Association Rules and Item-Based Collaborative Filtering for Amazon.com Baseline Recommender System

Vinh Nghiem To '22 and Dr. Sofia Visa (Advisor)

Abstract

Recommendation algorithms are best known for their use on e-commerce Web sites, where they use input about a customer's interests to generate a list of recommended items. In this study, the items that customers purchase and explicitly rate to represent their interests are applied into the recommendation models to personalize the online store for each customer. The dataset used for building the models is extracted from Amazon Customer Reviews, which contains customers' ratings in over two decades since the first review in 1995. We apply to a subset of this large dataset the association rule technique and the item-based collaborative filtering to identify additional items that a given customer might like to buy. Studying the top recommendation helps us better understand how complementary goods and brand loyalty are applied in e-commerce. Moreover, we discuss theoretically sentiment classification and analysis methods that can be useful in future work.

The Dataset: Amazon_US_Reviews

In a period of over two decades since the first review in 1995, millions of Amazon customers have contributed over a hundred million reviews to express opinions and describe their experiences regarding products on the Amazon.com website. This makes Amazon Customer Reviews a rich source of information for academic researchers in the fields of Natural Language Processing (NLP), Information Retrieval (IR), and Machine Learning (ML), amongst others. Over 130+ million customer reviews are available to researchers in TSV files in the amazon-reviews-pds S3 bucket in AWS US East Region. Each line in the data files corresponds to an individual review.

The dataset we are using is the Amazon Home Improvements, Kitchen, and Personal Care product reviews which is a subset of the large Amazon Product review. The dataset is already stored in the TensorFlow database and can be loaded directly using the *tfds* API from Tensorflow. Once the dataset is loaded, we convert it into a *Pandas* data frame using ***tfds.as_dataframe*** API.

customer_id	helpful_votes	marketplace	product_category	product_id	product_parent	product_title
b'28952152'	0	b'US'	b'Health & Personal Care'	b'B00NNTBF22'	b'281393741'	b'PRO-15 Advanced Strength Probiotics: 3x the ...

review_body	review_date	review_headline	review_id	star_rating	total_votes	verified_purchase	vine
b"When looking for a probiotic, time released ...	b'2015-02-23'	b'Good all purpose probiotic'	b'R1RGZC9IBTJ4LS'	4	0	1	1

Figure 1. An example of a review in the dataset

To ensure the quality of review's ratings, only reviews that are verified purchase are considered. In the data shown herewith, the number of unique customers is 1,285,915, the number of unique products is 450,860, the average number of reviews per customer is 1.51 and the average number of reviews per product is 4.30. While this is a small sample, it is well-representative of the data and presents an example of how a subset can be useful in dealing with much larger datasets.

```
print('Number of unique customers: ', df.customer_id.nunique())
print('Number of unique products: ', df.product_id.nunique())
print('Review per customer: ', len(df)/df.customer_id.nunique())
print('Review per product: ', len(df)/df.product_id.nunique())
```

```
Number of unique customers:  1285915
Number of unique products:  450860
Review per customer:  1.5069798548115545
Review per product:  4.2981147141019385
```

Figure 2. Calculations for brief data overview

Association Rules

Association Rule is one of the most common data mining techniques used to model market basket analysis. Retailers use this to increase sales by better understanding customer purchasing patterns. In daily life, supermarket for example, dairy items are placed in the same aisle, fresh fruits and vegetables are in the same shopping area, and beverages form another set. Organizing items in a scientific manner not only saves the customers' shopping time but also increases stores' profit as it encourages customer cross-item buying. Association rules help reveal such relationships between items [1].

