World Academy of Science, Engineering and Technology
International Journal of Economics and Management Engineering
Vol:15, No:4, 2021

Analysis Model for the Relationship of Users, Products, and Stores on Online Marketplace Based on Distributed Representation

Ke He, Wumaier Parezhati, Haruka Yamashita

***Abstract*—**Recently, online marketplaces in the e-commerce industry, such as Rakuten and Alibaba, have become some of the most popular online marketplaces in Asia. In these shopping websites, consumers can select purchase products from a large number of stores. Additionally, consumers of the e-commerce site have to register their name, age, gender, and other information in advance, to access their registered account. Therefore, establishing a method for analyzing consumer preferences from both the store and the product side is required. This study uses the Doc2Vec method, which has been studied in the field of natural language processing. Doc2Vec has been used in many cases to analyze the extraction of semantic relationships between documents (represented as consumers) and words (represented as products) in the field of document classification. This concept is applicable to represent the relationship between users and items; however, the problem is that one more factor (i.e., shops) needs to be considered in Doc2Vec. More precisely, a method for analyzing the relationship between consumers, stores, and products is required. The purpose of our study is to combine the analysis of the Doc2vec model for users and shops, and for users and items in the same feature space. This method enables the calculation of similar shops and items for each user. In this study, we derive the real data analysis accumulated in the online marketplace and demonstrate the efficiency of the proposal.

***Keywords*—**Doc2Vec, marketing, online marketplace, recommendation system.

I. Introduction

IN recent years, the number of consumers who make purchase on the internet has been continuously increasing owing to the digitalization of society. This holds true for online marketplaces. Online marketplace, rather than conventional online store, dominates ecommerce. The online marketplace acts as a middleman between users and stores and facilitates the sale of goods between the two parties through its extensive network of websites. This directly leads all kinds of brands and retailers to settle in and operate in marketplace to present their products. Therefore, compared to traditional e-commerce sites, online marketplaces can accumulate complete purchase history data about which customers purchase which product in which store.

In recent years, language models have been used in marketing strategies. For example, the word2vec [1] model in the field of natural language processing learns semantic representations of words by identifying relationships between words based on the hypothesis that occurrences of words can be predicted from words in the same context. Each word can be represented by one point in the feature space, and similar words can be gathered nearby. In marketing, this algorithm has been used to classify products based on item vectors [2]. Furthermore, "Doc2Vec" [3], proposed in the field of document classification, is an extension of word2vec. This model learns not only words but also the relationships between documents and enables the use of mathematical expressions for the representation of documents. Using this feature, the model is often used to extract semantic relationships between documents and words [4]. It has also been used in marketing to extract the relationship between users (positioned as Document) and products (positioned as Word) [5].

The use of customer purchase data within a single store has been discussed in previous research [5]. In online marketplaces, there are two types of consumers: those who select a product and then search for a store, and those who select a store and then decide on a product. Therefore, the relationship between products, stores, and consumers is complex, and it is difficult to analyze user behavior patterns. Thus, the purpose of this study is to define consumer preferences based on the purchased products and the corresponding stores and to comprehensively analyze the relationships among stores, products, and users.

Users have great diversity in purchase histories based on related products and stores. Therefore, the analysis based on Doc2vec should be preferable in terms of interpretability and for evaluating marketing strategies. The conventional Doc2vec model was formulated based on two elements. However, the relationship of focus here includes three elements—users, stores, and products. Thus, in this study, we need a model that represents these three elements together. Therefore, we propose a method that combines Doc2vec models for users and stores, as well as users and products, to comprehensively analyze the relationships among these three elements. The proposed method not only classifies products and stores but also predicts the products and stores that match consumers' preferences. In addition, the proposed method is expected to be used in marketing strategies and recommendation systems by identifying stores and products that have high similarity among users. In this study, purchase history data provided by Rakuten Market is analyzed using the proposed method [10]. The results show the effectiveness of the analysis using this method for the data accumulated in online marketplaces.

Ke He and Wumaier Parezhati are grad students and Haruka. Yamashita is a professor in Department of Information and communication Sciences, Sophia University. 7-1 8-261, Kioi-cho, Chiyoda-ku, Tokyo, Japan. (e-mail: k-he-9j2 @eagle.sophia.ac.jp, h-yamashita-1g8@sophia.ac.jp).

