**Location-based Collaborative Filtering**

**and Frequent Itemset**

**A Restaurant Recommendation System for Yelp**

Thanh Tung Nguyen (ID: 40042891)

Huy Nguyen (ID: 40023289)

*Date: April 12, 2019*

# Abstract

|  |
| --- |
| Before going out for a meal, Yelp has been one of the most popular choice for customers to check for restaurants quality. To help users make better choices, we use techniques and principles of recommendation systems to create an application which makes predictions based on the user similarities. Using Yelp’s dataset, we develop an enhanced collaborative filtering using location as a key criterion for generating recommendations and then extract collaborative and content-based features to identify customer and restaurant profiles. Besides, we also provide frequent itemset references to users subsequently based on their chosen restaurants. We would evaluate our algorithm using Root metrics Mean Squared Error and Mean Absolute Error, we then evaluate and compare the algorithms. Due to limitation of time and resources, our scope of work will be narrowed to businesses within Canada. |

## Keywords

collaborative filtering, frequent itemset, recommendation, location-based, location, Yelp dataset.

# Introduction

A vast database of reviews, ratings, and general information provided by the community about businesses, Yelp provides consumers with a myriad of options and information even when searching for an especially specific service or goods niche. However, although all required information may be present to make an informed choice, it is often still difficult by just looking at the raw data. Reading all the reviews of a single business alone is time consuming and requires more effort than the average user is willing to expend. As a result, we believe users could greatly benefit from a recommendation system

## Context

Headings of subsections should be in Georgia 11-point bold italics with initial letters capitalized (Heading 2). (Note: for sub-sections and sub-subsections, words like ‘the’, ‘of’, ‘a’, ‘an’ are not capitalized unless it is the first word of the heading.)

### Objectives

Headings for sub-subsections should be in Georgia 10-point bold with initial letters capitalized (Heading 3). Please do not go any further into another layer/level.

## Problem to Solve

Headings of subsections should be in Georgia 11-point bold italics with initial letters capitalized (Heading 2). (Note: for sub-sections and sub-subsections, words like ‘the’, ‘of’, ‘a’, ‘an’ are not capitalized unless it is the first word of the heading.)

### Related Work

Recommendation systems facilitate users to make better choices while dealing with huge amount of data by recommending to them items that they would like.

There has been a lot of work done on designing recommender systems during the last two decades. Amazon.com [3] and Netflix [4] are two popular applications of recommender systems. [5] Presents an online social network-based recommender system that extracts user‟s interests for jobs and then makes recommendations to them accordingly. It is focused on two very popular social networks Facebook and LinkedIn. [6] Implements Naive Bayes to retrieve hidden data from stored database and compares the user values with trained data set. Then mapping of patient‟s attributes with stored database entries is done and probabilistic values are analysed for decision making. It can answer complex queries for diagnosing heart disease and thus assist healthcare practitioners to make intelligent clinical decisions which traditional decision support systems cannot. Sentiment analysis or opinion mining, an imperative research area of natural language processing, involves the extraction and identification of the attitude of a speaker or writer about a certain subject matter [7]. Opinion is generally combination of words, sentences, or documents. Opinion mining is based on the reviews of the other users. Sentiment analysis is used to classify each opinion as positive or negative.[8] Research paper proposed a novel document quality classification approach, which extracts sentiment value from SentiWordNet and accumulates the different sentimental influence of each word based on a document level. According to the experimental results, this proposed approach, which extract sentimental knowledge from SentiWordNet, outperform the approach in which SentiWordNet is not used for all categories with an exception, which is spam category. [9] Proposed system uses SentiWordNet library. The data from the reviews first removing stop words, then stemming by Porter Stemmer algorithm and then that reviews are tagged by their respective parts of speech. Then the score of review is calculated by pair of part of speech and rank in SentiWordNet.

# Materials and Methods

## Materials

Headings of subsections should be in Georgia 11-point bold italics with initial letters capitalized (Heading 2). (Note: for sub-sections and sub-subsections, words like ‘the’, ‘of’, ‘a’, ‘an’ are not capitalized unless it is the first word of the heading.)

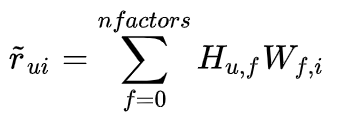
### Sub-subsections

Headings for sub-subsections should be in Georgia 10-point bold with initial letters capitalized (Heading 3). Please do not go any further into another layer/level.

## Methods

### Alternating least Square Collaborate-filtering

Collaborative filtering is a method to make predictions for a user based on his or her preferences already exist in the database. There are different approaches in collaborate-filtering, for example user-based, item-based or matrix factorization approach. Matrix factorization is a new solution comparing to the other two approaches due to its ability to solve sparse data problem, which is made famous after the Netflix Prize Challenge. In linear algebra, a matrix factorization can be defined as a factorization of a matrix into a product of matrices. As for collaborative filtering, matrix factorization algorithms applied by decomposing user-item predicted rating interaction matrix into a product of two lower dimensionality rectangular matrices, namely as user matrix and item matrix. While the row represents user and the column are latent factors in user matrix, it is the opposite role for item matrix. A matrix factorization matrix used to predict rating r of user u for item i can be presented as a formula as follow:



where H is the user matrix and W is the item matrix.

