Frequent Item Sets Mining for Recommender Systems

CMSC 5741 Group 7 - Project Proposal

Ziwen LU
1155155161@link.cuhk.edu.hk
Department of Information Engineering
The Chinese University of Hong Kong

Yan WU
1155148594@link.cuhk.edu.hk
Department of Information Engineering
The Chinese University of Hong Kong

Yaling ZHANG
1155147233@link.cuhk.edu.hk
Department of Information Engineering
The Chinese University of Hong Kong

Bowen FAN
1155155953@link.cuhk.edu.hk
Department of Information Engineering
The Chinese University of Hong Kong

ACM Reference Format:
Ziwen LU, Yan WU, Yaling ZHANG, and Bowen FAN. 2020. Frequent Item Sets Mining for Recommender Systems: CMSC 5741 Group 7 - Project Proposal. In *Proceedings of ACM Conference (Conference'17).* ACM, New York, NY, USA, 2 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

The retail industry took a distinguished turn with the flourish of online shopping. With the speed and convenience of online retail, it has become easier for consumers to get what they want when they want it. Moreover, due to the influence of COVID-19, people are more inclined to shop online recently. However, the online shopping industry can be cutthroat. This is why understanding online shopping statistics is more important now than ever to get ahead of the competition.

1.1 Motivation

Customers always have great expectations from brands they are interested in. In this case, providing customized product recommendation to different customers will increase the retailers' competitiveness. Therefore effective recommendation system which filters a large scale of information becomes necessary. However, the ongoing rapid expansion of online shopping and the diversity of customers' interests makes it difficult to conduct recommendation. To further illustrate such difficulty, firstly, the number of digital buyers worldwide keeps climbing every year. In 2019, an estimated 1.92 billion people purchased goods or services online. During the same year, e-retail sales surpassed 3.5 trillion U.S. dollars worldwide, and according to the latest calculations, e-commerce growth will accelerate even further in the future [1]. Secondly, customers' shopping behavior can be affected by different factors such as regions, personal habits, global events, etc. For example, In a world-wide statistics, the top online categories for purchasing are fashion (61%), travel (59%), books and music (49%), IT (47%), and events (45%), but in the Asia Pacific, the most popular online industries are packed groceries (40%), home care (37%), fresh groceries (35%), and video gaming (30%) [2]. Moreover, retail platforms have undergone an unprecedented global traffic increase between January 2019 and June 2020, surpassing even holiday season traffic peaks. Overall, retail websites generated almost 22 billion visits in June 2020, up from 16.07 billion global visits in January 2020. This is of course due to the COVID-19 which has forced millions of people to stay at home in order to stop the spread of the virus [3]. Considering all these factors above, it does make sense to analyze a large scale of data set and extract inspiring advice for nowadays' recommendation systems.

1.2 Objectives

Many other academic researches tend to focus on improving the efficiency or capability of recommender systems. Based on several topics of CMSC5741, this project aims to present more dimensions in regards to item recommendations. For instance, what brands should be recommended to shoppers, at what range of price shoppers are inclined to buy a product, when should recommendations be given to shoppers according to their shopping preference, what kind of similar items should be recommended, etc.

1.3 Deliverables

This project will provide following items:
1. A prototype system that recommend goods based on the frequency of items purchased by similar users
2. Clusters of similar users according to their purchase behavior, if possible

1.4 Relevance to the Course

This project will be relevant to the following topics taught in our course:
1. MapReduce:Upon the purchase data(i.e. product ID)is achieved about each user, then giving the product ID to each user to make the user becomes a basket.
2. Frequent Itemsets: Analyze all commodity purchase data and find some goods that the support threshold is over X. In those frequent itemsets, other goods could be the recommend goods for customers.

Conference'17, July 2017, Washington, DC, USA

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM…$15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

3. Clustering: Obtain the clusters of similar users,if possible.
4. 4V Feature: Enough volume of datasets will be included.

