Extracting Rating Dimensions from Hidden Topics in Text Reviews: a Better Recommendation System

Xinzhe Yang, Xuwen Shen

November 9, 2017

1 Research Context

The same rating from different users usually stands for different meanings. Individual users may assign different weights to such aspects when determining their overall score.For example, a spendthrift hotel reviewer might assign a low weight to "price" but a high weight to "service", thus explaining why their overall rating differs from a miserly reviewer who is only interested in price. There exists hidden information in reviews that leads to the final rating. By extracting the information, we can get where a restaurant shines and what it needs to improve. If we have k topics in reviews then we extract k (the same number as the number of topics in Yelp reviews) dimensions in rating for each topic respectively and then use the k dimensional rating to compute the recommendation score for an individual.

The importance of customer reviews has been proven in terms of giving insights to businesses and also providing customers personalized services. LDA has been widely used to analyze the latent meanings in reviews and study customer behaviors.

However, few of previous researches have been studying both the topic models of a specific user and those of a specific restaurant with rating break-downs.

Currently, Yelp has a recommendation that does not take the user's preference and potential matching with restaurants into account. Our recommendation will benefit customers by providing a more personalized user experience.

2 Research Objective

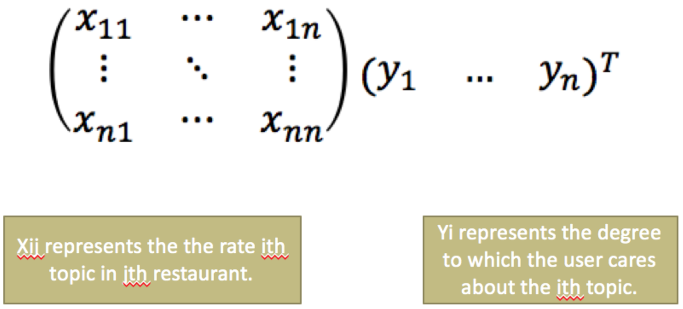
The ultimate goal for the project is to create a recommendation system which recommends restaurants to a specific user given the user's preference and the restaurants' rating with respect to the user's preference. To achieve the goal, the most important part is to extract dimensions for both overall rating for restaurants and user preference using information from reviews. Information we care about includes what a specific user cares about, food or view, price or service and what factors lead to the overall rating for a specific business.

3 Method

The ultimate goal for the project is to create a recommendation system which recommends restaurants to a specific user given the user's preference and the restaurants' rating with respect to the user's preference. To achieve the goal, the most important part is to extract dimensions for both overall rating for restaurants and user preference using information from reviews. Information we care about includes what a specific user cares about, food or view, price or service and what factors lead to the overall rating for a specific business.

The $n \times p$ matrix W in the above figure represents the rating score of each topic for each restaurant. Each row represents a restaurant and in that row, each column represents the rating score of a specific topic.

The $p \times 1$ matrix M on the right represents the user's preference for each topic. Multiplying the two matrices gives the personalized score of each restaurant for each user.



$$\begin{pmatrix} x_{11} & \cdots & x_{1n} \\ \vdots & \ddots & \vdots \\ x_{n1} & \cdots & x_{nn} \end{pmatrix} (y_1 \quad \cdots \quad y_n)^T$$

Xij represents the the rate ith topic in jth restaurant.

Yi represents the degree to which the user cares about the ith topic.

$$f(\mathcal{T}|\Theta, \Phi, \kappa, z) = \sum_{r_{u,i} \in \mathcal{T}} \underbrace{(rec(u,i) - r_{u,i})^2}_{\text{rating error}} - \mu \underbrace{l(\mathcal{T}|\theta, \phi, z)}_{\text{corpus likelihood}}.$$

First of all, to find the column of each user's preference, run LDA on all of text reviews of all restaurants and generate k topics. Then for each user, we extract all the reviews as a corpus and calculate the probability that each word belongs to each of the k topics. Finally we add up and normalize to get the preference column of each user where there is a 0.0-1.0 weight for each of the k topics.

