Project Proposal

Sainan He, 20635929
Yuzhou Wang, 20609396

February 18, 2016

Make Successful Movie

Overview

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 refers to the process of analyzing large databases, discovering novel, interesting, and potentially useful patterns, and applying algorithms to extract information, that is dealing with “knowledge discovery in databases.” It usually involves six tasks: anomaly detection, classification, clustering, association analysis, regression, and summarization. The most widely used application of data mining is prediction. For example, some of the major e-commerce sites, like Amazon and Netflix, have system to generate personalized recommendation to their customers. They need to predict what the customer are most interested in. The prediction is based on known attributes of the customer such as age, gender, view and purchase history, etc.

Data mining are supported by many domains, like statistics, machine learning. In particular, by providing efficient data storage, index structure and data query, database system plays an important role in data mining. Before data mining transforms massive data into meaningful patterns and rules, it is necessary to get plentiful and integrated data. With the progress of technology, relational databases with large capacity and well-organized structure are to store the required data used for data analysis and prediction support. After data preparation, such as transformation and data reduction, new data in the database will be clean without inadequacy ones, outlets or useless data. Based on the data after preprocessing, there are various kinds of algorithms to realize data mining, such as discriminant analysis, cluster analysis, decision trees and neural network.

In our project, we will do some researches on data mining area involving in both literature background research and application-oriented work.

Literature Review

We have read several related papers on data mining, including:

[1] Feng Chen, Pan Deng, Jianfu Wan and etc. Data mining for the internet of things: literature review and challenges. International Journal of Distributed Sensor Networks, 2015.

This is a review article surveying data mining through 3 perspectives. From the knowledge view, it illustrate 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.

[2] 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.

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 didn't 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.

[3] Daniar Asanov. Algorithms and Methods in Recommender Systems. Berlin Institute of Technology, Berlin, Germany, 2011.

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 neighborhood, 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.

Goal

Our project seeks to apply data mining on discovering and extract useful information from the IMDB datasets, such as evaluating how successful a movie is, what factor impact their rankings and predicting user ratings of a future movie.

Dataset Description

IMDB (Internet Movie Database) is an excellent resource to find rich information about movies. It provides more than 50 files of data which consist of many different attributes about movies, such as movie title, genre, actors/actresses, directors, year, rating, budget, etc.

Methodology

Our project mainly includes the following works:

1) Data Collection: IMDB data are available online through IMDB alternate interface.
2) Data Preparation. The IMDB provides more than 50 separated plain text data files which are natural language but not machine readable. It is difficult to perform data mining directly, so our first step would be data transformation.

3) Data Storage. In order to analyzing and mining data, we need to create a database and related tables for storing these data.

4) Data Mining. Use data mining strategies for movie evaluation and prediction, including feature selection, classification, etc. A further work is to predict user rating of the movies that they haven't review yet, and recommend new movies to users by prediction rankings.

Reference

[1] Feng Chen, Pan Deng, Jianfu Wan and etc. Data mining for the internet of things: literature review and challenges. International Journal of Distributed Sensor Networks, 2015.

[2] 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.

[3] Daniar Asanov. Algorithms and Methods in Recommender Systems. Berlin Institute of Technology, Berlin, Germany, 2011.