Restaurant Recommendation System[1]

Gomudurai, Tharun Niranjan Pandian
Master of Science in Software Engineering, ASU
tgomudur@asu.edu

Hussain, Yousuf Syed Mohammad
Master of Science in Software Engineering, ASU
syousufh@asu.edu

Iyer, Vignesh
Master of Computer Science, ASU
vkiyer@asu.edu

Manikarnike, Mukund
Master of Computer Science, ASU
mmanikar@asu.edu

Singh, Bharat
Master of Science in Software Engineering, ASU
bsingh21@asu.edu

Venkatesh, Vaibhav Deshu
Master of Science in Computer Engineering, Arizona ASU
vdvenkat@asu.edu

ABSTRACT

As the number of users interacting with web based applications increase, the number of such applications increase and there is going to be a constant demand for fine-tuning how content is served up to each user with the aim of maximizing user satisfaction. This paper illustrates how such custom content can be served up to users by using a recommendation system. Certain techniques like collaborative filtering, content based recommendation are explored and their performance is compared by using measures like root mean square error and mean absolute error. The recommendation system uses restaurant data from Yelp provided by Yelp as part of Round 7 of the dataset challenge.

Keywords

Recommendation, Collaborative, Content-based, Yelp-Data

1. INTRODUCTION

In the past decade and a half, the internet has exponentially grown, resulting in a large number of users and large amount of content generated by every user alone and also through inter-user interactions. As is the trend now, users look for a customized set of results to be output of each search query that they provide. Given the amount of content available and how it is predicted to grow, this would very soon become a common requirement on the internet. The sole purpose of a recommendation system is to customize the content served to every user based on the requirements of each user. The following subsections talk about the dataset used for the purpose of recommendation and the organization of further sections of the paper.

Yelp Dataset Challenge [1]

Yelp collects a large amount of data for businesses that are registered on the platform. The data collected includes the kind of business run by each organization, the amount of check-ins at the particular place, number of reviews for the same, the user interest in the business and more such information. The Yelp dataset challenge has been ongoing since March 2013 in several rounds. The data used for this project is that made available for Round 7 of the challenge which started in February 2016 and goes on until June 2016. Some of the topics that the challenge hopes to be addressed through submissions are estimation of culture trends, location mining and urban planning, seasonal trends and many more such aspects. This project chooses to implement a restaurant recommendation system using the provided dataset.

Yelp Dataset

The dataset consists of several components, which are shown in **Table 1**. Each of these components have several attributes that describe them.

Component	Record Size	Size on Disk (GB)
Business	206,534	0.064
Review	8,102,234	18
User	552,340	0.22
Check-in	55,570	0.024
Tip	606,820	0.11
Photos	Not evaluated because it isn’t used.	

Table 1 Dataset Description

[1] Source Code for the project is available at github.com/bks2009/RecommendationSystem.

Data items in each of these components cover data coming from all regions in the world. The following few subsections describe the structure of each of these components and how many data items each of these components contain. Each of these components are provided in JSON format. The data provided as part of Photos isn’t used for recommendation purposes as part of this paper.

Business

This JSON contains information about businesses like their primary identification, the neighborhood they’re in, the category they fall under, the average star rating given to the business by users and many more. The data item takes the format as shown in **Table 2**.

```
 {
     'type': 'business',
     'business_id': (encrypted id),
     'name': (business name),
     'neighborhoods': [(hood names)],
     'full_address': (localized address),
     'city': (city),
     'state': (state),
     'latitude': latitude,
     'longitude': longitude,
     'stars': (star rating, round to 0.5),
     'review_count': review count,
     'categories': [(category names)]
     'open': True / False (Closed/Open),
     'hours': {
         (day_of_week): {
             'open': (HH:MM),
             'close': (HH:MM)
         },
         ...
     },
     'attributes': {
         (attr_name): (attr_value),
         ...
     },
 }
```

Table 2 Business JSON

Review

This JSON contains information about every review that every user has given every business that is on yelp. In addition to text reviews, it also contains other details like the star rating, number of votes for the particular review and so on. The data takes the format shown in **Table 3**

```
 {
     'type': 'review',
     'business_id': (encrypted id),
     'user_id': (encrypted user id),
     'stars': (star rating, round to 0.5),
     'text': (review text),
     'date': (date, like '2012-03-14'),
     'votes': {(vote type): (count)},
 }
```

Table 3 Review JSON

User

This JSON contains information about every user including the user ID, the number of reviews contributed by the user and many more details. The data item takes the format as shown in **Table 4**

```
 {
     'type': 'user',
     'user_id': (encrypted user id),
     'name': (first name),
     'review_count': (review count),
     'average_stars': (floating point),
     'votes': {(vote type): (count)},
     'friends': [(friend user_ids)],
     'elite': [(years_elite)],
     'yelping_since': (date, '2012-03'),
     'compliments': {
         (compliment_type):
 (num_compliments_of_this_type),
         ...
     },
     'fans': (num_fans),
 }
```

