Mining Massive Data Sets Midterm Report

1st 522H0036 - Luong Canh Phong
Faculty of Information Technology
Ton Duc Thang University
Ho Chi Minh City, Vietnam
522H0036@student.tdtu.edu.com

2nd 522H0092 - Cao Nguyen Thai Thuan
Faculty of Information Technology
Ton Duc Thang University
Ho Chi Minh City, Vietnam
522H0092@student.tdtu.edu.com

3rd 522H0075 - Tang Minh Thien An
Faculty of Information Technology
Ton Duc Thang University
Ho Chi Minh City, Vietnam
522H0075@student.tdtu.edu.com

4th 522H0167 - Truong Tri Phong
Faculty of Information Technology
Ton Duc Thang University
Ho Chi Minh City, Vietnam
522H0167@student.tdtu.edu.com

5th Instructor: Nguyen Thanh An
Faculty of Information Technology
Ton Duc Thang University
Ho Chi Minh City, Vietnam
nguyenthanhan@tdtu.edu.com

***Abstract*—In the age of big data, the ability to mine and extract valuable information from massive datasets can give the user an unparalleled edge against the competition. Therefore, this requirement made by the lecturer is designed to simulate one of the three most fundamental challenges in data mining. Through these series of tasks, we will explore some algorithm implementations and solve different problems as well as explore their trade-offs. Each task is a different algorithm to explore and implement with their corresponding datasets. Through these tasks, we will gain some practical insight and experience in working with these algorithms as well as a better understanding of their pros and cons to be able to cater to each dataset based on their characteristics.**

I. Introduction

This report is divided into three large sections corresponding to the first three tasks provided by the lecturer. We will explore and present our findings while putting the proposed algorithms into practice.

Task 1 proposes utilizing the A-Priori algorithm in a Hadoop MapReduce program to discover groups of customers shopping on the same date as well as interacting with Hadoop Distributes File System (HDFS) to store files. By applying these methods, we will be able to understand how to extract patterns from large datasets locally.

The second task focuses on implementing the Park-Chen-Yu (PCY) algorithm using Object-Oriented Programming (OOP) principles and PySpark DataFrame to identify frequent item pairs and generate association rules from customer purchase data stored in Google Drive. The implementation, while generating association rules, also has to follow object-oriented programming principles inspired by PySpark's Frequent-Pattern Growth (FPGrowth) class.

In the third task, we will implement and compare the MinHashLSH algorithm and an alternative of our choice - in this case, a manual method of calculating Jaccard distance. Both of these approaches should achieve the same goal of searching for similar pairs of dates where the Jaccard distance is above a predetermined threshold. After that, we will visualize their runtime with their threshold ranging from 0 to 1 with 0.1 increments to gauge their performance and outline some characteristics between both approaches. Through these implementations, we demonstrate practical applications of data mining techniques with a given dataset. With these findings, we highlight the trade-offs between various aspects across different algorithms within the given time and constraints.

II. First Task: A-Priori Algorithm for Frequent Customers

III. Second Task: PCY Algorithm for Frequent Items

IV. Third Task: MinHashLSH for Similar Dates

V. Contribution

The following table represents the contribution of each member, note that whichever member handles whichever task will also write the report for that task.

TABLE I
Member Contributions

ID	Member	Contribution	Progress
522H0036	Luong Canh Phong	Task 1 and Handling Report	100%
522H0092	Cao Nguyen Thai Thuan	Overseer and Report Support	100%
522H0075	Tang Minh Thien An	Task 3	100%
522H0167	Truong Tri Phong	Task 2	100%

VI. Self-evaluation

The following table is our self-evaluation on our tasks:

TABLE II
Self-evaluation

Task	Task Requirements	Completion Ratio
Task 1	A-Priori Algorithm for Frequent Customers	100%
Task 2	PCY Algorithm for Frequent Items	95%
Task 3	MinHashLSH for Similar Dates	90%
Task 4	Report	100%

VII. Conclusion

References