Support, ***Confidence***, and ***Lift*** are the three main metrics when studying association rules. ***Support*** gives an idea of how frequent itemset is in all transactions. ***Confidence*** defines the likeliness of an item added to the cart given that the customer already bought the antecedents. ***Lift*** controls for the frequency while measuring the conditional probability to avoid bias conclusion from only *Confidence* value.

$$Support(\{X\}) = \frac{Number\ of\ customers\ who\ bought\ X}{Number\ of unique\ \ customers} \tag{1}$$

$$Confidence(\{X\} \rightarrow \{Y\}) = \frac{Number\ of\ customers\ who\ bought\ both\ X\ and\ Y}{Number\ of\ customers\ who\ bought\ X} \tag{2}$$

$$Lift(\{X\} \rightarrow \{Y\}) = \frac{Confidence(\{X\} \rightarrow \{Y\})}{Support(\{Y\})} \tag{3}$$

itemA	itemB	freqAB	supportAB	freqA	supportA	freqB	supportB	confidenceAtoB	confidenceBtoA	lift	categoryA	categoryB
b'Natizo Stainless Steel Measuring Spoon Set -...	b'Liquid Vitamin D Drops - 2oz D3 100 Iu Per D...	38	0.001278	404	0.013588	299	0.010056	0.094059	0.127090	9.353406	b'Kitchen'	b'Health & Personal Care'
b'Ozeri ZB19-W Rev Digital Bathroom Scale with...	b'Ozeri Green Earth Textured Ceramic Nonstick ...	35	0.001177	123	0.004137	184	0.006188	0.284553	0.190217	45.981575	b'Health & Personal Care'	b'Kitchen'
b'Ozeri Green Earth Textured Ceramic Nonstick ...	b'Ozeri ZB19-W Rev Digital Bathroom Scale with...	31	0.001043	184	0.006188	123	0.004137	0.168478	0.252033	40.726538	b'Kitchen'	b'Health & Personal Care'
b'Natizo Stainless Steel Measuring Spoon Set -...	b'Athelas Neutraceuticals Natural Triple Stren...	31	0.001043	404	0.013588	204	0.006861	0.076733	0.151961	11.183787	b'Kitchen'	b'Health & Personal Care'
b'Natizo Stainless Steel Measuring Spoon Set -...	b'PRO-15 Advanced Strength Probiotics: 3x the ...	29	0.000975	404	0.013588	298	0.010023	0.071782	0.097315	7.162079	b'Kitchen'	b'Health & Personal Care'

Figure 3. Association Rules from different categories (sorted by frequency)

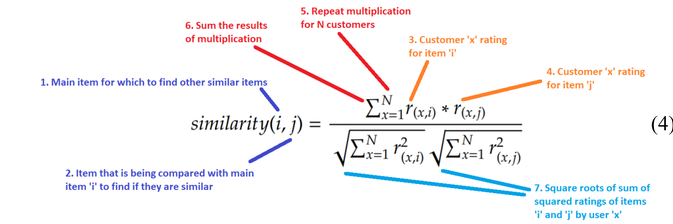
The brand royalty theory is strengthened by an example in the second and third rows of Fig. 3, where customers who bought "Ozeri Green Earth Textured Ceramic Nonstick Pan" are also confident in buying "Ozeri ZB19-w Rev Digital Bathroom Scale with Electro-Mechanical Weight Dial" and vice versa. Although these two products are from different categories (Kitchen and Health & Personal Care), they are daily used lifestyle products for the modern home. Therefore, once they are satisfied with their first purchase with Ozeri brand, many customers believe that other kitchen, bath, entertainment, and personal amenities goods provide similar high quality usage experience.

Rows 1, 4, and 5 in Fig. 3 shows that customers who bought vitamin/supplement goods also have the incentive to buy stainless steel spoons to use with (complementary goods). Capturing the brand loyalty along with the example above, firms selling those personal care products should include a spoon in their future product and sell them as a bundle at a higher but reasonable price. Informed customers giving good ratings for the previous goods will still buy the bundle since they think that the spoon included can help them consume the servings with more exact amount.