Then to find the rating scores for subtopics for each restaurant in the matrix W, we use the following formula to train a model to predict the score of subtopics. The loss function is essentially the least squread summation of recommended score minus real rating score.

4 Research Road Map

4.1 Experiment Design

Our objective is to learn hidden dimensions of behind a overall rating for a specific restaurant by combining latent rating dimensions, such as the topic learned by topic modeling methods like LDA. First we run topic model to get topics from reviews written by a user and topics from reviews on a restaurant. Then we combine topics we learned from the user and the restaurants as topic factors. Notice that the user preference and topics of restaurants is the same number, as well as topic factors. Finally, we learn the hidden dimension of rating by minimizing the mean squared error function defined by difference between the overall rating calculated by scores of hidden dimensions and the real rating of a review from one user for a restaurant.

In order to do that, we collected all the reviews from one user to get topics from the text as well as the probability of the topics. In effect, the probability of the topic in all the reviews from each user can be the "preference" of the user. After we got the topics of the user, we subsequently used the same model on all the reviews for one specific restaurant. However, the results we get from LDA lack the adjectives that can distinct the positive or negative quality of the topic. So our next goal is to learn the positive or negative quality of topics.

After achieving the goal of finding the distinct quality of topics, we can then learn the hidden dimension of rating based on the topic and the overall rating.

4.2 Experiment Validation

For each user, predict the rating score for the restaurants that the user has already visited before and compare with the actual rating score.

4.3 Ideal Plan

Use LDA to get topics frequently talked in all the reviews of a restaurant and all the reviews from one user respectively. Then set transformation between rating of each k topics for ith restaurant and topic distribution and summarization we get from LDA. Finally minimize the loss $\sum_{i=1}^{i=n}(r_{i,p} * u_p - o_i)$, where $r_{i,p}$ is the trained rating of pth topic for nth restaurant, u_p is the preference of the pth topic for a user and o_i is the original rating.

4.4 Fallback Plan

In the ideal plan, we aim to get a good summarization of topics from reviews and give rating dimensions according to the topic summarization we get. However, if the evaluation shows the method is not as good as the method where rating dimensions is just trained by minimizing the loss function with gradient descent, our fallback plan is to compare the two methods and figure out why it doesn't work.

4.5 Research Timeline

Oct 21 - Oct 28: Use the methods from previous studies to extract topics from reviews of a specific business with quantified score from 1.0-5.0 for each topic

Oct 28 - Nov 4: Extract hidden topics from each user's reviews and generate a preference list for each user. Build the recommendation system by figuring out the matrix sizes and coefficients

Nov 4 - Nov 11: Perform evaluation on the recommendation system by using the trained the model to predict the score of restaurants that users have already visited and then compare it with baseline.

Nov 12 - Nov 19: Train rating dimensions with user preference and frequency of topics which users care about for restaurants from LDA.

Nov 20 - Nov 27: Improve existing models by improving topics we get from LDA and then compare it with baseline. Explain why the method is better or worse.

Nov 28 - Dec 1: Finish writing of the official paper and presentation.

4.6 Resources

We use the datasets of yelp reviews, business and user to extract topics from reviews for popular restaurants (restaurants with most reviews) active users (users with most reviews).

References

[1] *Wang, Chong, and David M. Blei. "Collaborative topic modeling for recommending scientific articles." Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, 2011. APA Guo, Yue, Stuart J. Barnes, and Qiong Jia. "Mining meaning from online ratings and reviews: Tourist satisfaction analysis using latent dirichlet allocation." Tourism Management 59 (2017): 467-483..*

[2] *Linshi, Jack. "Personalizing Yelp star ratings: A semantic topic modeling approach." Yale University (2014)..*

[3] *Multi-label text classification with a mixture model trained by EM.*

[4] *Hierarchical Topic Models and the Nested Chinese Restaurant Process.*