II. Doc2Vec

Doc2vec [6] is an analysis method based on language neural networks that are widely applied in the fields of word processing and classification. For example, an article searching system that uses Doc2vec to convert article summaries into distributed representations can search related summaries by calculating similarity. A combination with SVM can detect unknown or bad malware from readable files [7]. A classifier is trained to classify a product description based on a doc2vec-based feature that is augmented in various ways [5]. A feature extraction method using "visual words" automatically learned from video analysis outputs and the Doc2Vec paradigm was proposed [8]. A content-based Bangla news recommendation system used paragraph vectors, also known as doc2vec [9]. The structure of a model that trains the parameters of a neural network is shown in Fig. 1. In order to predict the word to be correctly appear as a target word from the document and surrounding words, the value of the hidden layer obtained from the document information and the surrounding words of the target word calculated from the optimized parameters become the distributed representation of the target word in the feature space.

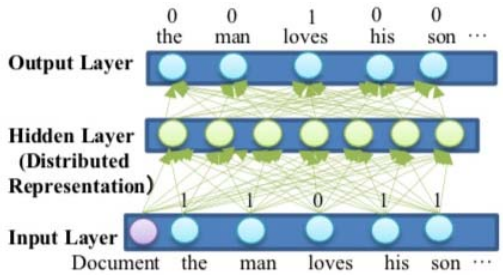

Fig. 1 Distributed representation of "the man loves his son"

For a set of words $(w_1, w_2, w_3 \dots \dots w_k)$ and document information d, we set Window_size (length of surrounding words) to $M(M > \frac{k-1}{2})$. If we denote a predicted word by w_c and the corresponding surrounding words by $w_{c1}, w_{c2}, \dots, w_{c2M}$, the probability of occurrence of the predicted word is given by:

$$P(w_c) = p(w_c | w_{c1}, w_{c2}, \dots, w_{c2M}, d) \quad (1)$$

The neural network is trained to maximize the probability of (1). In addition, a linear expression from the input layer to the hidden layer is substituted for the surrounding words and document information, and the value of the hidden layer becomes a distributed representation. Fig. 1 is an example of the distributed representation for the document "the man loves his son" when the target word is "loves" and $M = 2$.

III. Proposed Model

A. Concept of the Proposal

The purpose of this study is to propose a model to analyze the relationships among users, stores, and products based on Doc2vec. Since consumers may choose a store, and then decide products, or decide the product, and then select stores, there is a variety of relationships between the three input variables. When using this model as a recommendation one, it is advisable to simultaneously calculate stores and products that are highly similar in order to recommend combinations of stores and products to consumers. Therefore, we propose a method for learning a distributed representation that combines two types of Doc2vec models and unifies the feature space to exist.

B. Formulation

First, we denote the consumer set as $(u_1, u_2, u_3 \dots \dots u_j)$, the store set as $(s_1, s_2, s_3 \dots \dots s_k)$, and the product set as $(i_1, i_2, i_3 \dots \dots i_l)$. Next, we train a neural network to the maximum probability of occurrence of s_c based on the target store s_c, the surrounding stores $s_{c1}, s_{c2}, \dots, s_{c2M}$, and consumer u_j.

$$P(s_c) = p(s_c | s_{c1}, s_{c2}, \dots, s_{c2M}, u_j) \quad (2)$$

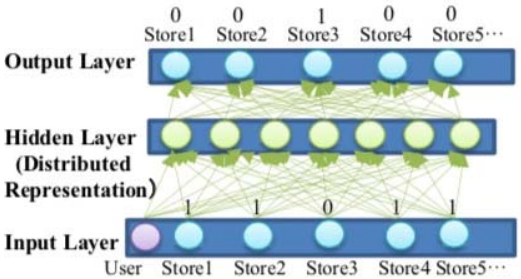

Fig. 2 Distributed representation of "Customer-Store"

The intermediate representations of u_j and s_c are $\boldsymbol{v}_j$ and $\boldsymbol{t_c}$ (both are horizontal vectors). Likewise, let i_c be the target product, and $\boldsymbol{v}'_j$ and $\boldsymbol{N_c}$ (both are horizontal vectors) be the intermediate expressions of u_j and i_c for the peripheral products $i_{c1}, i_{c2}, \dots, i_{c2M}$, and u_j. On the other hand, we define the similarity between consumers based on cosine similarity as:

$$\text{Similarity} = \frac{\boldsymbol{w}_j \boldsymbol{w}'_j}{|\boldsymbol{w}_j||\boldsymbol{w}'_j|} \quad (3)$$

Thereby, we can calculate the similarity between consumers considering both stores and products, and extract stores and products with high similarity to a given consumer. These results are expected to be used in marketing measures.

