CSE 258: Web Mining and Recommender System

Your Personal Stylist

Shunsuke Iwata
Mao-Ho Wang
Yada Klueabvchit
Andrew Santos

ABSTRACT

Your Personal Stylist is a proposed solution that aims to improve customer shopping enthusiasm and online shopping experiences. This report details and explains the model used for recommending shop items that users have a high possibility to purchase and the comparison between several approaches to recommendation.

The experimental results show that by measuring user similarity, specifically between purchase and rating history, one is able to more precisely predict user ratings for items that have not yet been purchased by the user and coincidentally curate a list of (new) items that the user is highly likely to purchase and rate high.

Results from this study are not limited to the fashion industry alone. The models can be used in many other industries where user-item ratings are available and rating predictions could influence item recommendation to the users.

I. INTRODUCTION

Rent the Runway is one of the greatest disruptors in the trillion-dollar fashion company– offering rental services for designer clothes at affordable prices when historically these items were bought and sold at luxurious price points and owned-to-wear. The company offers subscription-based services to its customers, with the most popular plan enabling customers to wear and enjoy 8 items per month, amounting to $3,000 worth of luxury clothing for only $11 a piece.

This revolutionized the fashion industry by making these designer clothes much more accessible and much less of a commitment. By essentially expanding the market and reaching a new market of users who likely have less knowledge and experience with the brands and clothing offered, it is essential for a business like Rent the Runway to make use of sophisticated recommender systems to present users with items they would like so they continue their subscription and carry on with the service.

II. DATASET

The dataset used contains the measurements of clothing fit from RentTheRunway. The data is cited from the report called Decomposing fit semantics for product size recommendation in metric spaces by Rishabh Misra, Mengting Wan, Julian McAuley RecSys, 2018 and has the following feature vectors:

1) fit
2) user_id
3) bust size

4) item_id
5) weight
6) rating
7) rented for
8) review_text
9) body type
10) review_summary
11) category
12) height
13) size
14) age
15) review_date

This dataset contains a total of 192,544 example vectors and aside from body measurements such as bus size, height, and weight, it also includes the text feedback (review_text), review date from each item's purchase/rental, and finally the rating given to the item by the user on a scale of 1-10.

Data Pre-Processing and Exploratory Data Analysis

The approach we observed involved first pre-processing the data followed by exploratory data analysis to help pinpoint significant features to be later used for making predictions. As for these predictions, the researchers made use of different strategies including Pearson Similarity, Jaccard Similarity, Bag of Words (Cosine Similarity), and tf-idf (Cosine Similarity). The goal was to compare the similarity scores between the different approaches and finally recommend the best-performing approach to Rent the Runway.

We began pre-processing the data by tackling null values in the dataset. Several rows had null values in the rating column and such cannot be used for the model predictions. These null ratings were dropped upon discovery during EDA.

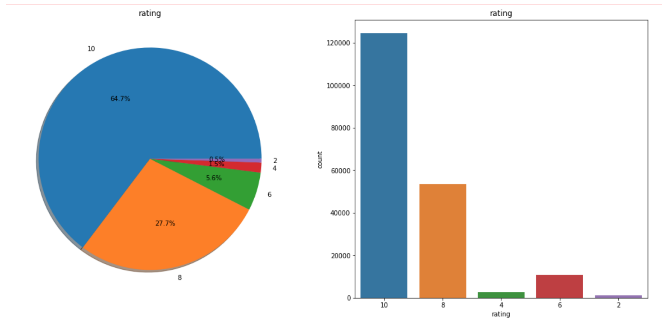


Figure 1A: Rating Distribution

Exploratory data analysis revealed that nearly 65% of items were rated a 10 by users and another large portion of around 28% were rated 8 [1A].

