Balanced Diet Grocery Recommendation Algorithm

Panxi Chen[1] and Zehua Wang[2]

University of Michigan, Ann Arbor, MI

1 Introduction

With the development of science and technology and the progress of society, people have more convenient lifestyles, but what follows is that under the pressure of work or study, the energy that can be spent on their physical and mental health is less and less. So we came up with an idea, can we use products of rapidly evolving technology to help people care for their physical and mental health. This idea leads to shopping recommendations.

Online shopping is an excellent proof of the development of technology. After many attempts, we believe that most shopping apps or websites only recommend items that customers like or prone to repurchase. People usually search on the Internet (another tremendous technological creation) or go to the hospital to seek medical help after they feel unwell, but the efforts made at this time can only be called remedial.

In today's fast-paced lifestyle, people usually don't pay attention to the nutrients they consume in their daily diet, especially now that there are so many fast food options[13], which leads people to focus on what they want to eat instead of what they need to eat. Prolonged unbalanced nutritional intake can lead to overweight and obesity, heart disease and stroke, type 2 diabetes, cancer, etc[11], so reminding people when they shop can effectively prevent such things from happening.

In summary, we want to focus on balanced diet grocery recommendations. By retrieving part of the consumer's consumption records and the nutritional value of the purchased products, the intake ratio of various nutrients is calculated and compared with the optimal nutritional intake ratio to obtain the nutrients that customers need to supplement. Extract information from consumption records and obtain consumers' consumption habits, and recommend products with nutrients that consumers lack.

2 Data

We obtained a dataset containing multiple purchase records of users from the official website of instacart[1] as our main analysis object. The original dataset contains over 3 million orders from more than 200,000 customers. We first filter the product categories in the dataset to pick out products in the category of grocery, and then filter the 1,500 most frequently purchased products from the above grocery products. Then we obtained the nutritional content of each product from the website Nutrition Value[9] through the crawler, cleaned it up and used it as the data for calculating the nutritional intake of customers. We use BERT model[4] to match the name of product from nutrition value dataset to instacart dataset and manually check the names match. Finally, we obtain the nutritional intake standards of adults from FDA[5] as an important indicator for our balanced diet recommendation algorithm.

The dataset will be randomly split into train, validation, and test with ratio 6:1:3. To ensure no information leak, the customers and their orders will be different between train and test dataset. The number of customers' shopping records ranges from 1 to 99, we filter out customers with too few shopping records (less than 10). We randomly labeled 30% orders for each customers for evaluation in three datasets. We use cosine similarity to compare the preference list generated by the model trained with unlabeled orders with the label orders to check the model performance.

After data cleaning, we have over 2.2 million orders from around 87,000 customers. The number of products we used is 1518, and we take 31 nutrition into consideration.

Type	Orders	Customers	Products	Nutrition
Train	1318728	52189	1518	31
Validation	8698	20012	1518	31
Test	663210	26094	1518	31
Total	2204854	86981	1518	31

Table 1: Cleaned Data Instance Overview

3 Related Work

In the paper[2] reviewing nutrition recommendation systems (NRS) and their characteristics for the first time. Nutritional informatics is a novel approach to help healthcare, and the recommendation system is an effective way to put this approach into practice. And the results of the article show that people usually use NRS on mobile, and the types of recommendation systems are mostly hybrid recommendation systems and knowledge-based recommendation. The recommendation system mentioned in the paper is actively used by users, but we are

committed to giving nutrition recommendations to users unknowingly. Ideally, the habits of some users with unhealthy diets will gradually change.

We further research on more research papers related to Diet Recommendation System. Specifically, researchers provides a customized diet recommendation service for preventing and managing coronary heart disease in health care services[8], propose a Personalized Diet Recommendation System for Cancer Patients to help patients manage their daily food intake[7], proposed Food Recommendation System (FRS) by using food clustering analysis for diabetic patients[10], uses the forward chaining and backward chaining method to provide recommendations for infant nutrition in accordance with the age and allergies[3], make a diet food recommendation system for diabetic patients based on ontology[12]. And we found one thing in common that they all have nutrition/diet/food recommendation system for specific populations, such as neonates, cancer patients, chronic disease patient. The algorithm we want to develop is a recommendation system that open to the general public for daily use, which is different from those recommendation systems used for medical needs.

4 Methodology

We want to calculate customer's nutritional intake ratio after finding their preference product lists, then by comparing customer's nutritional intake ratio with the standard, we recommend products from the types of nutrients customers need to supplement.

We first trying to find the preference product list for each customer. When sorting out the dataset, we found that the product names in the customer shopping records contain too much details, and the nutrition of the products often depends on the category of the products, which makes it a key challenge to classify the products in the shopping records. And we decided to use word embedding as baselines to complete this challenge. Word embedding is the technique of converting a word into a numerical representation whose variable type is a vector. And the numerical presentations could be use to identify similarity or dissimilarity between words. Each word can be mapped to a vector, which is then learned using the neural network, and find characteristics of this word relative to the entire text.

After researching on techniques, we pick BERT to produce word embedding for making categorical classification. The latest language model BERT (Bidirectional Encoder Representations from Transformers) is a language model deceloped by researchers at Google AI Language, which could be used for natural language processing (NLP) pre-training. BERT borrows word2vec's CBOW pattern. Compared to BERT, word2vec is an older algorithm learns word associations from a text corpus by using a neural network. Technique word2vec selects

specific vectors that capture the semantic and syntactic qualities of words to represent each word, so we can use cosine similarity to indicate similarity and dissimilarity between words, and it has two main architectures: the Skip-gram model uses a word as input to predict the context of the word; the CBOW model takes the context of a word as input to predict the word itself. Therefore, we choose the BERT upgraded from CBOW model to predict a target word from a list of context words (category). The power of word2Vec lies in its ability to group together vectors of similar words, which aligns with our goal of putting similar products (items of the same class) into the same groups. And BERT applies bidirectional training of Transformer, an encoder mechanism, to generate a language model, which could learn contextual embedding for words deeper than word2Vec. After the Transformer encoder read the entire sequence, the sequence of tokens are embedded into vectors and then processed in the neural network. BERT also avoid loss of accuracy due to some situations that limit context learning by using Masked LM and Next Sentence Prediction as training strategies.