Table 4 User JSON

Check-in

This JSON contains information about the check-ins made on Yelp to every business, the time at which it is done and more such information. The data item takes the form as shown in **Table 5**

```
 {
     'type': 'checkin',
     'business_id': (encrypted id),
     'checkin_info': {
         '0-0': (number of checkins from
 00:00 to 01:00 on all Sundays),
        '1-0': (number of checkins from
01:00 to 02:00 on all Sundays),
        ...
        '14-4': (number of checkins from
14:00 to 15:00 on all Thursdays),
        ...
        '23-6': (number of checkins from
23:00 to 00:00 on all Saturdays)
    }, # if there was no checkin for a
hour-day block it will not be in the dict
}
```

Table 5 Check-in JSON

Tip

Yelp contains a feature where users can provide tips to similar users based on what they're looking for. This JSON contains information about each tip that every user has provided. The data item is as shown in **Table 6**

```
{
    'type': 'tip',
    'text': (tip text),
    'business_id': (encrypted id),
    'user_id': (encrypted user id),
    'date': (date, like '2012-03-14'),
    'likes': (count),
}
```

Table 6 Tip JSON

Paper Organization

The following sections of this paper are organized as follows

- Section 2 provides a description of the approaches used to build a recommendation system and the specifics of the implementation techniques used. It also describes how a subset of the dataset was chosen from the large dataset provided by Yelp.
- Section 3 presents certain theoretical analyses of all the algorithms and the results obtained from each of the methods used to perform recommendation algorithms
- Section 4 provides a few conclusive points on the project including strong, weak points of the approaches and key takeaways from the project.
- Section 5 acknowledges all the people and sources that helped in the betterment of the results obtained for the project.
- Section 6 provides a few pointers on possible work that can be done in this regard in the future.

2. DESCRIPTION

2.1 Problem Formulation

The problem formulation of a typical recommendation system is as shown in **Table 7**

	R_1	R_2	R_3	R_4
User 1	4.0	?	3.6	4.5
User 2	3.5	5.0	?	2.5
User 3	?	2.3	3.2	2.5

Table 7 Recommendation System Problem Formulation

The problem here is, given a set of ratings for each user for each restaurant and given that a user hasn't rated a few restaurants, to predict the possible ratings a user would give for a restaurant which he/she hasn't rated yet. By doing so, one would be able to find the possibility of recommending the particular restaurant to the user. This is a simple presentation of the recommendation problem and can become more and more complex as we add features through which each restaurant is described. This paper will describe all the approaches starting from the simple approaches to complex ones.

2.2 Recommendation algorithms

This section describes the recommendation algorithms used in this paper.

2.2.1 Benchmarking algorithm

In recommendation systems, there isn't any ground truth available to compare with and hence evaluation of the predictions becomes difficult and sometimes not possible. Hence, the approach used to solve this problem is to use a benchmarking algorithm. The usage of a benchmarking algorithm entails the following

1. Estimate error measures on the provided data using trivial algorithms by dividing the training set provided into a training set and a validation set.
2. These error measures can then be used as a benchmark against which all better algorithms are compared against which will provide a way to measure whether the recommendation system does a good job or not.

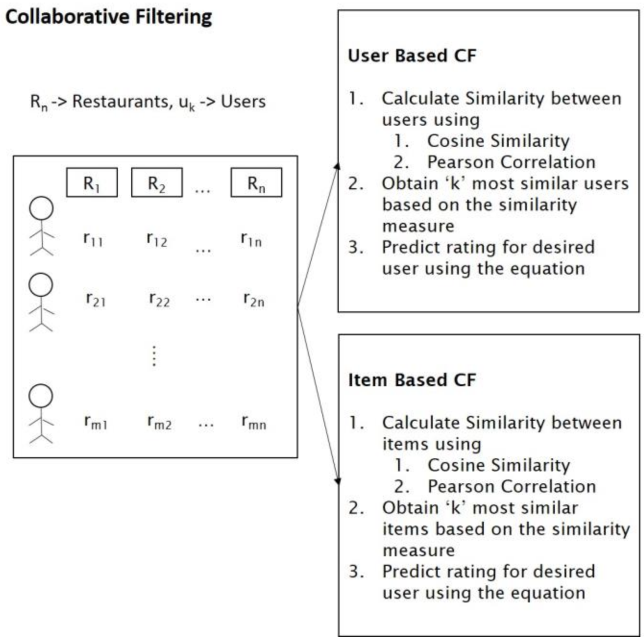
There are of course other ways to measure the success of a recommendation system like

1. User Satisfaction from the provided recommendations
2. User engagement with the content served up based on the recommendations provided.

However, the recommendation system designed for the purpose of this paper has a static dataset provided by Yelp and has no means to measure user satisfaction on the results provided. Hence, the benchmarking algorithm is employed. The technique used to perform benchmarking is the BASELINE approach used in [1].

2.2.2 Collaborative Filtering

The collaborative filtering approach tries to solve a typical recommendation system problem based on the ratings as shown in **Table 7**.



