; classification technique to apply into the movie recommendation system.

3.0 Research Question

- What are the attributes involve based on the movie recommendation system?
- How can this movie recommendation be done?
- What type of predictions are used in the movie recommendation system?

4.0 Project Scope

In this project, a research will be conducted on the movie recommender system algorithms which are K-Nearest Neighbors (KNN), Linear regression and Support Vector Machine (SVM) which three of them are under the classification technique and

find out the most efficient technique to apply on the movie recommender system. The system would help to analyze the user's behavior. Then, it would predict the movies that user might interested and suggest for the user.

5.0 Project Limitation

This project would only be focusing on classification techniques which are K-Nearest Neighbors (KNN), Linear regression and Support Vector Machine (SVM), no clustering techniques would be used. Other than that, the dataset is from MovieLens. Itcontains all the movies that is verified by IMDb.

6.0 Research Methodology

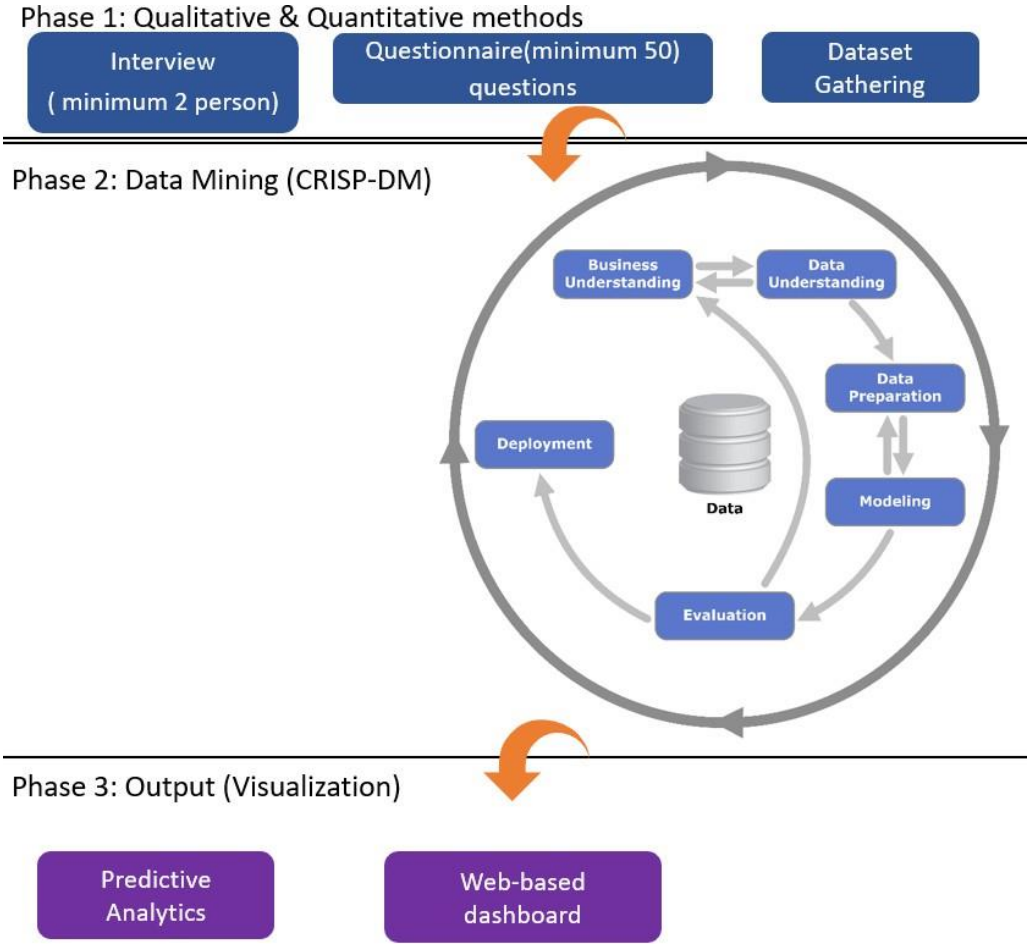

Fig.1. Project phase

CRISP-DM, Cross-Industry Standard Process for Data Mining, is a cross-industry data mining standard process developed by SPSS, Daimler, and other companies. Besides that, it is a process model with six stages, which naturally describes the data science lifecycle. It is a guidance that can help researcher to plan, organize, and implement data science projects. This project will apply CRISP-DM methodology, quantitative method, and qualitative method.

