Big Data Technology and Applications Project proposal

CMSC 5741: Group 1

Junjie LIANG*
Department of Computer Science and Engineering
The Chinese University of Hong Kong
1155155009@link.cuhk.edu.hk

Pengliang SUN*
Department of Computer Science and Engineering
The Chinese University of Hong Kong
1155148009@link.cuhk.edu.hk

Qingdan ZHENG*
Department of Computer Science and Engineering
The Chinese University of Hong Kong
1155154714@link.cuhk.edu.hk

Hongquan ZHANG*
Department of Computer Science and Engineering
The Chinese University of Hong Kong
1155148260@link.cuhk.edu.hk

1 Background And Motivation

Singles Day or 11.11 is the biggest shopping day of the year for singles, as Chinese e-commerce players led by Alibaba offer massive discounts on everything on Taobao. We would like to know more about the user behaviors on Taobao before the festival happens so that we can provide insights about the recommendation.

In order to cope with the massive data produced by Alibaba, we adopted big data technologies to do analyse and modeling, as well as mining the relationship between users, shopping items and categories.

2 Research Question

2.1 Question And Tasks

In this project, we use dataset offered by Alibaba, as mentioned in section 2.2, to do analysis and modeling. Since the data is massive, it is difficult to run the algorithms real-time on PC, and we adopted some big data services from AWS. We mainly focus on the following tasks:

- Adopting MapReduce framework to do data prepossessing, like removing NAN data and noisy data, picking data which related to out tasks, etc.
- Adopting MapReduce framework to implement the Apriori algorithm[1], and find frequent itemsets and other results that contribute to building recommender system.
- Finding a proper implementation for K-means[2] to do big data analysis, and try to mine the relationship between users, shopping items and categories.
- Trying to build a recommender system[3].

2.2 Dataset

This project uses the dataset, User Behavior Data from Taobao for Recommendation[1], offered by Alibaba. The data is generated by 1 million users who have behaviours including click, purchase, adding item to shopping cart and item favouring during November 25 to December 03, 2017. The size of the data is totally 3.42GB. Table[1] is the detailed descriptions of each field. More details about the dataset are illustrated in Table[2].

Dimension	Number
Users	987,994
Items	4,162,024
Categories	9,439
Interactions	100,150,807

Table 2: Summary of the data

2.3 Related Courses Topics

This projects is related to the courses topics as following:

- MapReduce. Transform the data format in order to effectively apply our major algorithms.
- Frequent Itemsets. Find out the internal relationship among the purchased items.
- Clustering. Group the objects with the same attributes.
- Recommender System/Matrix Factorization. Build a model for providing better recommendations.

*Every authors contributed equally to this research.

[1]https://tianchi.aliyun.com/dataset/dataDetail?dataId=649

Field	Explanation
User ID	An integer, the serialized ID that represents a user
Item ID	An integer, the serialized ID that represents an item
Category ID	An integer, the serialized ID that represents the category which the corresponding item belongs to
Behavior type	A string, enum-type from ('pv', 'buy', 'cart', 'fav')
Timestamp	An integer, the timestamp of the behavior

Table 1. Descriptions of each field

3 Plan

3.1 Techniques And Algorithms

We planned to used the following techniques and algorithms in this project:

- MapReduce: To do data processing.
- Apriori Algorithm: To compute the frequent itemsets.
- K-means: To do clustering and mine some relationships.
- Collaborative Filtering: To build a recommender system.

3.2 Main Function And Analysis Results

- Results of maximum frequent itemsets of purchased items or the corresponding categories.
- Implicit relations among users, items and categories.
- Recommendations for the top sale items.

3.3 Rough Timelines

- Nov. 15: AWS environment setup
- Nov. 20:
  - MapReduce data processing
  - MapReduce to implement Apriori
- Nov. 25: Implement K-means for MapReduce and big data
- Dec. 03: Recommender system
- Dec. 08: Report

3.4 Deliverables

- Final report
- Presentation slides
- Demo video
- Source code

References

[1] Sudhakar Singh, Rakhi Garg, and P. K. Mishra. Review of apriori based algorithms on mapreduce framework. *CoRR*, abs/1702.06284, 2017.

[2] P. S. Bradley, Usama Fayyad, and Cory Reina. Scaling clustering algorithms to large databases. In *Proceedings of the Fourth International Conference on Knowledge Discovery and Data Mining*, KDD'98, page 9–15. AAAI Press, 1998.

[3] Han Zhu, Xiang Li, Pengye Zhang, Guozheng Li, Jie He, Han Li, and Kun Gai. Learning tree-based deep model for recommender systems. *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery Data Mining*, Jul 2018.