Learning Semantic Representation from Restaurant Reviews: A Study of Yelp Dataset

Sanqiang Zhao[1], Shuguang Han[1], Rui Meng[1], Daqing He[1], Danchen Zhang[1]
[1]University of Pittsburgh

Abstract
Users' preference such as rating only provides uni-dimension information, but reasons behind users' preference may be related to various aspects of an item, such as the types, certain attributes. By observing user-generated review always provides such rich information, we proposed an item representation based on review data. This approach supports semantic operation, which could potentially enables more recommendation scenarios. Our experiments further demonstrated that this approach gained much better performance than classical item representation methods.

Keywords: Semantic representation; contextual information modeling; recommender System
DOI: Citation info is to be added.

Contact: saz31@pitt.edu, shh69@pitt.edu, rum20@pitt.edu, dah44@pitt.edu, yuc73@pitt.edu, daz45@pitt.edu

1 Introduction

Recommender systems have been widely developed in recent years for eliminating the problem of information overload and providing personalized recommendations of items (e.g., books, accommodations, musics and news articles) (Koren, Bell, Volinsky, et al., 2009). *User* and *Item* (the to-be-recommended products) are two important concepts in a recommender system. Modern recommendation techniques usually attempt to learn users' tastes based on their item preference histories (often utilizing users' ratings on items), and make recommendations based on a group of like-minded users who share the preference on the same items. However, the preference information may suffer from data shallowness — users' preference such as rating to an item only provides a one-dimension assessment of the item, but the reasons behind the users' preference may be related to any combinations of various aspects of the item, such as the types, certain attributes, its relationship with other items, and even surrounding contextual environment. Therefore, external data resources that provide additional information and full-coverage of preference aspects are favored to enhance the current recommendation performance.

We think that the user-generated review for items is one such resource. Particularly, since the number of users in a commercial recommender system usually far exceeds the number of items, there are more information for items than users. In terms of reviews, we do find that many items consist of tens or hundreds of reviews in many large-scale datasets, and each piece of review further contains tens or hundreds of words. More importantly, content modeling for review texts can provide a better and complementary understanding of why and how users like the items. Therefore, we focus on developing a better content-based modeling approach in this paper. Previous studies have tried both the simple bag-of-word approach (Pazzani & Billsus, 2007) and the latent topic modeling approach that tries to capture the semantic meaning of content (Wang & Blei, 2011). However, the latent semantic analysis for short text such as reviews usually does not yield a good performance (Hong & Davison, 2010), which motivates us to find alternative solutions.

Recent efforts building on top of the modeling of the contextual information for content have achieved substantial performance boosts in many text modeling tasks (Mikolov, Chen, Corrado, & Dean, 2013). One successful example is the development of Word2Vec (Mikolov et al., 2013), which models each word into a low "semantic" dimension, with a dense vector based on the word's usage contexts. Inspired by this idea, we thought that each item in a recommnder system can also be represented by a lower "semantic" dimension based on its context. Here, we use item reviews for representing its context since it directly related to how people assess the item. The new representation of items enables us to perform a set of semantic operations. In our later analysis, for example, we find that by aggregating the representations of a Chinese restaurant in Scottsdale, Arizona and a Casino in Las Vegas, we could locate a restaurant in Las Vegas and serves Chinese

food. We will provide more detailed explanations about such representation in §2. We believe such operations serve more recommendation scenarios.

To further understand how and whether this modeling approach would work in real-world scenarios, we follow the standard recommendation experiment protocol and attempt to predict user ratings. Specifically, we hold out a certain amount of user ratings for testing, and the rest for training. By comparing our new representation approach with the traditional representation approaches, we can then examine the effectiveness of our model. Again, due to the relatively sparsity of user information in the commercial recommender system, we focused on the item-based recommendation in our paper (Lee & Seung, 2001; Salakhutdinov & Mnih, 2011; Hoyer, 2004).

2 Our Approach: Contextual Representation of Items through Reviews

As mentioned above, the bag-of-word representation often encounters the vocabulary mismatch problem, whereas latent topic modeling can, to some extent, solve this problem by mapping each word into a lower "semantic" dimension. However, latent topic modeling does relatively poorly on handling word contextual information. Therefore, our approach aims to handle both semantic representation and contextual information of an item.