Item-Based Collaborative Filtering

In the early 1990s, collaborative filtering started to develop as a solution for the contemporary situation of users getting overloaded with disorganized and cumbersome display of information provided on the Internet. Collaborative filtering technique is defined to be an approach of setting up a recommender system, which assesses historical interactions and behaviors of the users and then regulates the link between the users and their items of interest. Item-based collaborative filtering (IBCF) is one kind of recommendation method which looks for similar items based on the items users have already liked or positively interacted with. IBCF suggests an item based on items the user has previously consumed and consists of two parts. Item-based techniques first analyze the user-item matrix to identify relationships between different items, and then use these relationships to indirectly compute recommendations [2].

The similarity score between two items is calculated as shown in Equation 4.



(4)

Figure 4. Cosine similarity equation, details labeled

The cosine similarity uses cos(θ) to measure the distance between two items (representing by two vectors), so as cos(θ) increases, θ decreases. Therefore, if the similarity value is closer to 1, we have θ closer to 0, and the closer the two vectors are. In other words, these two items are likely to be rated similarly.

When using similarity to predict the ratings that a customer would give for the items that have not been purchased before, we have the following result:

```
Customer b'10011509' bought:
      product_id                                      product_title  \
0  b'B00NNTBF22'  b'PRO-15 Advanced Strength Probiotics: 3x the ...
1  b'B00OG5O37Q'  b'Immune: Hyperbiotics Daily Immune & Wellness...
2  b'B00RH5I8U0'  b'Green Tea Extract Supplement with EGCG for W...
3  b'B00RH5K26I'  b'Digestive Enzymes Plus Prebiotics & Probioti...

             product_category rated_score
0  b'Health & Personal Care'            5
1  b'Health & Personal Care'            5
2  b'Health & Personal Care'            5
3  b'Health & Personal Care'            5
Recommended items:
      product_id                                      product_title  \
0  b'B00CUJMDNG'  b'Natizo Stainless Steel Measuring Spoon Set -...
1  b'B00DPNCEY4'  b'Liquid Vitamin D Drops - 2oz D3 100 Iu Per D...
2  b'B00KFETGW4'  b"Nature's Fruit ThermoCleanse (30 Servings, 6...
3  b'B00LPR2ZAU'  b'Athelas Neutraceuticals Natural Triple Stren...
4  b'B00ONH0S8U'  b'Hair Growth Vitamins Supplement - 5000 mcg B...

            product_category  recommended_score
0           b'Kitchen'            2.508418
1  b'Health & Personal Care'            2.493798
2  b'Health & Personal Care'            2.848355
3  b'Health & Personal Care'            2.564384
4  b'Health & Personal Care'            3.545424
```

Figure 5. Top five predicted rating for customer b'10011509'

Previously, customer b'10011509' bought probiotics, daily immune hyperbiotics, green tea extract supplement, and digestive enzymes and is very satisfied with those product, rating them all 5/5. Based on this history of purchase, IBCF program recommends other Health & Personal Care products that customer b'10011509' might be interested in, including vitamin D, fruit's extract, and hair growth supplement, etc. These supplements are often rated similarly to the products that the customer bought, and in practice, they can help in compensating negative side effects of using digestive products over a long time. Therefore, this recommendation is reasonable in terms of real-life circumstance.

Conclusion

This study presents exploratory research that uses and illustrates intelligent data mining approaches such as association rules and item-based collaborative filtering in the context of an e-commerce baseline recommender system. By examining a user's purchase history, we learned how businesses uses complementary good and brand loyalty theory to increase their profit and enhance customer's experience.

References

[1] Chengqi Zhang and Shichao Zhang. *Association Rule Mining*. Springer, 2002.

[2] Badrul Sarwar, George Karypis, Joseph A Konstan, and John Riedl. Item-based collaborative filtering recommendation algorithms. *Proceedings of the 10th International Conference on World Wide Web, WWW 2001*, pages 285–295, April 2001.