Collaborative Filtering

R_n -> Restaurants, u_k -> Users

	R_1	R_2	…	R_n
	r_{11}	r_{12}	…	r_{1n}
	r_{21}	r_{22}	…	r_{2n}
	⋮			
	r_{m1}	r_{m2}	…	r_{mn}

User Based CF

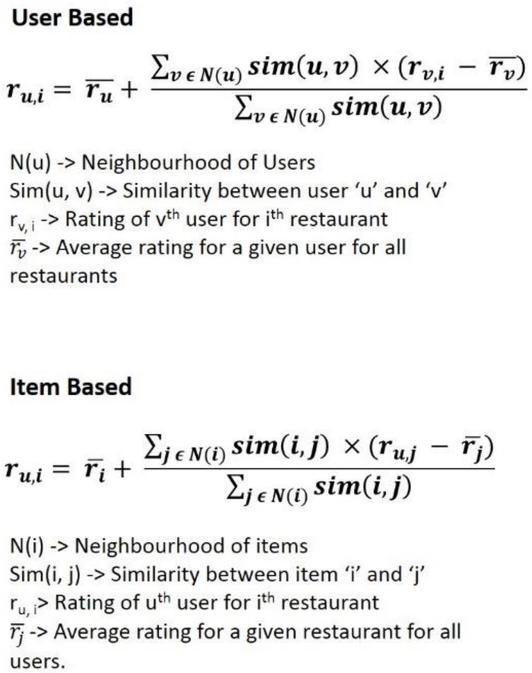
1. Calculate Similarity between users using
   1. Cosine Similarity
   2. Pearson Correlation
2. Obtain ‘k’ most similar users based on the similarity measure
3. Predict rating for desired user using the equation

Item Based CF

1. Calculate Similarity between items using
   1. Cosine Similarity
   2. Pearson Correlation
2. Obtain ‘k’ most similar items based on the similarity measure
3. Predict rating for desired user using the equation

Figure 1 Typical Collaborative Filtering System

Collaborative filtering can solve the recommendation problem using User-based or Item-based similarity approaches. **Figure 1** illustrates the approach taken to solve the recommendation problem using both approaches. **Figure 2** provides the prediction model that is used to predict the rating of an item for a particular user.



User Based

$$r_{u,i} = \overline{r_u} + \frac{\sum_{v \epsilon N(u)} sim(u,v) \times (r_{v,i} - \overline{r_v})}{\sum_{v \epsilon N(u)} sim(u,v)}$$

N(u) -> Neighbourhood of Users
Sim(u, v) -> Similarity between user ‘u’ and ‘v’
$r_{v,\, i}$ -> Rating of v^{th} user for i^{th} restaurant
$\overline{r_v}$ -> Average rating for a given user for all restaurants

Item Based

$$r_{u,i} = \overline{r_i} + \frac{\sum_{j \epsilon N(i)} sim(i,j) \times (r_{u,j} - \overline{r_j})}{\sum_{j \epsilon N(i)} sim(i,j)}$$

N(i) -> Neighbourhood of items
Sim(i, j) -> Similarity between item ‘i’ and ‘j’
$r_{u,\, i}$> Rating of u^{th} user for i^{th} restaurant
$\overline{r_j}$ -> Average rating for a given restaurant for all users.

Figure 2 Collaborative Filtering – Prediction Model

2.2.3 Content-based Filtering

<Content-based Recommendation Explanation>

2.3 Recommendation Lifecycle

This section includes a detailed description of all steps that were involved in solving the recommendation problem starting from data cleaning, going through implementation and

<Include the following information

- Cleaning data
- Feature selection
- Algorithm used
  - If required explanation of the algorithm using figures.
- Implementation specifications>

3. RESULTS

<Results here>

3.1 Theoretical Analysis

<Theoretical analysis if any

- Talk about evaluation techinques
- How the model should do given the data provided.>

3.2 Experimental Results

<Experiments performed.

- Include graphs for any kind of results generated
    - ○ Starting from baseline experiments carried out to end results.

>

4. CONCLUSION

<Address strong and weak points>

<Key Learnings from the project>

5. ACKNOWLEDGEMENTS

We would like to express our sincere thanks to Dr. Hanghang Tong for his efforts in discussing interesting topics in class that motivated us to do well in this project.

During the course, we also referred to several of Dr. Tom Mitchell's video lectures available on his CMU Website and the lectures of Dr. Andrew NG on Coursera which we found extremely useful and would like thank them for their insightful lectures which were truly thought provoking.

We would also like to thank the team at Yelp that made this project possible by making the dataset available publicly as part of a challenge.

6. FUTURE WORK

<Future Work here>

7. REFERENCES

[1] Sumedh Sawant, Gina Pai, Yelp Food Recommendation System, Stanford University

[2] Yelp Dataset Challenge

https://www.yelp.com/dataset_challenge

[3] Yelp Engineering Blog – Dataset Challenge http://engineeringblog.yelp.com/2013/10/yelp-dataset-challenge-winners-round-two-now-live.html

[4] Reza Zafarani, Mohammad Ali Abbasi, Huan Liu, Social Media Mining An Introduction