Business Understanding- In this stage is to understand the problem to be solved. It reallyis an iterative process of discovery with the data understanding phase. This purpose ofthis stage is to figure out from a business perspective. There are some specific stages inthis phase, which is determine business objectives, assess situation, determine data mining goals, and produce the project plan. For example, identify the project objective such as to execute descriptive analysis and predictive analysis, to develop recommendations movie system based on existing history information such as movie ratings and taglines on the user's preferences and to evaluate movie recommendation system in order to find out the most effective and accurate through Supervised LearningTechniques. Next, identify the research questions like What are the attributes involve based on the movie recommendation system? How can this movie recommendation be done? What type of predictions are used in the movie recommendation system? Other than that, identify the problem issues. The main problem issue is difficulty for user to find movies that he/she interested within the large volume of data so then the recommendation system is proposed to solve the problem. There are tons of algorithms that is applied to the recommendation system. KNN, SVM and Linear regression are the classification techniques that are going to apply to the recommendation for this project to find out which is the most efficient algorithm.

Data Understanding- This stage is to understand the strengths and limitations of the data. It's also important to know that data cost money and that some data may not be available, it is important to evaluate the costs and benefits of all different potential datasources. There are few specific steps in this stage, which is collecting initial data, describe the data, explore the data, and verify the quality of data. There are two datasetsare going to be used in this project. One is MovieLens dataset and the other dataset is IMDB dataset. Both datasets are from Kaggle. The MovieLens dataset describe ratings and free-text tagging activities from MovieLens, a movie recommendation service. It contains 20000263 ratings and 465564 tag applications across 27278 movies. These data were created by 138493 users between January 09, 1995 and March 31, 2015. There are six csv files which are tag.csv, rating.csv, movie.csv, link.csv, genome_scores.csv and genome_tags.csv. Then, the IMDB dataset consists of 105339 ratings applied over 10329 movies. There are only two files which are ratings.csv and movies.csv. Both datasets are suitable for this project research as it consists large volume of data so that it ensures the accuracy of the results.

Data Preparation- The data preparation stage covers the entire work of integrating the original rough data into the final data set. The data preparation or pre-processing data work in implemented multiple times, and the order of implementation is not predetermined. The tasks of the first stage mainly include tabulation, remove missing values, remove irrelevant data, transformation of data variables, and data cleaning to adapt to modeling tools. According to the relevance of the mining target, data quality and technical limitations, select the data used for analysis, and further clean and transform the data, construct derivative variables, merge the data, and format the data according to the requirements of the tool. Dimension reduction will be applied on this stage, it is a technique that used to reduce the number of input variables in dataset. Themain reason why this technique will be used is because huge numbers of input variables can cause bad performance in machine learning.

Modeling- At this stage, a variety of modeling methods will be selected and used, andthe parameters of the model will be calibrated to the most ideal values through

construction and evaluation. Typically, there are multiple methods to choose from for the same type of data mining problem. If there are multiple technologies to be used, then in this task, each technology to be used must be treated separately. Some modeling methods have specific requirements for the form of data. Therefore, at this stage, it is sometimes necessary to return to the data preparation stage to perform certain tasks. The most common techniques used for movie recommendation are Collaborative Filtering technique and Content-Based Filtering technique. Collaborative filtering is a method that can filter out things that a person would prefer based on reactions from related users. It operates by searching for a wide number of people and discovering a smaller selection of users with preferences identical to a particular user. On the other hand, content-based filtering uses the item function to recommend other items similar to the user's favorite item based on the user's previous actions or clear feedback. Besides that, there are 3 classification techniques will be used which is K-nearest neighbors (KNN), Linear Regression and Support Vector Machine (SVM). The K-Nearest Neighbors (KNN) algorithm is a classification algorithm, and it is also the simplest andmost understandable machine learning algorithm. It is suitable to predict what user would rate the movies before they had rated them. Then, Linear regression is a statistical analysis method that uses mathematical statistics and regression analysis to determine the quantitative relationship between two or more variables. It is used to predict the value of a variable based on the value of another variable. It is helpful in prediction to find the pattern, or the inner relationship among users' rating habits if there exists a pattern in rating records between any given two users. Support Vector Machine (SVM)is a generalized linear classifier that performs binary classification of data according tosupervised learning, and its decision boundary is the maximum-margin hyperplane. SVM uses the hinge loss function to calculate empirical risk and adds a regularization term to the solution system to optimize structural risk. It is a classifier with sparsity and robustness. SVM can perform non-linear classification through the kernel method, which is one of the common kernel learning methods.