The following graph [1B] shows how many null values are contained in each variable

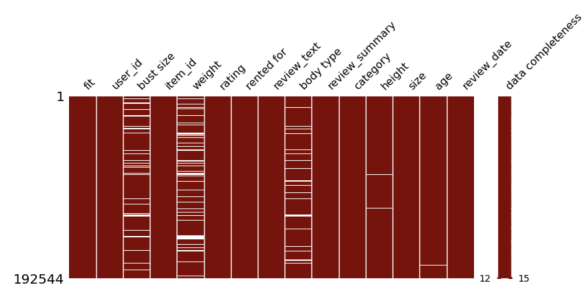


Figure 1B: Null Value Matrix

It was discovered that multiple columns contained NULLs, which had to be addressed in the data cleaning part of the process.

After removing null records and converting character variables to numbers, the following graph illustrates the correlation between the factor variables.

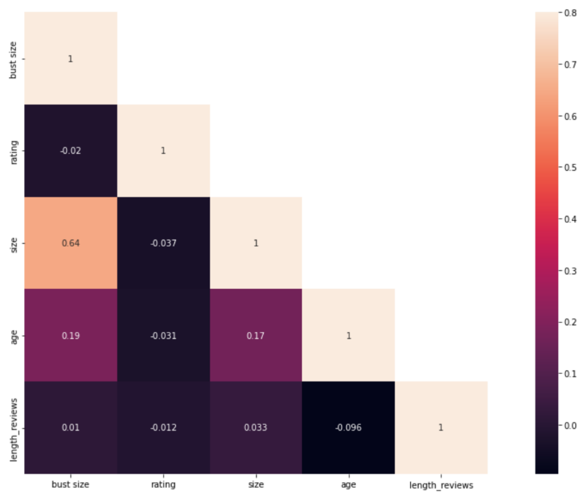


Figure 1C: Correlation Heatmap

A heatmap showing correlations [1C] reveals little correlation between rating and other numerical variables. Because of this, the researchers decided to focus primarily on similarity models based on purchase history alone along with text mining to predict ratings.

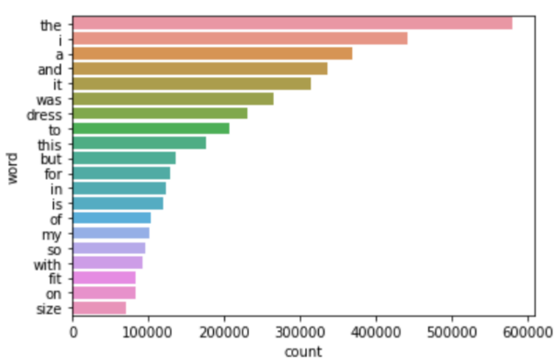


Figure 1D: Word Count Distribution

Diving more into sentiment analysis, the top ten most frequently used words in the reviews are articles and parts of speech. Standing out amongst these is the word dress, the seventh most common word, used 230,376 times– almost 40% of the most commonly used word, the [1D].

For more effective text mining, TF-IDF was performed to find influential words by ignoring commonly used English stop words. The results are displayed in the graph below [1E].

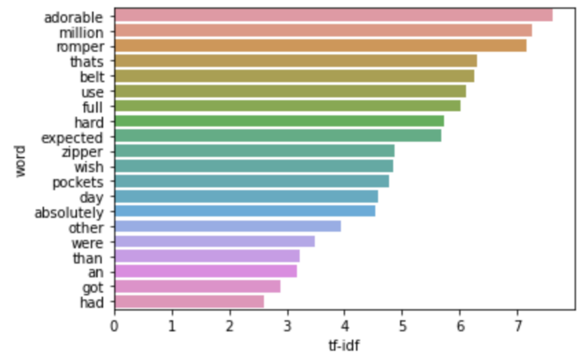


Figure 1E: TF-IDF Influential Words

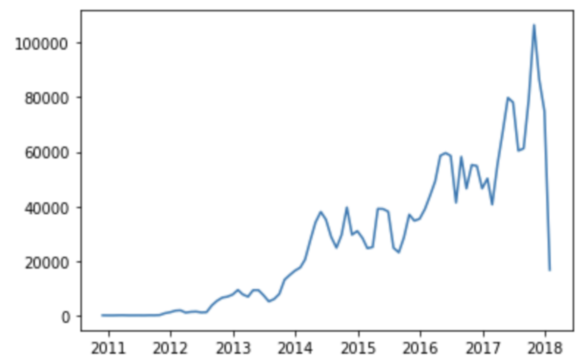


Figure 1F: Purchasing Over Time

To visualize the item purchase interaction over time, a time-series line graph was created and can be seen above. Business has been growing over the years and had really started to pick up in 2014.