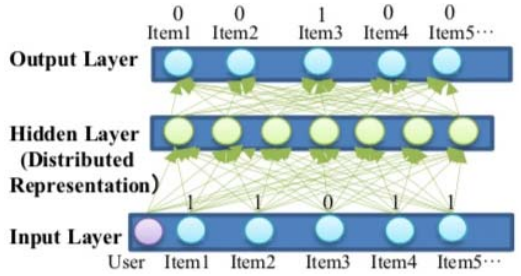

Fig. 3 Distributed representation of "Customer-Product"

In order to learn two variance representations $\boldsymbol{v}_j$ and $\boldsymbol{v}'_j$ of

u_j, we define a horizontal vector $\boldsymbol{w_j} = (\boldsymbol{v_j}, \boldsymbol{v_j'})$. This is the variance representation of the consumer:

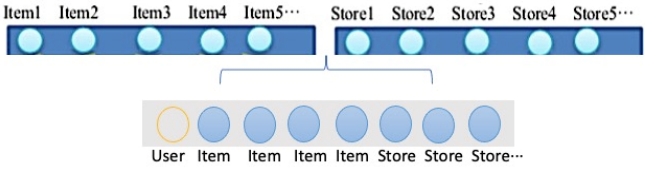

Fig. 4 Combining the Distributed representation of two models

IV. Data Analysis

A. Analysis Setting

In order to confirm the validity of the proposed method, we apply the proposed method to the evaluation history data of Rakuten ICHIBA in 2014 [10]. The evaluation history is considered a part of the purchase history data and is analyzed manually. The number of consumers was 961, the number of stores was 769, the number of products was 886, and the dimensionality of the variance representation $d = 100$ and $W = 2$.

B. Results

Table I shows that products with a high degree of similarity to sugar-free food A are healthy foods, and products with a large degree of sensory similarity to sugar-free food A were selected.

TABLE I
Top 4 Similar Products with to Sugar Free Food A

Product	Similarity
Red Tea A	0.9909
Organic food A	0.9864
Soy product A	0.9834
Pesticide-free tea A	0.9825

The similarity between healthy products is high, which is suitable for recommendations. Table II shows products with high similarity to bread. They are selected because they are staple food items and thus have a high degree of sensory similarity.

TABLE II
Top 4 Similar Products to Bread A

Product	Similarity
Ramen A	0.9394
Tea A	0.9296
Diet-friendly food A	0.9140
Soy product B	0.9082

The results are similar to those for food that are suitable for recommendation. Table III shows that stores with the highest similarity to drug store A include drug stores and beverage stores. This finding is consistent with our premise.

The results in Table IV show that the stores that have the greatest similarity to store A, which deals with food products, are almost all stores that deal with food products, which is intuitively consistent.

TABLE III
Top 4 Similar Stores to the Pharmaceutical Store A Selling Health Products

Store	Similarity
Store A selling soft drink	0.9776
Store A selling supplements	0.9769
Store A selling health products	0.9750
Store B selling soft drink	0.9741

TABLE IV
Top 4 Similar Stores to the Store A Selling Foods

Store	Similarity
Store B selling foods	0.9306
Store C selling foods	0.9287
Store A selling special local products	0.9218
Store D selling foods	0.9160

Table IV shows that there is a store that sells special products which is possible to direct users.

TABLE V
Characteristics of Consumers Who Are Highly Similar to Those Who Shop of Alcohol and Food Products in Drug Store A and Food Store E

Store	Similarity	Product
Drug store B, Food store B, Food store C	0.9362	Instant product A, soft drink A, Food A
Pet food store A, Food store D, Universal store A	0.9305	Dry Food A, Soy product A, Dry food B
Low priced supply store A, Drug store C, Food store E	0.9281	Cosmetics A, Daily necessities A, Cool drinks A
Food store F, Digital book store A, Discount supply store A	0.9271	Seafood A, Electronic books A, Drinking water A

Using the proposed model, we obtain user distributed representation that includes information of products and stores. By calculating users that are similar to a given user, we can calculate shops and products that match the preferences of those users. Table V shows four consumers with high similarity to consumers who shop mostly at drug store A and food store E, as well as shops and products with high similarity. Results indicate that the proposed method can extract similar stores and products appropriately and also suggest the validity of the proposed method.