Evaluation- In this stage, evaluate the model and review the steps executed to construct the model to ascertain that it properly achieves the business objectives. A key objective is to determine if there is some important business issue that has not been sufficiently considered. At the end of this phase, a decision on the use of the data mining results should be reached Whereas the assess model task of the modelling phase focuses on model accuracy and the model's ability to generalize, the evaluation phase looks more broadly at which model best meets the business and what to do next.

Development- Unless the client has access to its results, the model is not particularly useful. The complexity of this stage varies greatly. The acquired knowledge will need to be organized and presented in a way that customers can use. However, depending on requirements, the deployment phase can be as simple as generating a report, or as complex as implementing a repeatable data mining process throughout the enterprise. After that, a dashboard will be used to visualize all the analyzed data and find out the hidden patterns of the data and come out with a better business strategy to help with the revenue of an organization.

Qualitative Research is a set of deductive logic and many historical facts, it starts from the contradictions of things, describes the things studied. To conduct qualitative research, must directly grasp the main aspects of the characteristics of things based oncertain theories and experience, and temporarily ignore the quantitative differences in

homogeneity. The data is known as non-numerical data. For example, text, video, photographs, or audio recordings. The data can be collected through interviews. Minimum interview 2 person. It can use structured or unstructured interviews and ithas no fixed answers.

Quantitative Research involves the task of objectively gathering and evaluating numerical data for the purpose of identifying, forecasting, or monitoring variables of interest. Quantitative research intends to examine the causal interaction between factors, make forecasts, and generalize findings to larger populations. It can be done by experiment method such as questionnaire. Minimum 50 questions would be conducted. For example, execute a survey with tons of questions that has rating scale or 'yes' or 'no' to gather data from user. The other method is data analysis. Statistics can turn the quantitative data into useful information, it could help with decision making byanalyzing the graph after getting the information from users.

7.0 Target Audience

The target audience would be the users that used movie websites. They would be benefited because the users would save a lot of time to search for the movies they like. Instead of searching for the movies, the system would automatically provide a suggestion list for them. There are the movies that the users interested or new movies inside the list.

8.0 Result and Discussion

Data model evaluation will be conducted to ensure the accuracy and efficiency of the model. Besides that, user testing will be executed for the web-based dashboard system to avoid errors and bugs that might be occurred. For the user testing part, all the pages and functions will be tested and evaluated such as validation function and user's experience towards the proposed system.

After training and testing the dataset, K- Nearest Neighbor, Random Forest and Decision Tree have the highest accuracy score which is 100%. Followed by Support Vector Machines and Linear Regression, 35.56% and 35.57. This shows that this dataset is suitable to perform its maximum potential by using the K-Nearest Neighbor. Therefore, K-Nearest Neighbor has been selected to build the movie recommender system model based on user's ratings as the ratings values are suitable for applying the supervised learning technique K-Nearest Neighbor.

Techniques		
K-Nearest Neighbor	**Support Vector Machine**	**Linear Regression**
Accuracy: 100%	**Accuracy**: 35.56%	**Accuracy**: 35.47%
The accuracy score of K-Nearest Neighbor is 100%. The result is the most accurate compared to the other 2 techniques. Therefore, it is suitable to use to implement the movie recommender with this dataset by using this technique.	The accuracy score of Support Vector Machine is 35.56%. The result is less accurate. Therefore, it is not suitable to use implement the movie recommender system model with the dataset by using this technique.	The accuracy score of Linear Regression is 35.47%. The result is the least accurate among these 3 techniques. Therefore, it is not suitable to used to implement the movie recommender system model with the dataset by using this technique.