III. MODEL - PREDICTIVE TASK

We aimed to compare the Mean Squared Error across different approaches in using similarity for predictions and to highlight the best performing (lowest MSE) model. Take note that we decided to assign null ratings with 0 as there are no zero ratings in the dataset, which is unlikely. Therefore, we decided to convert null ratings to zero. The following are the results of these.

Jaccard Similarity

$$\text{Jaccard}(i, j) = \frac{|U_i \cap U_j|}{|U_i \cup U_j|}.$$

After splitting the dataset into train and test sets, we created dictionaries for both users per item and items per user. Making use of Jaccard similarity across item reviews by user and the overall rating mean, prediction results returned an MSE of 2.381. Interestingly, the MSE of the rating mean as the prediction for all rows resulted in an even lower MSE of 2.109.

Pearson Correlation

items rated by both users average rating by u

$$\text{Sim}(u, v) = \frac{\sum_{i \in I_u \cap I_v} (R_{u,i} - \bar{R_u})(R_{v,i} - \bar{R_v})}{\sqrt{\sum_{i \in I_u \cap I_v} (R_{u,i} - \bar{R_u})^2 \sum_{i \in I_u \cap I_v} (R_{v,i} -}}$$

Using Pearson Correlation as we deal with numerical ratings, predictions resulted in a slightly better MSE of 2.087.

Bag of Words (Cosine)

$$Cos\theta = \frac{\vec{a} \cdot \vec{b}}{\|\vec{a}\| \ \|\vec{b}\|} = \frac{\sum_1^n a_i b_i}{\sqrt{\sum_1^n a_i^2} \sqrt{\sum_1^n b_i^2}}$$

Vector a and b are Bag of Words

By combining User's history reviews together, we calculate the similarity between each user's total reviews with Cosine Similarity. The predictions resulted in MSE of 2.073

TF-IDF (Cosine)

$$\mathit{tf}\text{-}\mathit{idf}(w, d, \mathcal{D}) = \mathit{tf}(w, d) \times \log_2 \left(\frac{|\mathcal{D}|}{1 + \mathit{df}(w, \mathcal{D})} \right)$$

$$\mathit{tf}'(w, d) = \delta(w \in d)$$

$$\mathit{tf}''(w, d) = \frac{\mathit{tf}(w, d)}{\max_{w' \in d} \mathit{tf}(w', d)}.$$

$$\text{idf}(q_i) = \log \left(\frac{|\mathcal{D}| - \mathit{df}(q_i, \mathcal{D}) + 0.5}{\mathit{df}(q_i, \mathcal{D}) + 0.5} + 1 \right).$$

By making use of TF-IDF we are able to find words that appear most frequently in the review_text field of the dataset and use this to determine the

Using TF-IDF with Cosine Similarity, predictions resulted in MSE using user similarity is 2.49, and item similarity is 2.43.

Item2Vec

$$\log p(w_o | w_i) \simeq \log \sigma(\gamma'_{w_o} \cdot \gamma_{w_i}) + \sum_{w \in \mathcal{N}} \log \sigma(-\gamma'_w \cdot \gamma_{w_i}).$$

Using Item2Vec with Skip-gram, predictions resulted in 2.47. Item2Vec performs embedding of items based on user activity history, which did not bring about accurate MSE. This indicated that a time series analysis of the items purchased by the users in this data set would not be informative

IV. RELATED LITERATURE

The dataset was taken from a study titled *Decomposing Fit Semantics for Product Size Recommendation*
in Metric Spaces by Rishabh Misra, Mengting Wan, and Julian McAuley. The goal of their research was to use customer fit feedback to identify the true fit of clothing. They made use of ordinal regression procedures to properly handle and classify labels and address the company's label imbalance issues.

