CSCI 5510 Project Proposal

**The recommendation system of merchandises**

Group **No.17**

HUANG, Xuming 1155037043

Isaacwym@gmail.com

YE, Zishi 1155040506

felixleafer@gmail.com

QUAN, Cong 1155039072

quancong1990@163.com

# Motivation

Since the emphasis of the Big Data Analytics course is analyze massive dataset and familiarize us with modern techniques and tool chains in analyzing big data. So our project will focus on utilizing the modern techniques to analyze great amount of data and synthesizing the result of our analysis. A well-behavior merchandises recommendation system needs to mine valuable information among massive dataset and build relations between different clusters. We would analyze the relationship among great amount of dataset which are related to customers’ purchasing behaviors and social activities. And we would also find the relationships between people, societies and merchandises in order to recommend relatively appropriate merchandises to customers.

With the help of Hadoop, we implement Map-Reduce framework to deal with the massive dataset in order to retrieve the connections and relations between various merchandises types and different social groups. Since analyzing the data and transactions in the area of E-commerce market is a challenging task, we are more likely to apply the knowledge derives from the course to collect more than 10 gigabytes of typical data and mine their interrelations to accomplish our project.

# Background

In recent years, E-commerce has become increasingly popular around the world. Lots of people prefer purchasing merchandises online rather than in real store. Many famous E-business companies such as Taobao, Amazon have sprung up and developed rapidly in last ten years. There are thousands of hundreds transactions happened in these companies’ website every minutes. In other words, these companies face with high competition, too. Moreover the requirements of customers are continuously increasing. So E-business companies have to improve their services and catch up with customers’ thoughts. A merchandises recommendation system which makes use of customers’ purchasing behaviors and thus build connection between several kinds of merchandises and certain customer group, can wisely recommends the merchandises which are more likely to be selected by certain customers as needed.

Even though Taobao and Amazon embrace mature recommendation system which can recommend almost “right” merchandises to certain customers according to their transaction records, and their systems are able to recommend their users several merchandises which were bought by people who share similar purchasing behaviors with them. Our recommendation system will focus on finding the relationship between certain customer groups and merchandises through mining the massive dataset. We will also accomplish parts of functions of the recommendation system which are implemented by modern famous E-business companies.

# Techniques and Tools

**Python language, Mysql and etc.**

**Hadoop, Map-Reduce, Stream mining, Collaborative filtering, Cluster analysis, Valuation and etc.**

# Input and output

**Input**:

1. merchandises’ information, including name, brand, prize, variety and so on.

2. transaction information, including customer’s name, goods’ name, quantity, date and so on.

**Output**:

The whole system, including interrelation graph of merchandises, similar merchandises recommended, bundle offers, preference information and so on.

Objective1: The System is able to recommend merchandises which are related to user’s purchasing behaviors, to users, with analysis upon their purchasing behaviors, transaction records and social behaviors.

Objetive2：Generate straightforward graph and table to show the result of our analysis. Exhibit the partition of various user and merchandises based on the cluster analysis of their features.

Objective3: The system is durable while dealing with massive dataset (at the level of 10 gigabytes) and offer approximately precise recommendation.

Objective4: Completely practice making use of the distributed computing that we have learnt in class to handle and analyze the interrelations of the data.

# Timeline

|  |  |
| --- | --- |
| Item | period |
| Fundamental learning and revise proposal | Oct.1st - Oct.9th |
| System structure and algorithm design | Oct.10th – Oct.15th |
| Data collection | Oct.16th – Oct.19th |
| Main codes implementation | Oct.20th – Nov.5th |
| Testing and report | Nov.6th – Nov.8th |
| Bug fixing and performance optimization | Nov.9th – Nov.21st |
| Final report writing | Nov.22nd – Dec. 2nd |