**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 choices 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

As a very popular platform for choosing restaurants, Yelp contains an immense database of reviews, ratings, and other information provided by the community about businesses. Yelp’s restaurant ratings are being becoming a key performance indicator of to access whether a restaurant is successful and popular. However, reading all the reviews and ratings of a single business and compare them with others could take so much time from users. Therefore, we decided to build a recommendation system based on restaurant ratings which could greatly benefit Yelp users and food lovers by removing all the challenges processing a huge amount of data to make an informed choice.

Recommendation systems have been widely implemented in e-commerce websites to recommend and provide customers with suggestive information helping them decide which products to buy. For example, media-services providers, such as Netflix, iTunes and Spotify, utilize recommendation systems to suggest songs, videos, movies to users based on their previous choices and taste. Given this general theme, techniques and principles of recommendation systems as well as taking users’ locations into consideration, we shall create a simple restaurant recommendation app to help Yelp users make quick and better choices

### Related Work

Thanks to the vast amount of information contained in the Yelp dataset, numerous past projects and researches and tried to use it to give food and restaurant recommendations to users. For example, Sumedh Sawant and Gina Pai [1] used the network-based-inference collaborative filtering algorithm to develop a recommendation system on the Yelp Dataset Challenge dataset. Sawant also had another Collaborative Filtering recommendation system project on Yelp dataset using Weighted BiPartite Graph Projection [2]. In addition, with regards to using location for personalized POI recommendations in mobile environments, we found a research paper of done by Tzvetan Horozov, Nitya Narasimhan, Venu Vasudevan [3] proposing GeoWhiz app - a real-world deployment of our restaurant recommender system for location-based points of interest (POI). In this project, we apply collaborative filtering and frequent itemset recommendation methods on restaurants ratings and aims to work on a location-based analysis instead of a nationwide user-based analysis in order to provide suggestions to Yelp restaurants.

# Materials and Methods

## Materials

Our primary dataset is the Yelp's businesses retrieved from [www.kaggle.com/yelp-dataset/yelp-dataset](http://www.kaggle.com/yelp-dataset/yelp-dataset) that contains actual business, user, and users’ review data from North America region. From the original dataset, we extracted business information located in Canada as our predefined scope of work. The Canadian businesses were found using postal codes. The Canadian postal code listing is obtained from Google Fusion Tables at <https://fusiontables.google.com/>. Some basic characteristics of the dataset are summarized as below:

|  |  |
| --- | --- |
| **Original Dataset** | **Number of entries** |
| * Number of businesses | 192,609 |
| * Number of reviews | 6,685,900 |
| * Number of users | 1,637,138 |
| **Canada** |  |
| * Number of Canadian businesses | 50,644 |
| * Number of Canadian reviews. | 1,063,142 |
| **Canadian Postal Codes** |  |
| * Number of postal codes | 889,320 |

For this project, we combined the review, business and user datasets and picked diverse feature set for developing the machine learning model. The structure of the different datasets was available in individual json files. We co­related the data in the files, joined, denormalized them and extracted the desired features.

## Methodology

The theory and formal definitions of our methods for this project are presented in this section before we present our app algorithm in the next section.

### Collaborative Filtering

Collaborative filtering is a method of making automatic predictions (filtering) about the interests of a user by collecting preferences or taste information from many users (collaborating).

For our project, we want to use matrix factorization, a more sophisticated machine learning technique used in recommender system, in collaborative filtering. We will discuss what is matrix factorization and how is it implemented in Spark.

In collaborative filtering, matrix factorization is the leading-edge solution for sparse data problem. Matrix factorization is a factorization of a matrix into a product of matrices. In the case of collaborative filtering, matrix factorization algorithms work by decomposing the user-item interaction matrix into the product of two lower dimensionality rectangular matrices. One matrix can be seen as the user matrix where rows represent users and columns are latent factors. The other matrix is the item matrix where rows are latent factors and columns represent items.

In the sparse user-item interaction matrix, the predicted rating user u will give item i is computed as:

### Frequent Itemset

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.

## Algorithm

### 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.

### 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.

# Results

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.

* Complete/full papers: Must not exceed 10 pages all-inclusive (approx. 5,000 words, including author names, abstract, figures, tables, references, appendices).
* Emergent Research Forum (ERF) papers: Must not exceed 5 pages all-inclusive (approx. 2,500 words, including author names, abstract, figures, tables, references, appendices).

## Subsections

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.

# 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.

## Subsections

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.

**Title**

Your paper’s title should be in Georgia 20-point bold. Ensure proper capitalization within your title (i.e. “The Next Frontier of Information Systems” versus “the next frontier of Information systems.”

## Normal or Body Text

Please use a 10-point Georgia font (similar to Times New Roman, but more easily read online) or, if it is unavailable, another proportional font with serifs. The Georgia font is also available on Macintosh. Please use sans-serif or non-proportional fonts only for special purposes, such as source code text (SpecialStyle). [References to Georgia font from this point forward should be interpreted as “Georgia or equivalent.”]