In Spark, Alternating Least Square (ALS) is also a matrix factorization algorithm that runs in a parallel fashion. ALS is implemented in Apache Spark Machine Learning (ML) library and built for a large-scale collaborative filtering problems

### Frequent pattern algorithm

Frequent pattern (FP) growth algorithm is one of many ways to do frequent itemset without using candidate generations. The algorithm will firstly compress the input database before creating an FP-tree instance representing frequent items. Then, the compressed database is divided into a set of conditional databases where each one is associated with one frequent pattern. Finally, each of those conditional database will be mined separately. A unique important feature of FP Growth is that it computes short patterns recursively and separately before adding all of them up into one long frequent patterns. Using this strategy, FP-Growth algorithm manages to reduce search costs by searching for frequent pattern.

# Results

To test our results, we follow cross validation when checking our performance with several evaluation metrics. Specifically, we break the Yelp dataset into two chunks of 80% and 20% each. The first set of 80% is used to train the system and the second set of 20% issued to test the system. For evaluation metrics, popular metrics are used to measure the performance of recommendation systems where a user gives an item a rating are Root Mean Squared Error (RMSE) and Mean Absolute Error (MAE)

|  |  |  |
| --- | --- | --- |
| **when distance = 3km and rank = 70** | | |
|  | RMSE | MAE |
| basic als recommender | 2.127391175 | 1.722178277 |
| global average recommende | 1.435300565 | 1.242512298 |
| als with bias recommender | 3.042489076 | 2.554701563 |

|  |  |  |
| --- | --- | --- |
| **when distance = 5km and rank = 70** | | |
|  | RMSE | MAE |
| basic als recommender | 2.147847095 | 1.717569799 |
| global average recommende | 1.435300565 | 1.242512298 |

|  |  |  |
| --- | --- | --- |
| **when distance = 10km and rank = 70** | | |
|  | RMSE | MAE |
| basic als recommender | 1.852597978 | 1.460206047 |
| global average recommende | 1.329057787 | 1.112019584 |

## Data explanation

After observing the result, it is clear that something went wrong in either the algorithm or the dataset we used. After careful review of the algorithm with different dataset, we could find nothing that is responsible for the result. Hence, we start to suspect our dataset suffer from data sparseness problem. Nonetheless, we proceed to continue trying FP-growth at min support and confidence at 0.5. Unexpectedly, the algorithm returned empty df. Hence, we keep on lowering min s and min c until 0.1. At this point, our suspicion about data sparseness increase sharply. So we try lower to 0. Unfortunately, after waiting for 15 minutes with my CPU working at 100% most of the time, it still does not give back any result, so I stop it. From this, we tried to look for others who work on the same dataset who also use collaborate filtering like us and we found one. Perhaps lucky for us, the paper from Stanford University has the same conclusion as us that the Yelp dataset suffers from data sparseness.

### Sub-subsections

Headings for sub-subsections should be in Georgia 10-point bold with initial letters capitalized (Heading 3). Please do not go any further into another layer/level.

# Discussion

Each type of submission (completed research papers, research-in-progress papers, and panels) has specific page length requirements. See additional requirements specific to each type of submission. Any submission that exceeds page length limits will be rejected without review. Paper length limitations are intended to encourage authors to publish full-length papers in journals or other outlets at a later date.

## Solutions to the issues found

Headings of subsections should be in Georgia 11-point bold italics with initial letters capitalized (Heading 2). (Note: for sub-sections and sub-subsections, words like ‘the’, ‘of’, ‘a’, ‘an’ are not capitalized unless it is the first word of the heading.)

### Sub-subsections

Headings for sub-subsections should be in Georgia 10-point bold with initial letters capitalized (Heading 3). Please do not go any further into another layer/level.

## Limitations of the methods used

Headings of subsections should be in Georgia 11-point bold italics with initial letters capitalized (Heading 2). (Note: for sub-sections and sub-subsections, words like ‘the’, ‘of’, ‘a’, ‘an’ are not capitalized unless it is the first word of the heading.)

### Sub-subsections

Headings for sub-subsections should be in Georgia 10-point bold with initial letters capitalized (Heading 3). Please do not go any further into another layer/level.

## Possible future work

Headings of subsections should be in Georgia 11-point bold italics with initial letters capitalized (Heading 2). (Note: for sub-sections and sub-subsections, words like ‘the’, ‘of’, ‘a’, ‘an’ are not capitalized unless it is the first word of the heading.)

### Sub-subsections

Headings for sub-subsections should be in Georgia 10-point bold with initial letters capitalized (Heading 3). Please do not go any further into another layer/level.

# Conclusion

It is important that you write for a general audience. It is also important that your work is presented in a professional fashion. This guideline is intended to help you achieve that goal. By adhering to the guideline, you also help the conference organizers tremendously in reducing our workload and ensuring impressive presentation of your conference paper. We thank you very much for your cooperation and look forward to receiving your nice looking, camera-ready version!

# REFERENCES (Ensure that all references are fully complete and accurate as per the examples)

|  |  |
| --- | --- |
| [1] | Schwab, K. "The Fourth Industrial Revolution"; Penguin: London, UK, 2017. |
| [2] | James Zou and Londa Schiebinger, "Design AI so that it's fair"; Nature, Vol 559, 19 Jul 2018. |
| [3] | Sam Shead, "Google DeepMind's Go AI Has Mastered Chess And Shogi"; Forbes [Online], 7 Dec 2018, https://www.forbes.com/sites/samshead/2018/12/07/google-deepminds-go-ai-has-mastered-chess-and-shog |
| [4] | Virginia Eubanks, "Automating inequality"; St. Martin's Press, New York, US, 2018. |
| [5] | Jeremy Hsu, "Humanity Must 'Jail' Dangerous AI to Avoid Doom", Live Science [Online], 1 Mar 2012, https://www.livescience.com/18760-humans-build-virtual-prison-dangerous-ai-expert.html |