Team 07 - Project Proposal

CS6220 - Data Mining Techniques - Fall 2017
Northeastern University

Nakul Camasamudram, Rosy Parmar, Rahul Verma, Guiheng Zhou

October 25, 2017

1 The Dataset

Our team will be using "The Instacart Online Grocery Shopping Dataset 2017". Instacart is an American company that operates as a same-day grocery delivery service.[1] This anonymized dataset contains a sample of over 3 million grocery orders from more than 200,000 Instacart users. For each user, the dataset has 4 to 100 of their orders, with the sequence of products purchased in each order. The week and hour of the day the order was placed, and a relative measure of time between orders is also available.[2].

File Descriptions:[3] Each entity (customer, product, order, aisle, etc.) has an associated unique id.

- **aisles.csv**

```
aisle_id,aisle
1,prepared soups salads
2,specialty cheeses
3,energy granola bars
...
```

- **departments.csv**

```
department_id,department
1,frozen
2,other
3,bakery
...
```

- **order_products__*.csv**: These files specify which products were purchased in each order. "order_products__prior.csv" contains previous orders for all customers. "reordered" indicates that the customer has a previous order that contains the product.

```
order_id,product_id,add_to_cart_order,reordered
1,49302,1,1
1,11109,2,1
1,10246,3,0
...
```

- **orders.csv**: This file tells to which set (prior, train, test) an order belongs.

```
order_id,user_id,eval_set,order_number,order_dow,order_hour_of_day,
    days_since_prior_order
2539329,1,prior,1,2,08,
2398795,1,prior,2,3,07,15.0
473747,1,prior,3,3,12,21.0
...
```

- **products.csv**

```
product_id,product_name,aisle_id,department_id
1,Chocolate Sandwich Cookies,61,19
2,All-Seasons Salt,104,13
3,Robust Golden Unsweetened Oolong Tea,94,7
...
```

2 Questions to be answered

Our goal is to create a product recommendation system using the Instacart data that would answer the following questions

- Given a set of products in a customer's basket, what is another associated set of products he/she is likely to buy?
- Given a customer, what products could be recommended to him/her so that a purchase would be made?

3 Algorithms

We plan on applying a subset of the below algorithms to explore relationships in the dataset and answer the above mentioned questions.

- Association rule learning: Apriori Algorithm, FP Growth
- Recommender systems: Collaborative filtering
- Other association exploration: Clustering, Word2vec

4 Division of work

- **Nakul**: Documentation, Domain Research, Algorithm Research and Usage, Find basic stats and answers, Attribute research and selections, Dimension Reduction
- **Rosy**: Documentation, Algorithm Research and Usage, Research and experimentation of libraries, Data cleaning
- **Rahul**: Documentation, Algorithm Research and Usage, Find basic stats and answers
- **Guiheng**: Documentation, Algorithm Research and Usage, Research and experimentation of libraries, Data cleaning

References

[1] https://scholar.google.com/citations?hl=en&user=09kJn28AAAAJview$_op$ = $list_workssortby = title$

[2] Stanley, Jeremy. "3 Million Instacart Orders, Open Sourced tech-at-instacart." Tech-at-instacart. May 03, 2017. Accessed October 11, 2017. https://tech.instacart.com/3-million-instacart-orders-open-sourced-d40d29ead6f2.

[3] Instacart. "Instacart Market Basket Analysis - Data." Instacart Market Basket Analysis. Accessed October 11, 2017. https://www.kaggle.com/c/instacart-market-basket-analysis/data.