# Sections

The heading of a section should be Georgia 13-point bold, left justified (Heading 1 Style in this template file). Sections should not be numbered.

## Subsections

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.

# Figures, Tables & Captions

Place figures and tables close to the relevant text (or where they are referenced in the text).

Captions should be Georgia 10-point bold (Caption Style in this template file). They should be numbered (e.g., “Table 1” or “Figure 2”), centered and placed beneath the figure or table. Please note that the words “Figure” and “Table” should be spelled out (e.g., “Figure” rather than “Fig.”) wherever they occur. The proceedings will be made available online, thus color figures are possible.

## Inserting Images

Occasionally MS Word generates larger-than-necessary PDF files when images inserted into the document are manipulated in MS Word. To minimize this problem, use an image editing tool to resize the image at the appropriate printing resolution (usually 300 dpi), and then insert the image into Word using Insert | Picture | From File...

Using tables to hold places can work very well in Word. If you want to copy a figure from another application (such as PowerPoint) and then paste to the place where you want your figure to be, make sure that (1) the figure stays in the position, and (2) it does not take up too much space. You can ensure the former by double clicking the figure, then go to “Layout” tab, and select “In line with text.” To ensure the latter, use “Paste Special,” then select “Picture.” You can resize the figure to your desired size once it is pasted.

## Table Style

Inserting a table in the text can work well. See Table 1 below. If you do not use this style, then you may want to adjust the vertical spacing of the text in the tables. (In Word, use Format | Paragraph… and then the Line and Page Breaks tab. Generally, text in each field of a table will look better if it has equal amounts of spacing above and below it, as in Table 1.)

|  |  |  |
| --- | --- | --- |
|  | Treatment 1 | Treatment 2 |
| Setting A | 125 | 95 |
| Setting B | 85 | 102 |
| Setting C | 98 | 85 |
| Table 1. A Very Nice Table | | |

# Language, Style, and Content

With regard to spelling and punctuation, you may use any dialect of English (e.g., British, Canadian, US, etc.) provided this is done consistently. Hyphenation is optional. To ensure suitability for an international audience, please pay attention to the following:

* Write in a straightforward style.
* Try to avoid long or complex sentence structures.
* Briefly define or explain all technical terms that may be unfamiliar to readers.
* Explain all acronyms the first time they are used in your text – e.g., “Digital Library (DL)”.
* Explain local references (e.g., not everyone knows all city names in a particular country).
* Be careful with the use of gender-specific pronouns (*he*, *she*) and other gendered words (*chairman*, *manpower*, *man-months*). Use inclusive language that is gender-neutral (e.g., *they*, *s/he*, *chair*, *staff*, *staff-hours*, *person-years*).

# 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!

# Acknowledgements (Optional)

Please do *not* add acknowledgements to your original submission because it may identify authors. Add any acknowledgements to the revised, camera-ready version of your paper.

## References and Citations

References are to be formatted using the **new** *MIS Quarterly* style (<http://www.misq.org/manuscript-guidelines> under MISQ References Format). References must be complete, i.e., include, as appropriate, volume, number, month, publisher, city and state, editors, last name & initials of all authors, page numbers, etc. If you use EndNote, be aware that different versions of the software change the styles, creating some inconsistencies. Please also be aware that the *MIS Quarterly* style provided by default in the EndNote X4 software (and all earlier versions) is for the **old** *MIS Quarterly* style. You may download the Endnote Style [here](http://ais.site-ym.com/resource/resmgr/Files/MISQ-Revised.zip).

Your references should comprise only published materials accessible to the public. Proprietary information may not be cited.

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

Ackoff, R. L. 1961. "Management Misinformation Systems," *Management Science* (14:4), pp. 147-156.

Benbasat, I., and Zmud, R. W. 2003. “The Identity Crisis within the IS Discipline: Defining and Communicating the Discipline’s Core Properties,” *MIS Quarterly* (27:2), pp. 183-194.

Bonini, C. P. 1963. *Simulation of Information and Decision Systems in the Firm*, Englewood Cliffs, NJ: Prentice-Hall.

Broadbent, M., Weill, P., O’Brien, T., and Neo, B. S. 1996. “Firm Context and Patterns of IT Infrastructure Capability,” in *Proceedings of the 14th International Conference on Information Systems*, J. I. DeGross, S. Jarvenpaa, and A. Srinivasan (eds.), Cleveland, OH, pp. 174-194.

Carroll, J. 2005. “The Blacksburgh Electronic Village: A Study in Community Computing,” in *Digitial Cities III: Information Technologies for Social Capital*, P. van den Besselaar and S. Kiozumi (eds.), New York: Springer-Verlag, pp. 43-65.