Recommendation System in E-Commerce

Group 4

Member names

SUN, Yimiao / 20566220

LAN, Hao / 20549894

TAI, Tsen Jung / 20567298

CAO Pei / 20576184

Hong Kong University of Science and Technology

---

1 Problem Description

Recommendation algorithms are to personalize the online store for each customer. They are known for their use on e-commerce websites, where they use input about a customer's interests to generate a list of recommended items.

Based on the real user behavior data and location information data of Alibaba's mobile e-commerce platform, this problem deals with personalization recommendation system for users. We aim to excavate the rich connotation behind the data, and accurately recommend suitable item for mobile users. To implement the recommendation system, we intend to use the memory-based approach which has few modeling assumptions and few tuning parameters to learn and is easy to explain to users. The task details are as follows:

- Task 1: Build a user-based recommendation algorithm to find the similarity between users, cluster users, and pick out users with similar preferences. (Collaborative Filtering, Clustering)[1]
- Task 2: Build a item-based recommendation algorithm to find the similarity between items, find the frequent pattern. After that we can aggregate the similar items and recommends them. (Collaborative Filtering, Frequent Pattern Mining)[2]

Finally, we will predict users' purchase behavior in the future based on their historical behavior data. We will compare performances of the user-based model and the item-based model by using the precision, recall and F1 score as the criterion.

2 Dataset Description

The dataset is provided by Tianchi Competitions[3], Alibaba Cloud and its size is approximately 1 GB. It records the behaviour that a user performed on an item at a specific time. The records are in csv format and are divided into two collections: *Item* and *User*. And there are 620,000 data of *Item* and 23,000,000 data of *User*. The tables have the following schemas:

Table	***Column***	***Description***
item	item_id	The unique identifier of an item.
	item_geohash	The encoded geolocation of the item; the value can be NULL
	item_category	The classification that an item belongs to.
User	user_id	The unique identifier of a user.
	item_id	The identity of the item that is bought by the user in the current behavioural record.
	behavior_type	The type of behaviour that the user performed on the item. The values 1, 2, 3, 4 stand for browsing, add-to-wish-list, add-to-cart and checkout.
	user_geohash	The encoded geolocation of the user; the value can be NULL
	item_category	The classification that an item belongs to.
	time	The time that the user took the action.

Table 1. the schemas of dataset

The dataset provides the possibilities of performing item clustering, user clustering and recommendation. Items can be grouped based on the time relation that a set of items are interested by the same user and the correlation on the set of items across different users. And the users sharing similar interests and behavioural patterns can be aggregated. Based on the constructed clusters and other analytical methods, a recommendation system can be built to cater to each user's need.

3 Technology Description

3.1 RDD

In this project, we will implement RDD for our data preprocessing to improve the efficiency of preprocessing.

3.2 Spark SQL

Spark SQL provides such interfaces that can provide more information about the structure of the data. In this project, we will use Spark SQL mainly for data analysis and data preprocessing after which we can have some insights from the data.

3.3 MLlib/sklearn/TensorFlow

We will implement MLlib in our project and if possible, we want to compare the performance of MLlib with that of sklearn module, a machine learning module in Python, and TensorFlow.

3.4 GraphX/Tableau/Matplotlib

In this project, we want to implement GraphX for our data analysis. We would also use Tableau or Matplotlib in Python as auxiliary tools.

---

[1] John S. Breese, David Heckerman, Carl Myers Kadie: Empirical Analysis of Predictive Algorithms for Collaborative Filtering. UAI 1998: 43-52

[2] Linden, G., B. Smith, and J. York. "Amazon.com recommendations: item-to-item collaborative filtering." IEEE Internet Computing 7, no. 1 (January 2003): 76-80. http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=1167344

[3] https://tianchi.aliyun.com/competition/entrance/231522/introduction