Our case of using customer-item text reviews for rating prediction can be applied to other datasets aside from Rent the Runway. Virtually any dataset with user-item relationships, ratings, and text-review can be fitted into the model and can benefit from the predictions via recommender systems. This includes but is not limited to e-commerce services and other subscription services such as netflix/hulu.

State-of-the-art method Recommendations

Recommendation methods prior to deep learning include older models such as

- *Collaborative Filtering*
- *Matrix Factorization*
- *Singular Value Decomposition*
- *Factorization Machine*

RecSys 2019 and 2020 papers on research and proposals for fair benchmarking of recommender systems

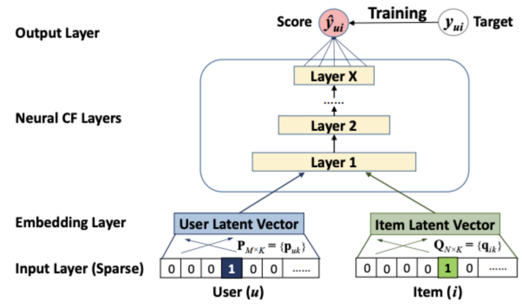
[Are We Really Making Much Progress? A Worrying Analysis of Recent Neural Recommendation Approaches]

[Are We Evaluating Rigorously? Benchmarking Recommendation for Reproducible Evaluation and Fair Comparison]

In the above paper, they tested the effectiveness and robustness (i.e., does it perform consistently across all data sets) against neural network-based recommendation methods that have become popular in recent years. The paper argued that many DNN-based recommendation methods cannot beat the older collaborative filtering and BPRFM methods, thus calling into question the DNN-based recommendation methods. However, some advanced approaches derived from traditional methods have been getting more promising results lately.

NCF（Neural Collaborative Filtering）

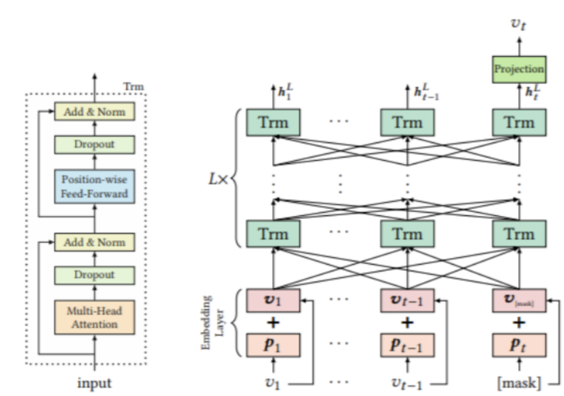
This is one of the advanced Matrix Factorization models. NCF is a nonlinear model that uses a multilayer perceptron to estimate the contents of a matrix from the combined user and item embeddings. It is trained by replacing the inner product representation part of Matrix Factorization with a neural network-based function.



NFC paper

BERT4Rec

This is one of the Sequential Recommendation models. In our research, we already adopted the Item2Vec model which is also a Sequential Recommendation. However, the following model saw a more accurate result. BERT4Rec's "BERT" stands for Bidirectional Encoder Representations from Transformers, a natural language processing model published in October 2018 in a paper by Google's Jacob Devlin et al. It has led to breakthroughs that have led to the saying "AI has surpassed humans," with the highest scores at the time across a wide variety of tasks. BERT4Rec is an application of this BERT model to recommendations.



(a) Transformer Layer. (b) BERT4Rec model architecture.

BERT4Rec paper

V. Results

Results of the study show that among the models listed above, the Bag of Words is the best performing model for the objective of using text-mining for rating prediction, with once again a MSE of 2.073. This finding indicates that users who write similar reviews also have similar preference for clothes. Thus, we can make future clothes recommendations to customers based on reviews they wrote before. It is interesting to note that the other models do not fall far behind in terms of performance and further testing and experimentation with other models mentioned as part of our recommendations could prove to yield even more accurate results.