CMSC 5741 project proposal

Proposed Project Title:

Ranking and Prediction of Amazon Fine Food based on Costumer's rating and review

Group ID	Student ID	Name	Email address
8	1155106860	Ng Wing Hin	1155106860@link.cuhk.edu.hk
8	1155002533	Lei Wen Feng	1155002533@link.cuhk.edu.hk
8	1155106571	Cheng Tin Chu	1155106571@link.cuhk.edu.hk
8	1155103957	Chan Hei Nam	1155103957@link.cuhk.edu.hk

1 Motivation:

Currently, we are all living in a social system under capitalism. Every enterprise is competing with each other with their own products, services and even user experience. With the rapid growth of the internet in recent year, user experience can be further improved via processing huge data on costumer's review and rating. The Large internet-based retailer, like Amazon, have an enormous number of products but obviously, not all of them are popular. Therefore, processing costumer's review and rating of a product is vital for improving user experience thus increasing profit.

In order to determine the popularity of a product, an explicit way is to process costumer's review and rating. Nonetheless, the number of responses on a product would not always be sufficiently large enough for reference. In other words, we have a sparse dataset with users versus rating. With the help of matrix completion helps to extend the sparse data and return an estimated complete data matrix for further processing. In the Amazon Fine

Food Reviews, there are user's scores on different products and our goal is to investigate the popularity of products. This will benefit online retailer to promote further actions to enhance user experience and yield more revenue. That is to say to shine a spotlight on popular items under "Recommended Items" to drive traffic; or to group less popular products for sale and promotions.

2 Topics related, Algorithm and deliverables:

This project is related to the following topics in descending order of level of relation:

- Probabilistic Matrix Factorization for analysis product's popularity
- Collaborative filtering
- Frequent itemset on reviews based on frequently used important word
- MapReduce for pre-processing the data into format we want

At the end of the project, we are expected to come up with the most and least popularity set of products. The result could also be used to predict an product popularity based on rating or reviews.

3 Dataset and demonstration method:

The tentative dataset is the **Amazon Fine Food Reviews**[1] data from Standford Network Analysis Project with 568,454 food reviews from 256,059 Amazon users on 74,258 products during Oct 1999 to Oct 2012.

The following is the basic statistics of this dataset:

Table 1: Table of general statistic

	Score	HelpfulnessNumerator	HelpfulnessDenominator
Mean	4.18	1.74	2.23
Std	1.31	7.64	8.29
Min	1	0	0
First Quartile	4	0	0
Median	5	0	0
Third Quartile	5	0	0
Max	5	866	923

Table 2: Table of Counting Unique

Count Unique	
Review Id	568454
product Id	74258
User Id	256059
HelpfulnessNumerator	231
HelpfulnessDenominator Id	234
Rating	5
Time	3168
Summary	295743
Review	393579

Table 3: Table based on Product

Number of Review of each ProductId	
Count	74258
Mean	7.66
Std	26.45
Min	1
First Quartile	1
Median	2
Second Quartile	5
Max	913

Table 4: Table based on User

Number of Review of each UserId	
Count	256059
Mean	2.22
Std	4.44
Min	1
First Quartile	1
Median	1
Second Quartile	2
Max	448

We would first be using MapReduce for processing all the data into a matrix. Idea behind is to divide the whole table into section for different Mappers, producing pair with key(userId, productId) and value(Score, Summary, Tex, HelpfulnessNumerator, HelpfulnessDenominator) and Reducers put them into a matrix.

Thus, we would be using matrix factorization techniques to perform Probabilistic Matrix Factorization and followed by optimization steps. To demonstrate our analysis result, the program would be able to suggest top 5+ popular items and some less popular items during the period when the data was recorded. Furthermore, with the result, we can do prediction on the popularity of a product based on user's review and rating.

4 Related work:

Building a prediction model by Guillaume Payen from Kaggle[2]

5 Timeline of project milestone:

The following is the rough project timeline :

Week	Work to be done
1	Designing the semantic of the program and implementation.
2	Implementation period.
3	Implementation, validation and testing period.
4	Presentation Slide did by latex and ready for demonstration.
5	Finalising project.

References

[1] J. McAuley and J. Leskovec, *From amateurs to connoisseurs:modelling the evolution of user expertise through online reviews*, WWW, 2013.

[2] Guillaume Payen, *Building a prediction model*, Kaggle, 2015.