Specifically, our model can be illustrated by Figure 1. Suppose that we have two items: item i and item j, and each of them receives one piece of review, R_i and R_j, respectively. R_i consists of three words (w_1, w_2, w_3), and R_j contains two words (w_4, w_5). In the beginning, we represent each word using a vector. The vector is obtained through pre-training a large data corpus based on Word2Vec (Mikolov et al., 2013). In this paper, we set the vector dimension as 200. On top of this word vector representation, we then aggregate word vectors to represent an item if a word has appeared in the item review. The aggregation process attempts to search for an optimal item representation (also represented by a vector, with the same dimension as a word vector) so that the item vector becomes more similar to these words that are in reviews and less similar those words that are not in reviews. More details of our approach can be referred to (Dai, Olah, & Le, 2015).

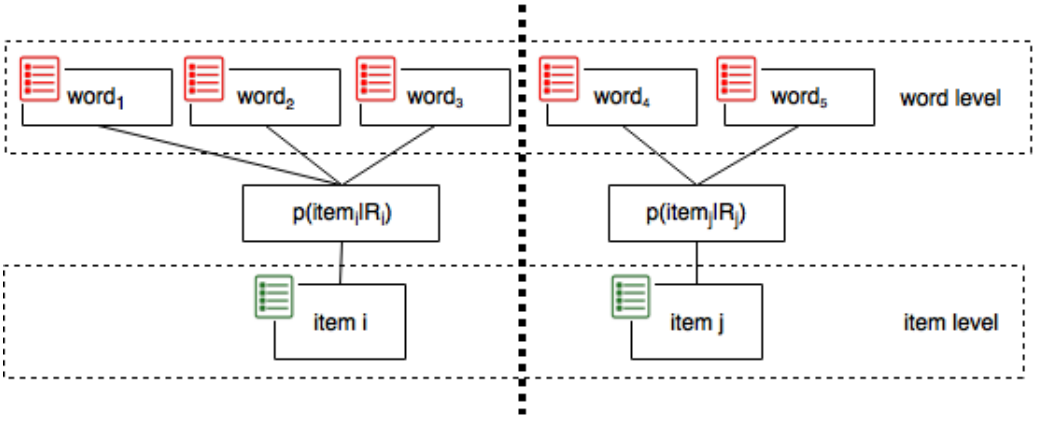

Figure 1: An illustration of our proposed approach, in which both words and items are represented by vectors. The word vector is pre-trained based on Word2Vec and the item vector is obtained by maximizing the item generation probability $p(item_i|R_i) = \sum_{w \in R_i} p(item_i|w)$.

3 Experiment

3.1 Dataset

Our experiment utilizes all of the restaurants and their review information from a large-scale Yelp Challenge Dataset[1]. In total, it consists of 24,974 restaurants located in 10 cities and across four countries, and in total

[1] https://www.yelp.com/dataset_challenge

1.3 millions of reviews. Based on these reviews, our model then generates the corresponding representation for each restaurant.

3.2 Understanding Item Representation

To better understand the learned representation for each restaurant, we conduct a simple qualitative analysis as shown in Table 1, in which we try to find the most similar restaurants for each query restaurant. Specifically, at first, we aggregate the item representations of the query restaurants. After that, we compute the cosine similarities of the aggregated representation with each of restaurants in our dataset. The top two or three restaurants are provided. According to Table 1, we find that through our representation, we could easily locate the most similar restaurants with shared attributes.

Querying a Chinese restaurant in Pittsburgh (e.g., china-palace-pittsburgh) enables us to locate other restaurants that also serves Chinese foods (e.g., long-jin-chinese-cuisine-las-vegas-2, yummy-yummy-chinese-restaurant-scottsdale) but in different locations. More interestingly, if we aggregate the representations of a Chinese restaurant in Scottsdale, Arizona (yummy-yummy-chinese-restaurant-scottsdale) with the representation of a Casino in Las Vegas (the-mirage-las-vegas-3), then we could locate a restaurant in Las Vegas and also serves Chinese food (long-jin-chinese-cuisine-las-vegas-2). We believe the above semantic operations would have many potential applications. For instance, a person who travels often can easily locate his desired restaurant in one city by providing his favorite restaurant in a different city.