2 RELATED WORK

As for the methods of dealing with selected frequent itemsets, [4] proposed a method on decision trees to further enhance the accuracy. Its' approach mainly relies on the relationship between customer purchase history and optimal recommendation goods.[5] has suggested a method of relevant set correlation to do clustering in recommender system. The similarity of two items depends on the number of common neighbours they have.

3 METHODOLOGY

3.1 Dataset

This project uses E-commerce behavior data from multi-category store available on[6] as datasets.The datasets contain 285 million users' purchasing events from a large multi-category online store for a duration of 7 months, dated from October 2019 to April 2020. Each dataset is shown in a CSV formart.The statistics of the datasets is summarized in Table 1.

Table 1: The Statistics of Datasets

Dataset	Size	Records
2019-Oct	5.27GB	42,448,765
2019-Nov	8.38GB	67,501,980
2019-Dec	8.71GB	67,542,879
2020-Jan	7.24GB	55,967,042
2020-Feb	7.14GB	55,318,566
2020-Mar	7.28GB	56,341,242
2020-Apr	8.62GB	66,589,269
Total Size of Datasets	52.64GB	
Total No. of Records	411,709,743	

3.2 Overview of Main Techniques and Algorithms

To handle such big volume of data, this project will first extract and sort topic relevant information with cloud MapReduce. The expected output will be UserId, ProductID, Product Brand and Product Categories.With such output, this project treats each user as a basket and applies A-priori algorithm to find the frequent itemsets from each user. The frequent itemsets give us information about what products are usually bought together and therefore can be used to recommend related products.

To be clearer, the process can be shown below:

- ★ Step1 Pre-processing: conduct MapReduce to select information needed.
- ★ Step2 Implementation: use frequent itemsets algorithm (Apriori/SON/PCY) to choose recommended goods.
- ★ Step3 Improvement: add decision trees algorithm to reach more accurate result.
- ★ Step4 Bonus: make full use of datasets to analyze customer behaviour of purchase and create visualized analysis.

3.3 Evaluation Methods

Instead of showing all the frequent itemsets at the output, this project will inquire a user what products he/she is interested in. Based on users' responses, the system can scan through frequent itemsets to find related products for recommendation. The users are encouraged to input more than one product, as the recommendation accuracy can be improved with multiple inputs. The option of implementing a decision tree when scanning through the frequent itemsets for a more reliable result is considered.

4 EXPECTATION

Table 2 shows the expected project milestone.

Table 2: Project Planner

Time	To do list
2020/10/27-2020/11/3	Search Dataset,Determine Topic Write Proposal
2020/11/4-2020/11/10	Set up Environment, Define Existing Algorithm
2020/11/11-2020/11/17	Implement in local environment, Debug
2020/11/18-2020/11/24	Implement on Hadoop platforms, Debug
2020/11/25-2020/12/2	Summarize project codes, Write Report and slides, Record demo video

Here are some immatured ideas about what to explore further in this project:
1. Do similar user clustering, so that the interface can present both recommended goods and how many similar users are purchasing this kind of products.
2. Analyze customer purchase behaviours via the information of purchasing events, price and brand preference.

REFERENCES

[1] Nestor Gilbert. 74 compelling online shopping statistics: 2020 data analysis market share, 2020.
[2] J. Clement. E-commerce worldwide - statistics facts, 2020.
[3] J. Clement. Covid-19 impact on global retail e-commerce site traffic 2019-2020, 2020.
[4] Daniel Nikovski and Veselin Kulev. Induction of compact decision trees for personalized recommendation. In *Proceedings of the 2006 ACM Symposium on Applied Computing*, SAC '06, page 575–581, New York, NY, USA, 2006. Association for Computing Machinery.
[5] Nkechi J. Nnadi. Applying relevant set correlation clustering to multi-criteria recommender systems. In *Proceedings of the Third ACM Conference on Recommender Systems*, RecSys '09, page 401–404, New York, NY, USA, 2009. Association for Computing Machinery.
[6] Michael Kechinov. E-commerce behavior data from multi category store, 2019.