The nutrient composition and nutrient content of each product are different. We take each nutrient as a classification condition. Each classification consists of products containing the nutrition corresponding to this classification, and is ranked according to the nutrient content of the same unit. When we filter out the types of nutrients that a user lacks and excess, we retrieve the products that the user lacks in nutrition at the top and the user's excess nutrition ranks low, or retrieve the user's lack of nutrition that ranks high and does not rank. Products present in the excess nutrients category. Then according to The Five Food Groups[6], we then divide commodities into five categories by nutrient type to simplify and reduce computational costs.

In order to seek the difference between the recommended list based on nutritional needs and the customer's original preference list, we calculate two cosine similarities for comparison: the customer's product preference list and the customer's nutritional needs product list.

5 Evaluation

We primarily evaluate the similarity between the recommendation shopping list and the purchase history. We use the test dataset and the pre-trained BERT model to generate two baselines for evaluations. The first baseline is the shopping list generated by randomly pick products. The second baseline is the shopping list generated by most frequent shopping products. Here, we pick 10 customers in test dataset as an example. All product names in test and baselines are embedding with BERT model, and calculate the mean cosine similarity as the metric. We also provide the average number of products per order for each customers. The average is round up as integers.

User ID	Avg. Number of Products per Order	Baseline 1	Baseline 2
3	5	0.6408	0.6825
13	4	0.5913	0.6450
17	2	0.6570	0.6298
22	4	0.6457	0.7028
27	5	0.6836	0.6501
38	5	0.6415	0.6583
43	6	0.5978	0.6366
48	7	0.6281	0.6559
52	2	0.6330	0.6702
57	3	0.5680	0.6627
Average	4.3	0.6290	0.6594

Table 2: Mean Cosine Similarity Under Two Baselines

6 Work Plan

We have completed the data cleaning of shopping records, classified products, matched the nutritional content of each category, calculated the proportion of various nutrients ingested by customers, and compared this proportion with the nutrient intake ratio standard to obtain the type of nutrition that each customer needed. We are currently utilizing the Bert model to obtain the preference list of each customer. Then we are going to extract and sort the nutrition of the products of each group in The Five Food Groups[6], obtaining the product categories that customers need for nutrition, and Update preference list based on nutrition needs. Calculating the cosine similarity for both recommendation list, we draw a conclusion after comparing it with the cosine similarities from randomly selected customers.

Section	Time
Get the nutrition type that customers need	Nov 18
Get the preference list of each customer	Nov 25
Update preference list based on nutrition needs	Dec 02
Get the final conclusion (cosine similarity)	Dec 09
Modify details and report content	Dec 09 - Dec 14
Final project report	Dec 14

7 Results

8 Conclusion

References

[1] *3 Million Instacart Orders, Open Sourced*. May 2017. URL: https://tech.instacart.com/3-million-instacart-orders-open-sourced-d40d29ead6f2.

[2] Shahabeddin Abhari et al. "A systematic review of nutrition recommendation systems: with focus on technical aspects". In: *Journal of biomedical physics & engineering* 9.6 (2019), p. 591.

[3] Ratih Nur Esti Anggraini, Siti Rochimah, and Kessya Din Dalmi. "Mobile nutrition recommendation system for 0–2 year infant". In: *2014 The 1st International Conference on Information Technology, Computer, and Electrical Engineering*. IEEE. 2014, pp. 272–275.

[4] Jacob Devlin et al. "Bert: Pre-training of deep bidirectional transformers for language understanding". In: *arXiv preprint arXiv:1810.04805* (2018).

[5] Center for Food Safety and Applied Nutrition. *Daily value on the New Nutrition Facts Label*. URL: https://www.fda.gov/food/new-nutrition-facts-label/daily-value-new-nutrition-and-supplement-facts-labels.

[6] National Health and Medical Research Council. *The five food groups*. Apr. 2021. URL: https://www.eatforhealth.gov.au/food-essentials/five-food-groups.

[7] Wahidah Husain et al. "Application of data mining techniques in a personalized diet recommendation system for cancer patients". In: *2011 IEEE Colloquium on Humanities, Science and Engineering*. IEEE. 2011, pp. 239–244.

[8] Jong-Hun Kim et al. "Design of diet recommendation system for healthcare service based on user information". In: *2009 Fourth International Conference on Computer Sciences and Convergence Information Technology*. IEEE. 2009, pp. 516–518.

[9] *Nutritional values for common foods and products*. Sept. 2022. URL: https://www.nutritionvalue.org/.

[10] Maiyaporn Phanich, Phathrajarin Pholkul, and Suphakant Phimoltares. "Food recommendation system using clustering analysis for diabetic patients". In: *2010 International Conference on Information Science and Applications*. IEEE. 2010, pp. 1–8.

[11] *Poor nutrition*. Sept. 2022. URL: https://www.cdc.gov/chronicdisease/resources/publications/factsheets/nutrition.htm.

[12] B Raj Kumar and K Latha. "DFRS: Diet food recommendation system for diabetic patients based on ontology". In: *Int J Appl Eng Res* 10 (2015), pp. 2765–70.

[13] Giovanni Sogari et al. “College students and eating habits: A study using an ecological model for healthy behavior”. In: *Nutrients* 10.12 (2018), p. 1823.