Query restaurant[2]	Most similar restaurants	Shared attributes
the-mirage-las-vegas-3	bellagio-hotel-las-vegas new-york-new-york-hotel-casino-las-vegas monte-carlo-hotel-and-casino-las-vegas	Located in Las Vegas, Casinos and Hotels
valle-luna-phoenix	carlos-o-briens-phoenix-phoenix valle-luna-phoenix-2 la-fonda-del-sol-scottsdale-2	Mexican food
which-wich-middleton	which-wich-charlotte which-wich-chandler-3	Sub-branch of which-wich
pkwy-tavern-las-vegas-2	carolina-ale-house-charlotte the-house-of-brews-gilbert duckworths-grill-and-taphouse-charlotte-2	Having sports bar
china-palace-pittsburgh	long-jin-chinese-cuisine-las-vegas-2 yummy-yummy-chinese-restaurant-scottsdale	Serving Chinese food
yummy-yummy-chinese-restaurant-scottsdale the-mirage-las-vegas-3	long-jin-chinese-cuisine-las-vegas-2	Located in Las Vegas and serving Chinese food

Table 1: An qualitative analysis of item representation through locating the most similar restaurants

3.3 Review-based Recommendation

To further understand whether this representation works in real-world scenarios, we follow the standard recommendation experiment protocol (Koren et al., 2009) and conduct item-based recommendations to predict users' ratings. At first, the dataset is split into 80/20 (80% for training and 20% for testing) based on the review posting time. Then, based on different approaches for modeling items, we locate the most similar items (top k items, where k is a parameter and we tried 3, 5 and 7) for each of these items. Finally, we predict user rating on each of the similar items as their averaged rating from other users. The prediction performance is evaluated based on the square error of the true rating and the predicted rating, i.e., root-mean-square error (RMSE).

Table 2 shows the result of our experiment, where two baselines (Bag-of-Words and LDA) and our approach are applied. They are utilized to find the top k similar items. The bag-of-word baseline computes similarity based on whether two items share exactly the same word, whereas LDA and our model tend to match top k documents based on semantic relations. For LDA, we also use 200 topics to align with our model. Meanwhile, this is also the common setting in many LDA applications (Blei, Ng, & Jordan, 2003).

[2]Readers can check the restaurant information by visiting https://www.yelp.com/biz/ plus a Yelp Id, for example, https://www.yelp.com/biz/yummy-yummy-chinese-restaurant-scottsdale.

RMSE	Bag-of-Words	LDA	Our Model
k = 3	2.841	2.731	1.507
k = 5	3.122	3.107	1.479
k = 7	3.308	3.276	1.472

Table 2: Results of the recommendation experiments. A small RMSE indicates a better performance.

As shown in Table 2, both LDA and our model outperform the bag-of-word approach, indicating the effectiveness of the semantic modeling of items. Our approach achieves significantly the best performance compared to LDA, denoting the necessity of modeling the semantic information based on contexts. In addition, our model tends to be insensitive to different configurations of k, which is a strong positive message to recommendation community since it is the most difficult parameter in a recommender system.

4 Conclusion

Most of the existing recommendation approaches remain rely on the simple user rating information, whereas such uni-dimension information cannot reveal many aspects of user preferences such as how and why a user prefers one item. Observing that item reviews often provide such rich information, this paper proposed an item representation based on review data, and the start-of-art word text modeling approach based on word contexts (Mikolov et al., 2013). This approach supports semantic operation, which could potentially empowers more recommendation scenarios. Our experiments further demonstrated that this approach gained much better performance than classical item representation methods based on words and semantic topics. We do think that the applications of our approach are not limited to the recommender systems. Similar ideas can be easily applied in any text-based system. We would like to explore more of these applications in the future.

References

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of machine Learning research*, *3*(Jan), 993–1022.

Dai, A. M., Olah, C., & Le, Q. V. (2015). Document embedding with paragraph vectors. *arXiv preprint arXiv:1507.07998*.

Hong, L., & Davison, B. D. (2010). Empirical study of topic modeling in twitter. In *Proceedings of the first workshop on social media analytics* (pp. 80–88).

Hoyer, P. O. (2004). Non-negative matrix factorization with sparseness constraints. *Journal of machine learning research*, *5*(Nov), 1457–1469.

Koren, Y., Bell, R., Volinsky, C., et al. (2009). Matrix factorization techniques for recommender systems. *Computer*, *42*(8), 30–37.

Lee, D. D., & Seung, H. S. (2001). Algorithms for non-negative matrix factorization. In *Advances in neural information processing systems* (pp. 556–562).

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Pazzani, M. J., & Billsus, D. (2007). Content-based recommendation systems. In *The adaptive web* (pp. 325–341). Springer.

Salakhutdinov, R., & Mnih, A. (2011). Probabilistic matrix factorization. In *Nips* (Vol. 20, pp. 1–8).

Wang, C., & Blei, D. M. (2011). Collaborative topic modeling for recommending scientific articles. In *Proceedings of the 17th acm sigkdd international conference on knowledge discovery and data mining* (pp. 448–456).