TABLE VI
Characteristics of Consumers Who Are Highly Similar to Those Who Shop for Stationery and Household Goods in Stationery Store A and General Store A

Store	Similarity	Product
Book store A, Digital book store A, Electronic store A	0.8082	Comic A, Textbook A, E-book A, Washing machine A
General merchandise store B, pet store A, pet care supply store A	0.8007	Camelot A, tool A, goldfish A, goldfish B, sink A
Goods Store A, Electricity Store B, Daily Goods Store A, Health Store A	0.7973	Cosmetics A, hygiene products A, deodorizers A, chargers A, nutrition materials A
Book store A, Electricity Store C, Food Store A, Health Management Food Store A, Furniture Store A	0.7963	Novel A, Novel B, Antenna A, Ramen A, Frozen lunch box A, Engineering table B

Table VI shows the four consumers with high similarity to consumers who shop mostly at Stationery Store A and General

Merchandise Store A, as well as the stores and products with high similarity. Thus, we can recommend both stores and products for consumers in online markets.

C. Consideration

From the results of the analysis in Tables I and II, it is possible to classify products in the semantic space and visualize them. In addition, it is possible to calculate the similarity that fits intuitively. We suggest that this property can be used as a recommendation. For example, it is possible to monitor products purchased by consumers and display the recommended products in real time. The results of the analysis in Tables III and IV show that stores with similar characteristics in the semantic space can be extracted well. In such cases, we can consider measures such as collaboration between stores that sell different products but have similar concepts. From the results of Tables V and VI, we can conclude that the products often purchased, and stores are often used by other users who have a considerable similarity with a certain user who is similar to the stores and products often purchased by the user. Therefore, it can be used to recommend other user stores and products to the user.

The relationship between the products and stores is obtained by combining the two models; therefore, they do not necessarily coincide. In some cases, the recommended store does not sell the recommended product. Therefore, it is necessary to consider whether the store sells the product. The results may vary depending on the dimensionality of the feature space and the changes in the activation function. The results of the analysis must be used to determine whether they are robust to the parameters. There are several other possible ways to combine these two models. This also needs to be verified.

V. Conclusion

In this study, we proposed a method for integrated analysis by constructing a relationship model between consumers, products, and stores using Doc2vec, which has been studied in the field of machine learning. In addition, we applied the proposed method to the evaluation history data of Rakuten ICHIBA in 2014 [10] and analyzed and examined the validity of the analysis.

While applying the proposed model to actual marketing, there may be some products that are actually sold in stores and some products that are similar to those extracted by the proposed model. Therefore, the next task is to extract combinations of stores and products that have a high purchase probability. Moreover, a method for combining two user vectors should be considered in the future.

Acknowledgment

In this paper, we used "Rakuten Dataset" provided by Rakuten, Inc. via IDR Dataset Service of National Institute of Informatics.

References

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

[2] Yoav Goldberg, Omer Levy.word2vec Explained: Deriving Mikolov et al.'s Negative-Sampling Word-Embedding Method. February 14, 2014

[3] Quoc Le, Tomas Mikolov, Distributed representations of sentences and documents. In: International conference on machine learning2014, pp. 1188-1196

[4] Lap Q. Trieu, Huy Q. Tran, Minh-Triet Tran. News Classification from Social Media Using Twitter-Based Doc2Vec Model and Automatic Query Expansion. DOI: 10.1145/3155133

[5] H. Lee, Y. Yoon, Engineering doc2vec for automatic classification of product descriptions on O2O applications, Electron Commer Res 18, 433-456 2018. https://doi.org/10.1007/s10660-017-9268-5

[6] J.H. Lau, T. Baldwin, An empirical evaluation of doc2vec with practical insights into document embedding generation. In:Proc. RepL4NLP 2016, pp. 78-86

[7] Diederik P. Kingma, Jimmy Lei Ba, ADAM: A method for stochastic optimization, ICLR 2015

[8] L. Chen, G. Feng, C. Leong, B. Lehman, M. Martin-Raugh, H. Kell, et al., "Automated scoring of interview videos using Doc2Vec multimodal feature extraction paradigm."Proc. of ACM ICMINov. 2016

[9] R. Nath Nandi, M. M. Arefin Zaman, T. Al. Muntasir, S. Hosain Sumit, T. Sourov, M. Jamil-Ur Rahman, "Bangla News Recommendation Using doc2vec,"International Conference on Bangla Speech and Language Processing (ICBSLP), Sylhet, Bangladesh, 2018, pp. 1-5, doi: 10.1109/ICBSLP.2018.8554679

[10] Rakuten, Inc. (2014): Rakuten Dataset. Informatics Research Data Repository, National Institute of informatics. (dataset). https://doi.org/10.32130/idr.2.0