Table.1. Comparison Result

9.0 Conclusion

Conclusion, the steps of developing the movie recommender system have been stated clearly. The objectives, problem statement, project scope, project limitation and target audience are precise so that this project could be developed successfully. Other than that, the methodology of this project is CRISP-DM methodology. It is a process model that describes six stages of the data science life cycle, it is a guidance that can help researcher to plan, organize, and implement data science projects. Then, the quantitative and qualitative research will be conducted in this project to gather data and information. The quantitative research can be done by experiment method such as questionnaire. On the other hand, the qualitative research can be done by interview.

References

Yin Zhang. "A Sentiment-Enhanced Hybrid Recommender System for Movie Recommendation: A Big Data Analytics Framework." 22 Mar 2018 <https://www.hindawi.com/journals/wcmc/2018/8263704/>. Accessed 25 Jan. 21

Phonexay Vilakone, Doo-Soon Park & Khamphaphone Xinchang. "An Efficient movie recommendation algorithm based on improved k-clique". 13 December 2018. <https://hcis-journal.springeropen.com/articles/10.1186/s13673-018-0161-6>. Accessed 25 Jan. 21

F. Furtado, A, Singh. "Movie Recommendation System Using Machine Learning".Int. J. Res. Ind. Eng. Vol. 9, No. 1 (2020) 84–98. 15 March 2020. <http://www.riejournal.com/article_106395_c6c0038f1bf5d4c421bd552d0541d6be.p df>. Accessed 25 Jan. 21

Ananya Agarwal, S. Srinivasan. "Movie Recommendation System". <https://www.irjet.net/archives/V7/i7/IRJET-V7I7199.pdf>. Accessed 25 Jan. 21

Nirav Raval, Vijayshri Khedkar. “A Review Paper On Collaborative Filtering Based Moive Recommedation System”. INTERNATIONAL JOURNAL OF SCIENTIFIC & TECHNOLOGY RESEARCH VOLUME 8, ISSUE 12, DECEMBER 2019. <http://www.ijstr.org/final-print/dec2019/A-Review-Paper-On-Collaborative- Filtering-Based-Moive-Recommedation-System-.pdf>. Accessed 25 Jan. 21

Manoj Kumar, D.K. Yadav, Ankur Singh, Vijay Kr. Gupta. “A Movie Recommender System: MOVREC”. International Journal of Computer Applications (0975 – 8887) Volume 124 – No.3, August 2015. <https://www.ijcaonline.org/research/volume124/number3/kmar-2015-ijca-904111.pdf>. Accessed 25 Jan. 21

Ramni Harbir Singh, Sargam Maurya, Tanisha Tripathi, Tushar Narula, Gaurav Srivastav. “Movie Recommendation System using Cosine Similarity and KNN”. International Journal of Engineering and Advanced Technology (IJEAT) ISSN: 2249 – 8958, Volume-9 Issue-5, June 2020. <https://www.ijeat.org/wp-content/uploads/papers/v9i5/E9666069520.pdf>. Accessed 30 Jan. 21

“What is CRISP DM?”. <https://www.datascience-pm.com/crisp-dm-2/>.

Xinyang Ge, Jia Liu, Qi Qi, Zhenyu Chen. “A New Prediction Approach Based on Linear Regression for Collaborative Filtering”. <https://www.microsoft.com/en- us/research/wp-content/uploads/2016/10/fskd11-1.pdf>. Accessed 30 Jan. 21

Xibin WANG, Fengji LUO, Chunyan SANG, Jun Zeng. “Personalized Movie Recommendation System Based on Support Vector Machine and Improved ParticleSwarm Optimization”. February 2017. <https://www.researchgate.net/publication/313229640_Personalized_Movie_Recomm endation_System_Based_on_Support_Vector_Machine_and_Improved_Particle_Swa rm_Optimization>. Accessed 5 Feb. 21

Moa Andersson and Lisa Tran. “Predicting movie ratings using KNN”. DEGREE PROJECT, IN COMPUTER SCIENCE, FIRST LEVEL STOCKHOLM, SWEDEN 2020. 8 June 2020.<https://www.diva-portal.org/smash/get/diva2:1464572/FULLTEXT01.pdf>. Accessed 5 Feb.